Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Agric Biol Environ Stat. Author manuscript; available in PMC 2016 April 22.

Published in final edited form as:

J Agric Biol Environ Stat. 2015 December ; 20(4): 598–613. doi:10.1007/s13253-015-0232-3.

Hierarchical Modeling and Differential Expression Analysis for RNA-seq Experiments with Inbred and Hybrid Genotypes

Andrew Lithio[*] and **Dan Nettleton**

Department of Statistics, Iowa State University

Abstract

The performance of inbred and hybrid genotypes is of interest in plant breeding and genetics. High-throughput sequencing of RNA (RNA-seq) has proven to be a useful tool in the study of the molecular genetic responses of inbreds and hybrids to environmental stresses. Commonly used experimental designs and sequencing methods lead to complex data structures that require careful attention in data analysis. We demonstrate an analysis of RNA-seq data from a split-plot design involving drought stress applied to two inbred genotypes and two hybrids formed by crosses between the inbreds. Our generalized linear modeling strategy incorporates random effects for whole-plot experimental units and uses negative binomial distributions to allow for overdispersion in count responses for split-plot experimental units. Variations in gene length and base content, as well as differences in sequencing intensity across experimental units, are also accounted for. Hierarchical modeling with thoughtful parameterization and prior specification allows for borrowing of information across genes to improve estimation of dispersion parameters, genotype effects, treatment effects, and interaction effects of primary interest.

1. INTRODUCTION

Over the past decade, many statistical methods have been developed for analyzing high throughput RNA sequencing (RNA-seq) data. RNA-seq enables the sequencing of entire transciptomes, yielding counts associated with the mRNA abundance corresponding to each gene or genetic feature. Due to the cost of RNA-seq, experiments typically have relatively few experimental units, yet still result in high dimensional data, since there are often tens of thousands of genetic features measured for each experimental unit. To detect Differentially expressed (DE) genes, RNA-seq data are commonly analyzed using frequentist or moderated frequentist methods, such as those implemented in *edgeR* (Robinson, McCarthy and Smyth, 2010), *DESeq* (Anders and Huber, 2010), and *limma* (Smyth, 2005), but because of the high dimensionality, fully Bayesian methods are not often used.

edgeR and *DESeq* both use a negative binomial model with a generalized linear model (GLM) framework. This allows each package to accommodate arbitrary fixed-effects models, but neither allows for the use of random effects. The two packages differ in estimation of the negative binomial dispersion parameter, but both take a shrinkage approach, estimating a common or trended dispersion for the entire data set, then shrinking

[*]lithio@iastate.edu.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the dispersion estimates of each feature towards that common estimate or trend. *DESeq2* extends the idea of shrinkage across genetic features to logarithmic fold change estimates to help account for high variance in fold change estimates for low-count genes (Love et al., 2014).

Methods originally developed for the analysis of microarray data, including *limma*, have been adapted for RNA-seq data (Law et al., 2014). To extend to count data, *limma* uses the *voom* procedure, calculating a non-parametric estimate of the mean-variance relationship to generate weights for a linear model analysis of log transformed counts with empirical Bayes shrinkage of variance parameters. Law et al. (2014) argue that this procedure, and the use of log-transformed normal models, allows for more accurate modeling of the mean-variance relationship, while also yielding better small sample properties and permitting the use of a wider range of statistical tools than procedures based on count models.

Alternatives to both the count-based GLM and the transformed normal theory classes of methods include non-parametric approaches such as *samr* (Li and Tibