HHS Public Access

Author manuscript

J Agric Biol Environ Stat. Author manuscript; available in PMC 2016 May 02.

Published in final edited form as:

J Agric Biol Environ Stat. 2015 December ; 20(4): 614–628. doi:10.1007/s13253-015-0230-5.

Empirical Bayes analysis of RNA-seq data for detection of gene expression heterosis

Jarad Niemi[*], **Eric Mittman**, **Will Landau**, and **Dan Nettleton**

Department of Statistics, Iowa State University, Ames, Iowa, U.S.A.

Abstract

An important type of heterosis, known as hybrid vigor, refers to the enhancements in the phenotype of hybrid progeny relative to their inbred parents. Although hybrid vigor is extensively utilized in agriculture, its molecular basis is still largely unknown. In an effort to understand phenotypic heterosis at the molecular level, researchers are measuring transcript abundance levels of thousands of genes in parental inbred lines and their hybrid offspring using RNA sequencing (RNA-seq) technology. The resulting data allow researchers to search for evidence of gene expression heterosis as one potential molecular mechanism underlying heterosis of agriculturally important traits. The null hypotheses of greatest interest in testing for gene expression heterosis are composite null hypotheses that are difficult to test with standard statistical approaches for RNA-seq analysis. To address these shortcomings, we develop a hierarchical negative binomial model and draw inferences using a computationally tractable empirical Bayes approach to inference. We demonstrate improvements over alternative methods via a simulation study based on a maize experiment and then analyze that maize experiment with our newly proposed methodology. This article has supplementary material online.

Keywords

Hierarchical model; Negative binomial; RNA-seq; Bayesian LASSO; Parallel computing; Hybrid vigor

1. Introduction

Heterosis exists when the expected value of a hybrid phenotype differs from the average of the expected phenotypic values of the hybrid's parents. The most interesting and useful form of heterosis, known as hybrid vigor, occurs when hybrid progeny display a mean phenotype that is superior to both parental phenotypic means. This heterosis phenomenon was scientifically documented in plants by Darwin (1876) and has long been used to improve agricultural production. One classic example involves hybrid maize offspring that are taller, faster to mature, and yield considerably more grain than their inbred parents (Hallauer and Miranda, 1981; Hallauer et al., 2010).

[*]niemi@iastate.edu.

Supplementary Materials

The supplementary materials include links to the data as well as code to reproduce the analysis in this manuscript.

Depending on whether large or small values of a phenotype are favorable, hybrid vigor can occur if the mean phenotype of a hybrid is greater than both parental means or less than both parental means. We refer to the former as high-parent heterosis (HPH) and the latter as low-parent heterosis (LPH). Note that heterosis, HPH, and LPH are similar to the quantitative genetics concepts of dominance, overdominance, and underdominance. However, the various forms of dominance are usually reserved for describing the association of mean phenotype with homozygous and heterozygous genotypes at a single genetic locus. Heterosis involves a comparison of inbreds (simultaneously homozygous at many loci) and hybrids (simultaneously heterozygous at many loci). For simplicity, throughout the remainder of the article, we will use the term *extreme heterosis* (EH) to describe the situation where the hybrid mean is more extreme than the parental means, i.e. we say there is EH if and only if either HPH or LPH holds.

Despite intensive study and successful use of heterosis in agriculture, the basic molecular genetic mechanisms remain poorly understood (Chen, 2013). One potential explanation, is EH of gene expression, i.e. enhanced (or suppressed) expression of one or more genes in hybrids compared to their inbred parents. Gene expression heterosis has been investigated in maize by Swanson-Wagner et al. (2006) and Springer and Stupar (2007) and EH of gene expression is conceptually depicted in the right column of Figure 1b in Chen (2013).

Recently, Ji et al. (2014) introduced an approach to assess gene expression heterosis using microarray data under the assumption that these data are continuous. They built a normal hierarchical model for microarray measurements of transcript abundance that allows borrowing of information across genes to estimate means and variances. They introduced an empirical Bayes framework that first estimates model hyperparameters, then estimates the posterior distribution for gene-specific parameters conditional on those hyperparameters, and finally computes heterosis probabilities based on integrals of regions under this posterior. This development was necessary due to the composite null hypotheses in tests for heterosis. These hypotheses, which many available methods do not fully accommodate, remain a challenge in the transition from continuous measurements of transcript abundance to count-based measurements that arise from RNA sequencing (RNA-seq) technology. Building on the work of Ji et al. with the normal data model, we construct a hierarchical model based on a negative binomial data model. We also utilize an empirical Bayes approach to obtain estimates of the hyperparameters and the posterior distributions for the gene-specific parameters conditional on those hyperparameters.

The remainder of the paper proceeds as follows. Section 2 presents the proposed hierarchical model, an empirical Bayes method of estimating the parameters, and the calculation of posterior probabilities of EH. Section 3 presents a simulation study based on a maize experiment and compares our approach to alternative methods. Section 4 analyzes a maize experiment where hybrid vigor is well established and identifies genes demonstrating EH of expression. Section 5 summarizes the work and suggests directions for future research.

2. Empirical Bayes identification of gene expression heterosis from RNA-seq read counts

We consider an RNA sequencing (RNA-seq) experiment that involves at least three genetic varieties: two parental varieties and a cross between these two varieties called the hybrid. For each variety, replicate RNA samples are isolated and assessed for quality. Complementary DNA (cDNA) libraries, consisting of short cDNA fragments derived from RNA, are constructed. Then, next generation sequencing technology is used to determine the *reads*, or nucleotide sequences, in the cDNA libraries. These reads are processed using bioinformatic algorithms to match the reads to genes, or specific gene transcripts, exons, microRNAs, etc. The results of read processing are summarized by a gene × sample matrix of counts. See Datta and Nettleton (2014) for more details on RNA-seq experiments and data from a statistical perspective, and see Paschold et al. (2012) for the biological background behind the use of RNA-seq to study gene expression heterosis.

To use RNA-seq counts to identify genes displaying EH of expression, we build a hierarchical model to borrow information across gene-variety means and across gene-specific overdispersion parameters, estimate the hyperparameters using an empirical Bayes procedure, and calculate empirical Bayes posterior probabilities for EH.

2.1 Hierarchical model for RNA-seq counts

Let Y_{gvi} be the count for gene $g = 1, \ldots, G$, variety $v = 1, \ldots, V$, and replicate $i = 1, \ldots, n_v$. We assume

$$Y_{gvi} \overset{ind}{\sim} \mathrm{NB}\left(e^{\mu_{gv}+\gamma_{vi}}, e^{\psi_g}\right) \quad (1)$$

where NB(ξ, e^{ψ}) indicates a negative binomial distribution with expectation ξ and variance $\xi + e^{\psi}\xi^2$, and *ind* indicates the observations are conditionally independent. As shown in equation (1), our data model involves gene-specific overdispersion ψ_g and a mean that depends on the gene-variety combination through μ_{gv} and on the sample though γ_{vi}. The μ_{gv} terms are of primary scientific interest; the γ_{vi} terms are normalization factors that account for differences in the thoroughness of sequencing from sample to sample.

Following Ji et al. (2014), we reparameterize the gene-variety mean structure into the genespecific parental average (ϕ_g), half-parental difference (α_g), and hybrid effect (δ_g). For our heterosis study where $V = 3$, we let $v = 1, 2$ indicate the two parental varieties and $v = 3$ indicate the hybrid. The reparameterization is

$$\phi_g = \frac{\mu_{g^1} + \mu_{g^2}}{2}, \quad \alpha_g = \frac{\mu_{g^1} - \mu_{g^2}}{2}, \quad \text{and} \quad \delta_g = \mu_{g^3} - \phi_g.$$

We assume a hierarchical model for the gene-specific mean parameters and overdispersion parameters. Initially, we assume the parental averages, half-parental differences, hybrid effect, and overdispersion parameters follow normal distributions, i.e.

$$\phi_g \overset{ind}{\sim} N\left(\eta_\phi, \sigma_\phi^2\right), \alpha_g \overset{ind}{\sim} N\left(\eta_\alpha, \sigma_\alpha^2\right) \delta_g \overset{ind}{\sim} N\left(\eta_\delta, \sigma_\delta^2\right), \quad \text{and} \quad \psi_g \overset{ind}{\sim} N\left(\eta_\psi, \sigma_\psi^2\right).$$

Empirical plots of estimated values of α_g and δ_g for our maize data set (described in Section 4) suggest that the distribution of these parameters are more peaked near zero and have heavier tails than a normal distribution allows. For differential expression between the two parental phenotypes, we would expect many genes to have small effects and only a few genes to have relatively large effects. For these many genes with small effects, we might expect the hybrid to act similar to its parents and therefore also have many genes where the hybrid effect is small and only a few genes where this hybrid effect is large. For these reasons, we also assessed Laplace (or double exponential) distributions for the half-parental difference and hybrid effect and thus implement a Bayesian LASSO (Park and Casella, 2008; Hans, 2009), i.e.

$$\alpha_g \overset{ind}{\sim} La\left(\eta_\alpha, \sigma_\alpha\right) \quad \text{and} \quad \delta_g \overset{ind}{\sim} La\left(\eta_\delta, \sigma_\delta\right)$$

where $a \sim La(\eta, \sigma)$ has a probability density function given by $La(a; \eta, \sigma) = \exp(-|a - \eta|/\sigma)/2\sigma$ with expectation η and variance $2\sigma^2$. Whether using normal or Laplace distributions, we assume *a priori* independence amongst the parental averages, half parental differences, hybrid effects, and overdispersion parameters.

2.2 Empirical Bayes

Initial attempts to perform a fully Bayesian analysis via Markov chain Monte Carlo (MCMC) failed due to high computational costs and poor mixing in the resulting chains. For example, we implemented the model in the statistical software Stan (discussed at the end of this section), ran the MCMC on a simulated data set with 10,000 genes for 2 months on a state-of-the-art linux server, and obtained potential scale reduction factors (Gelman and Rubin, 1992) that suggested we would need to run at least ten times as long to obtain convergence. Although there are certainly improvements that could be made to decrease computational costs and improve mixing, we opted for an empirical Bayes approach. This approach may be a reasonable approximation to a fully Bayesian approach when estimating models with large numbers of genes as the posterior distributions for the hyperparameters may be tightly peaked.

We categorize the parameters of the model in Section 2.1 into gene-specific parameters $\theta = (\theta_1, \ldots, \theta_G)$ where $\theta_g = (\phi_g, \alpha_g, \delta_g, \psi_g)$, normalization factors $\gamma = (\gamma_{11}, \ldots, \gamma^{VnV})$, and hyperparameters $\pi = (\eta, \sigma)$ where $\eta = (\eta_\phi, \eta_\alpha, \eta_\delta, \eta_\psi)$ and $\sigma = (\sigma_\phi, \sigma_\alpha, \sigma_\delta, \sigma_\psi)$. We obtain estimates for the hyperparameters and then base gene-specific inference on the posterior conditional on these estimates.

To obtain normalization factors $\hat{\gamma}$ we use the weighted trimmed mean of M values (TMM) method of Robinson and Oshlack (2010). We use edgeR to obtain genewise dispersion estimates, $\hat{\psi}_g$, and the generalized linear model methods to obtain estimates for the

remaining gene-specific parameters, i.e. $\hat{\phi}_g, \hat{\alpha}_g, \quad \text{and} \quad \hat{\delta}_g$ (Robinson et al., 2010). Using $\hat{\theta}_g = \left(\hat{\phi}_g, \hat{\alpha}_g, \hat{\delta}_g, \hat{\psi}_g\right)$ we estimate hyperparameters for the location and scale parameters in the hierarchical model using a central method of moments approach. For example, $\hat{\eta}_\phi = \sum_{g=1}^{G} \hat{\phi}_g / G \quad \text{and} \quad \hat{\sigma}_\phi^2 = \sum_{g=1}^{G} \left(\hat{\phi}_g - \hat{\eta}_{\hat{\phi}}\right)^2 / (G-1)$ (and similarly for $\hat{\eta}_\psi$ and $\hat{\sigma}_\psi$) and $\hat{\eta}_\alpha = \sum_{g=1}^{G} \hat{\alpha}_g / G$ and $2\hat{\sigma}_\alpha^2 = \sum_{g=1}^{G} (\hat{\alpha}_g - \hat{\eta}_\alpha)^2 / (G-1)$ (and similarly for $\hat{\eta}_\delta$ and $\hat{\sigma}_\delta$) for the model assuming Laplace distributions.

Conditional on the estimated normalization factors $\hat{\gamma}$ and hyperparameters $\hat{\pi}$ we perform a Bayesian analysis to re-estimate the gene-specific parameters and describe their uncertainty. Equation 2 shows that conditional on $\hat{\gamma}$ and $\hat{\pi}$ the gene-specific parameters are independent and therefore conditional posterior inference across the genes can be parallelized. In this equation, the densities for α_g and δ_g will depend on whether we are assuming normal or Laplace distributions.

$$p(\theta|y, \hat{\pi}, \hat{\gamma}) \propto \prod_{g=1}^{G} \left[\prod_{v=1}^{V} \prod_{i=1}^{n_v} \mathrm{NB}\left(y_{gvi}; e^{\mu_{gv} + \hat{\gamma}_{vi}}, e^{\psi_g}\right) N\left(\phi_g; \hat{\eta}_\phi, \hat{\sigma}_\phi^2\right) p\left(\alpha_g; \hat{\eta}_\alpha, \hat{\sigma}_\alpha\right) p\left(\delta_g; \hat{\eta}_\delta, \hat{\sigma}_\delta\right) N\left(\psi_g; \hat{\eta}_\psi, \hat{\sigma}_\psi^2\right) \right] \quad (2)$$

To perform the conditional posterior inference on the gene-specific parameters, we used the statistical software Stan (Stan Development Team, 2014b) executed through the RStan interface (Stan Development Team, 2014a) in R (R Core Team, 2014). Stan implements a variant of MCMC called Hamiltonion Monte Carlo (Neal, 2011) to obtain samples from the posterior in equation (2). We ran 4 simultaneous chains with random initial starting values for 1000 burn-in (and tuning) iterations followed by another 1000 iterations retaining every fourth sample (to reduce storage space) for inference. We monitored convergence using the potential scale reduction factor and effective sample size (ESS) for ϕ_g, α_g, δ_g, and ψ_g (Gelman and Rubin, 1992). If the minimum ESS was less than 1000, we reran the chains with double the iterations for both burn-in and inference. We continued this restarting and doubling until we obtained minimum ESS greater than 1000 for all parameters.

2.3 Gene expression heterosis

In the maize context that motivates this work, we are interested in extreme heterosis (EH), i.e. either low-parent heterosis (LPH) or high-parent heterosis (HPH), in gene expression. For a specific gene g, LPH occurs when expected expression in the hybrid is less than the expected expression of either parent, i.e. $\mu_{g3} < \min\{\mu_{g1}, \mu_{g2}\}$ or, equivalently, $\delta_g < -|\alpha_g|$, and HPH occurs when expected expression in the hybrid is greater than the exptected expression of either parent, i.e. $\mu_{g3} > \max\{\mu_{g1}, \mu_{g2}\}$ or, equivalently, $\delta_g > |\alpha_g|$. We evaluate these probabilities based on empirical Bayes estimates of their posterior probabilities, e.g.,

$$P(\mathrm{LPH}_g|y, \hat{\pi}, \hat{\gamma}) = P(\delta_g < -|\alpha_g \parallel y, \hat{\pi}, \hat{\gamma}) \approx \frac{1}{M} \sum_{m=1}^{M} \mathrm{I}\left(\delta_g^{(m)} < -|\alpha_g^{(m)}|\right) \quad (3)$$

where $\left(\delta_g^{(m)}, \alpha_g^{(m)}\right)$ is the m^{th} MCMC sample from the empirical Bayes posterior, and I(A) is 1 if A is true and 0 otherwise. HPH probability is defined similarly with the inequality reversed and without the negative sign. We construct a ranked list of genes according to the maximum of the gene's LPH and HPH heterosis probabilities. Geneticists can use this list to prioritize future experiments to understand the molecular genetic mechanisms for heterosis.

2.4 *Implementation in* ShrinkBayes

In addition to the approach above, we utilized ShrinkBayes to estimate EH probabilities with two modifications described here. ShrinkBayes utilizes integrated nested Laplace approximation (INLA) (Rue et al., 2009) in combination with empirical Bayes ideas (van de Wiel et al., 2014). One limitation with inferential methods based on INLA is that all distributions, except for the data distribution, must have tails as light or lighter than the normal density. Thus, we cannot implement the Laplace priors for the half-parental difference (a_g) and the hybrid effect (δ_g) and instead use normal priors in this situation. An additional limitation is that INLA provides approximations to marginal posteriors for parameters or linear combinations of parameters, but not an approximation to the full posterior. Since we are interested in non-linear quantities such as $P(\delta_g > |a_g||y)$, we cannot compute these directly using ShrinkBayes. Instead, for ShrinkBayes, we calculate EH probabilities conditional on posterior means for the half-parental difference and hybrid effect, i.e. $\hat{\delta}_g$ and $\hat{\alpha}_g$ For example,

$$P_{\text{shrinkBayes}}\left(\text{LPH}_g|y\right) \equiv P\left(\delta_g < -\alpha_g|y\right) \text{I}\left(\hat{\delta}_g < -\hat{\alpha}_g \leqslant 0\right) + P\left(\delta_g < \alpha_g|y\right) \text{I}\left(\hat{\delta}_g < \hat{\alpha}_g \leqslant 0\right).$$

HPH probability is defined similarly with all inequalities reversed. As before, we construct a ranked list of genes according to the maximum of the gene's LPH and HPH heterosis probabilities.

We will use the term eBayes to refer to the approach defined in Sections 2.1–2.3 and add parenthetical labels "normal" and "Laplace" to specify whether we are assuming normal or Laplace distributions for half-parental differences and hybrid effects.

3. Simulation study based on a maize experiment

To assess the efficacy of our method to identify genes demonstrating EH, we used a maize data set with parental varieties B73 and Mo17 and the hybrid variety (B73 × Mo17) (Paschold et al., 2012) to determine realistic parameter values for a simulation study. Section 4 describes the maize dataset in detail. We compared our method to approaches using the R packages edgeR, baySeq (Hardcastle and Kelly, 2010; Hardcastle, 2012), and ShrinkBayes (van de Wiel et al., 2014).

3.1 Constructing simulated data

We used the methods described at the beginning of Section 2.2 to obtain normalization factors $\hat{\gamma}$ and gene-specific parameter estimates $\hat{\theta}_g$ for all genes using the edgeR package (Robinson et al., 2010) from Bioconductor (Gentleman et al., 2004) applied to the maize

data on 27,888 genes with average count at least one and at most two zero read counts for each variety across the four replicates. This analysis produced sample-specific normalization factors of $\hat{\gamma}$ = (0.074, −0.059, −0.074, −0.014, −0.014, −0.124, 0.093, 0.063, 0.021, −0.037, 0.049, 0.021). The gene-specific parameter estimates were treated as the true parameter values for the simulation study so that our simulated datasets mimicked the existing structure among the gene-variety means of the original maize data.

Using these parameters and normalization factors, we simulated data according to the negative binomial model in equation (1) independently for each gene. For each simulation, we analyzed a subset of 25,000 genes selected randomly from genes with simulated counts at least one on average and with at most two zeros for each variety across replications. We repeated this simulation process 10 times for each of 4, 8, and 16 replicates per variety, reusing normalization factors when necessary.

For a particular gene, the truth was determined via the estimated values for α_g and δ_g Specifically, if $|\hat{\delta}_g| > |\hat{\alpha}_g|$ the gene was considered to have EH. For many heterosis genes, the value of $|\hat{\delta}_g|$ was only slightly larger than $|\hat{\alpha}_g|$. Thus there are many genes in these simulated data sets that are considered to have EH, but whose signal in the simulations will be extremely small. Conversely, there are many non-heterosis genes whose value of $|\hat{\delta}_g|$ was only slightly smaller than $|\hat{\alpha}_g|$ but whose simulated data will look similar to many EH genes. Therefore, we expect it will be difficult to accurately identify EH genes, but believe this level of difficulty is representative of real applications.

3.2 Alternative methods

We compared our method to that of Ji et al. (2014), which assumes normality in the response, by modeling the logarithm of the RNAseq count plus one adjusted by the normalization factor, i.e. $\log(Y_{gvi} + 1) - \gamma_{vi}$. Use of the normalization factor here provides two advantages: 1) counts are properly adjusted for the thoroughness of the sequencing of the sample and 2) for genes with no count variation within variety (which actually occurs in our maize data set), use of the normalization factors allows the approach of Ji et al. to still execute.

In addition to the approach of Ji et al., we modified two existing RNA-seq approaches, edgeR and baySeq, for use in the heterosis context. For each method, we attempted to provide a measure of the strength of EH for each gene such that large values of this measure indicate support for EH. edgeR can be used to test for differential expression between any two varieties based on the fit of a negative binomial log-linear model (Robinson and Smyth, 2007; Robinson et al., 2010). To construct a measure of EH, we computed the maximum likelihood estimates of the μ_{gv} parameters for all genes using edgeR's built-in Fisher scoring algorithm, and then used likelihood ratio tests to calculate two p-values for each gene: p_{g1} for testing $H_{g01} : \mu_{g1} = \mu_{g3}$ and p_{g2} for testing $H_{g02} : \mu_{g2} = \mu_{g3}$. Then, for each gene, we defined a new p-value denoted as p_g and set to $p_g = 1$ if the estimate of μ_{g3} falls between the estimates of μ_{g1} and μ_{g2} and $p_g = \max\{p_{g1}, p_{g2}\}/2$ otherwise. For all relevant significance thresholds ω near 0, it can be shown that rejecting the null hypothesis of no EH for gene g

whenever $p_g \leq \omega$ results in a test that is asymptotically size ω. We then use $1 - p_g$ as a final measure positively associated with strength of evidence for EH.

baySeq allows for a wider range of hypotheses for each gene, including H^* : $\mu_{g1} = \mu_{g2} = \mu_{g3}$, H^* : $\mu_{g1} = \mu_{g2}$, H^* : $\mu_{g1} = \mu_{g3}$, H^* : $\mu_{g2} = \mu_{g3}$, and H^* : all μ_g 's are distinct. In a technique similar to our application of edgeR, we used the posterior probabilities of these hypotheses to construct a measure of EH for each gene. We set this measure to zero if the maximum likelihood estimate, calculated using edgeR, of μ_{g3} is between the maximum likelihood estimates of μ_{g1} and μ_{g2}. Otherwise, the measure is the sum of the posterior probabilities of H^*_{g2} and H^*_{g5}, the two hypotheses that allow for EH.

3.3 Results

For the methods in Sections 2 and 3.2, we sorted genes according to the computed measure of the strength of evidence for EH. From these sorted lists, we constructed receiver-operating characteristic (ROC) curves to evaluate the ability of these methods to distinguish genes with EH, as defined in Section 3.1, from those without EH. A representative set of ROC curves is shown in Figure 1.

The ROC curves indicate modest performance, e.g. for a false positive rate of 5%, the best performing methods only achieved a true positive rate of just over 15%. This is consistent with our expectation discussed in Section 3.1 due to the low signal-to-noise ratio in these simulated data.

For this simulation, we can see that the approaches based on the model in Section 2.1, i.e. eBayes and ShrinkBayes, provide the best true positive rate for a given false positive rate. Also, as expected, as the sample size increases, our ability to distinguish genes with EH from genes without improves.

Figure 2 provides the area under the ROC curve (AUC) below a false positive rate of 0.1 across the 10 simulations for each of the 3 different sample sizes.

Similar to the single ROC curve, the eBayes and ShrinkBayes methods appear to outperform the other methods in terms of AUC. This improvement ranges from about a 20% improvement over Ji et al. to about a 100% improvement over edgeR (which was not designed for heterosis testing).

With 4 replicates per variety, there does not appear to be much of a difference between ShrinkBayes and the eBayes approaches, but as the number of replicates increases, the eBayes approaches appear to improve relative to ShrinkBayes. Two differences exist between the ShrinkBayes and eBayes approaches that could explain the difference in AUC: 1) ShrinkBayes utilizes a different empirical Bayes approach for estimating both the hyperparameters and the gene-specific parameters and 2) the measure of EH is slightly different due to INLA not providing a full posterior.

There also appears to be a pattern of the eBayes (Laplace) systematically performing better than eBayes (normal). We suspect this is because the Laplace distributions are better

approximations to the true underlying distribution for these parameters, and we discuss this in Section 5.

4. Searching for gene expression heterosis in the maize experiment

We used our method to analyze a maize data set (Paschold et al., 2012) of RNA-seq gene expression in parental lines B73 and Mo17 and the hybrid genotype (B73×Mo17) with a total of 39,656 genes. Each variety had four biological replicates measured with Illumina methodology and equipment. Reads were mapped to the whole reference genome using the short reads aligner, NOVOALIGN. For more specifics, please see Paschold et al. (2012).

We analyzed the data using all the methods compared in the previous section. The computation time on a desktop with two 4-core 3.6GHz Intel Xeon processor was 14 seconds for edgeR, 1.3 minutes for Ji et al., 10 hours for the eBayes approaches, and 17 hours for baySeq. In the eBayes approach, the vast majority of the time is spent on independent MCMC analysis for each gene. Thus we parallelized this step using doMC (Analytics, 2014) and plyr (Wickham, 2011). When parallelized across 5 cores, the eBayes approaches took about 2.5 hours. ShrinkBayes took 12 hours on a cluster node with two 8-core 2.6GHz Intel Haswell E5-2640 v3 processors where the code was also parallelized across 5 cores.

For the eBayes (Laplace), we estimated the hyperparameters to be $\hat{\eta}_\phi=3.44, \quad \hat{\sigma}_\phi=2.74$, $\hat{\eta}_\alpha=-0.05, \quad \hat{\sigma}_\alpha=0.36, \quad \hat{\eta}_\delta=0.08, \hat{\sigma}_\delta=0.33, \hat{\eta}_\psi=-2.35$, and $\hat{\sigma}_\psi=0.60$. We then performed independent MCMC analysis for each gene conditional on these hyperparameters. As with the simulation study, we ran 4 chains simultaneously and doubled the number of MCMC iterations until each gene-specific parameter had an effective sample size above 1,000.

Figure 3 provides point estimates of the gene-specific parameters from the initial edgeR estimation step and after the eBayes (Laplace) procedure described in Section 2.2. This figure shows shrinkage for large absolute estimates of α_g and δ_g from the edgeR estimation toward $\hat{\eta}_\alpha$ and $\hat{\eta}_\delta$ which are both approximately zero. The figure also shows decreased values for the overdispersion parameter with larger decreases for high and low values of overdispersion. Finally, very little, if any, shrinkage is observed for ϕ_g estimates.

With posteriors for all parameters, we can calculate empirical Bayes posterior probabilities for LPH and HPH. For each gene, the quantity of interest is the maximum of these two probabilities. For each gene with a high probability of either HPH or LPH, the magnitude of the effect is of interest, thus we calculate

$$\text{estimated} \quad \text{effect} \quad \text{size}_g \equiv \begin{cases} \hat{\delta}_g - |\hat{\alpha}_g| & \text{if} \quad \hat{\delta}_g > \quad |\alpha_g| \\ \hat{\delta}_g - |\hat{\alpha}_g| & \text{if} \quad \hat{\delta}_g < -|\alpha_g|. \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

This estimated effect size is the difference between hybrid mean and the nearest parent with negative values indicating LPH and positive values indicating HPH. If the hybrid mean is estimated to be between the parents, this effect is defined to be zero.

Figure 4 provides a volcano plot, in this case a bivariate histogram, to visualize the maximum of the probabilities of LPH and HPH versus estimated effect size.

The figure shows a ridge at an effect size of zero for probabilities below 0.5. Above a probability of 0.5, we see a prototypical volcano pattern, with increased EH probability corresponding to larger estimated effect sizes and no estimated effect sizes near zero for genes with high EH probability. We also see asymmetry, with larger negative effect sizes than positive effect sizes due to genes with hybrid counts near zero and relatively high parental counts. Genes with high estimated posterior probabilities of EH and large estimated effect sizes are candidates for further investigation.

5. Discussion

Geneticists speculate that gene expression EH is one possible explanation of hybrid vigor of traits, such as plant height or grain yield. Existing methods for identifying differential gene expression based on RNA-seq data are not directly applicable for detecting EH genes. Ji et al. (2014) introduced an empirical Bayes approach based on a hierarchical model for microarray data. We followed their approach, modified to allow for RNA-seq read counts as measures of transcript abundance. We developed an empirical Bayes approach based on obtaining estimates for hyperparameters followed by MCMC to estimate gene-specific parameters. The empirical Bayes posteriors can be used to estimate posterior probabilities of high and low parent heterosis. Through a simulation study, we demonstrated that this method outperformed alternative methods, and performed comparably well with a similar model in ShrinkBayes, which estimates the posterior via INLA. We then demonstrated the use of the methodology on a maize experiment in which phenotypic heterosis is well known.

Although our method appears to hold some advantage over existing methods, we believe our approach can be improved by refining the hierarchical model for the gene-specific parameter distribution. Figure 5 shows marginal and bivariate histograms for eBayes (Laplace) posterior means for the gene-specific parameters.

These figures show departures from marginal model assumptions, e.g. normality assumptions for ϕ_g and ψ_g, and independence assumptions for (ϕ_g, ψ_g) and (a_g, δ_g). The plot of ϕ_g versus ψ_g shows a pattern where the mean overdispersion decreases as the mean expression level increases. The plot of a_g versus δ_g shows a rotated V pattern where δ_g appears to be equal to $|a_g|$. This V pattern is consistent with Mendel's Law of Dominance where the hybrid has mean expression equal to the parent with higher mean expression.

In addition to improving the hierarchical distribution, we believe better estimates of the parameters of this distribution, i.e. the hyperparameters, will also improve detection of gene-expression heterosis. Our current method, based on moment matching of essentially independently estimated gene-specific parameters, provides consistent estimators as the number of replicates per variety increases. But typically the number of replicates per variety is quite small. In our data there are only four replicates per variety, and therefore asymptotic justifications are deficient. There are a variety of alternative estimation approaches to explore, e.g. expectation-maximization algorithms or fully Bayesian approaches.

Notwithstanding these improvements, we believe our approach is a computationally efficient method that can immediately aid scientists who are interested in identifying candidate genes involved in a genetic mechanism of heterosis.

This paper has focused on statistical methods for detecting EH in gene expression using RNA-seq data. EH at the transcript level is only one of multiple molecular genetic mechanisms that may play roles in establishing hybrid vigor. Complementation (Paschold et al., 2012), allele-specific expression (Bell et al., 2013; Wei and Wang, 2013), and other complex forms of genomic and epigenetic interaction (Chen, 2013) are all plausible as mechanisms partially responsible for phenotypic heterosis. Further study of these phenomena using modern genomic technologies and appropriate statistical methods should enhance our understanding of hybrid vigor.

Acknowledgements

The authors thank Andrew Lithio for help in implementing our model in ShrinkBayes. Research reported in this publication was supported by National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM109458. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Analytics R. doMC: Foreach parallel adaptor for the multicore package. R package version 1.3.3. 2014

Bell GD, Kane NC, Rieseberg LH, Adams KL. RNA-seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. Genome biology and evolution. 2013; 5:1309–1323. [PubMed: 23677938]

Chen ZJ. Genomic and epigenetic insights into the molecular bases of heterosis. Nature Reviews Genetics. 2013; 14:471–482.

Darwin, C. The effects of cross and self fertilisation in the vegetable kingdom. John Murray; 1876.

Datta, S.; Nettleton, D. Statistical Analysis of Next Generation Sequencing Data. Springer; 2014.

Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statistical Science. 1992; 7:457–472.

Gentleman RC, Carey VJ, Bates DM. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology. 2004; 5:R80. [PubMed: 15461798]

Hallauer A, Miranda F. Quantitative genetics in maize breeding. Iowa St. Univ. Press, Ames, IA. 1981

Hallauer, AR.; Carena, MJ.; Miranda Filho, J. Quantitative genetics in maize breeding. Vol. 6. Springer; 2010.

Hans C. Bayesian lasso regression. Biometrika. 2009; 96,:835–845.

Hardcastle TJ. baySeq: Empirical Bayesian analysis of patterns of differential expression in count data. R package version 2.0.50. 2012

Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010; 11:422. [PubMed: 20698981]

Ji T, Liu P, Nettleton D. Estimation and testing of gene expression heterosis. Journal of Agricultural, Biological, and Environmental Statistics. 2014; 19:319–337.

Neal R. MCMC using Hamiltonian dynamics. Handbook of Markov Chain Monte Carlo. 2011; 2:113–162. Chapman & Hall/CRC.

Park T, Casella G. The Bayesian lasso. Journal of the American Statistical Association. 2008; 103:681–686.

Paschold A, Jia Y, Marcon C, Lund S, Larson NB, Yeh C-T, Ossowski S, Lanz C, Nettleton D, Schnable PS, et al. Complementation contributes to transcriptome complexity in maize (*Zea mays*

L.) hybrids relative to their inbred parents. Genome research. 2012; 22:2445–2454. [PubMed: 23086286]

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2014.

Robinson M, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology. 2010; 11:R25. [PubMed: 20196867]

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139–40. [PubMed: 19910308]

Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 2007; 23:6.

Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2009; 71:319–392.

Springer N, Stupar R. Allelic variation and heterosis in maize: How do two halves make more than a whole? Genome research. 2007; 17:264–275. [PubMed: 17255553]

Stan Development Team. RStan: the R interface to Stan, version 2.5.0. 2014a.

Stan Development Team. Stan: A C++ library for probability and sampling, version 2.5.0. 2014b.

Swanson-Wagner R, Jia Y, DeCook R, Borsuk L, Nettleton D, Schnable P. All possible modes of gene action are observed in a global comparison of gene expression in a maize f1 hybrid and its inbred parents. Proceedings of the National Academy of Sciences. 2006; 103:6805–6810.

van de Wiel MA, Neerincx M, Buffart TE, Sie D, Verheul HM. ShrinkBayes: a versatile R-package for analysis of count-based sequencing data in complex study designs. BMC bioinformatics. 2014; 15:116. [PubMed: 24766777]

Wei X, Wang X. A computational workflow to identify allele-specific expression and epigenetic modification in maize. Genomics, proteomics & bioinformatics. 2013; 11:247–252.

Wickham H. The split-apply-combine strategy for data analysis. Journal of Statistical Software. 2011; 40:1–29.

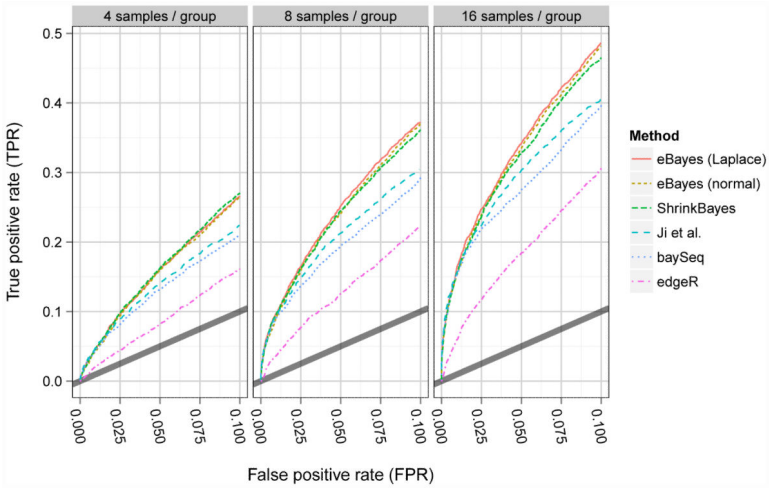

Figure 1.

Example ROC curves for false positive rates below 0.1 for the approaches using Ji et al., edgeR, and baySeq described in Section 3.2, the ShrinkBayes approach described in Section 2.4 and the eBayes approach described in Section 2.2 using both normal and laplace distributions for the half-parental difference and hybrid effect.

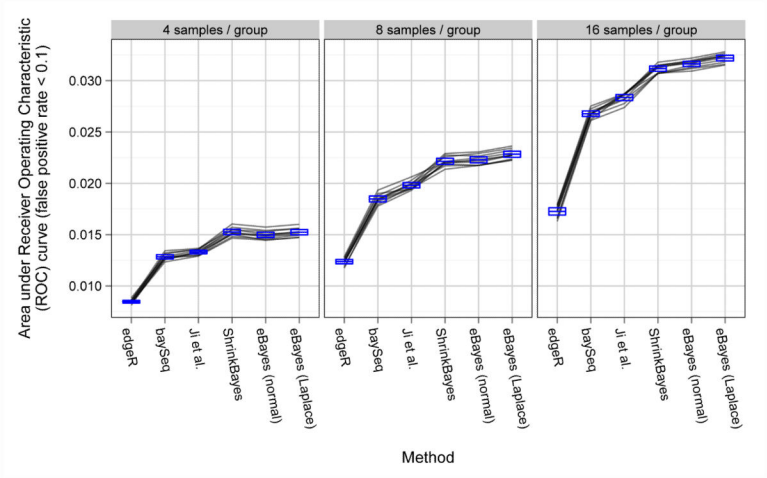

Figure 2.

Area under the ROC curves (AUC) below a false positive rate of 0.1 for 3 different replicates per variety for the approaches using Ji et al., edgeR, and baySeq as described in Section 3.2, the ShrinkBayes approach described in Section 2.4 and the eBayes approach described in Section 2.2 using both normal and laplace distributions for the half-parental difference and hybrid effect. Each line is a different simulation while the blue box indicates mean AUC (plus or minus one standard error).

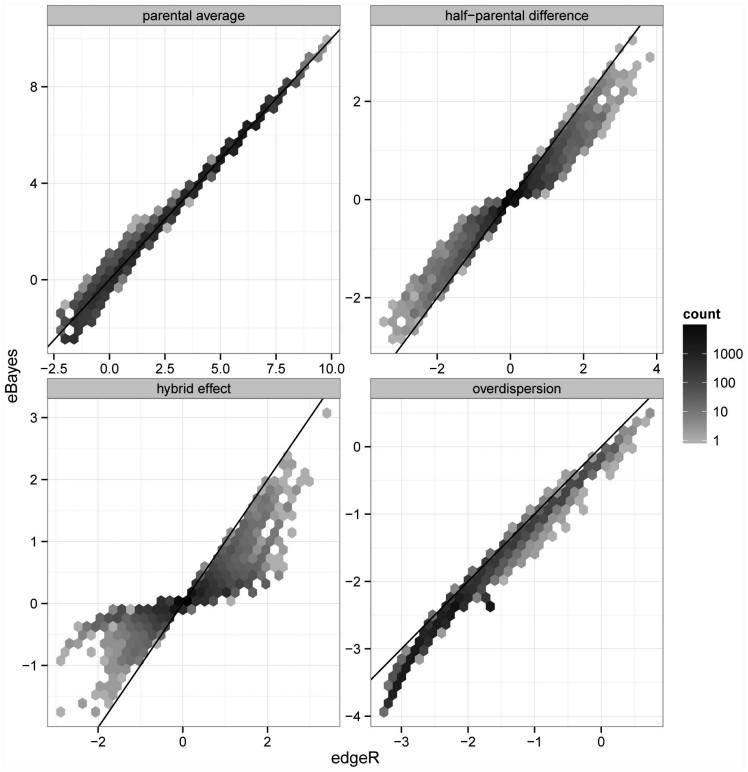

Figure 3.
Two-dimensional histogram of point estimates from edgeR and posterior means from eBayes (Laplace) along with the $y = x$ diagonal.

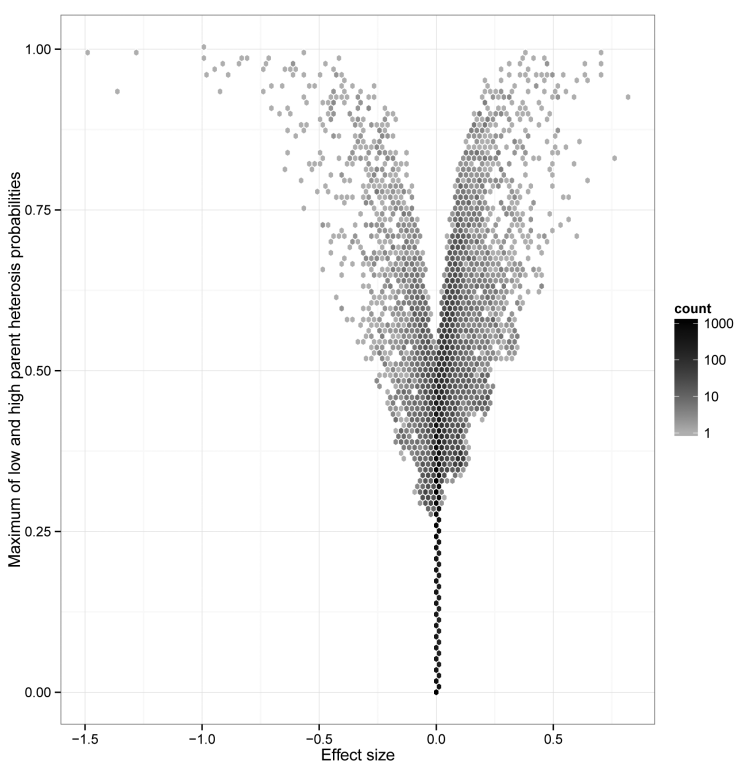

Figure 4.

A bivariate histogram of the maximum of the LPH and HPH probabilities versus estimated effect size defined in equation (4) for the B73 × Mo17 maize experiment using eBayes (Laplace).

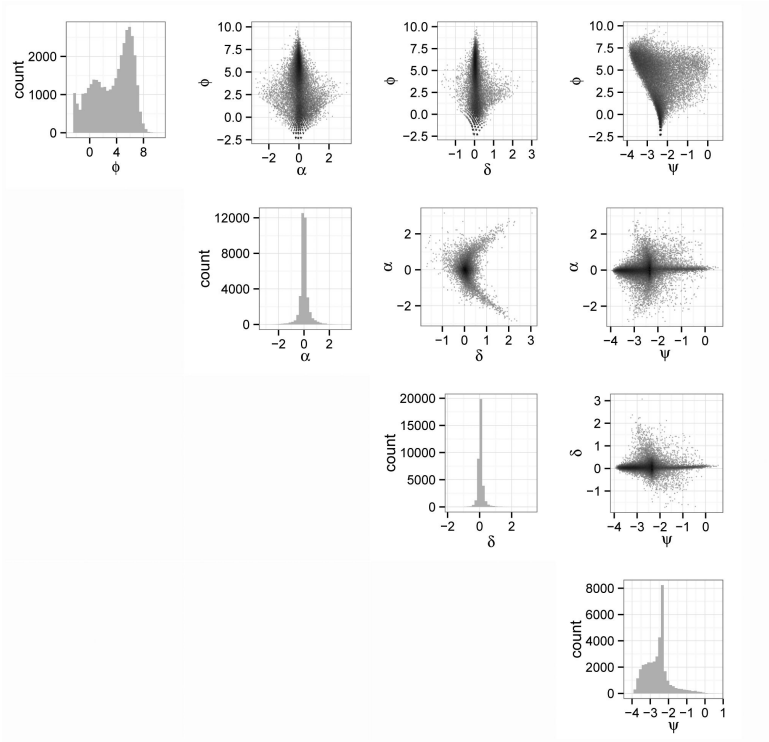

Figure 5.
Marginal and bivariate histograms of posterior means for gene-specific parameters for the B73 × Mo17 maize experiment.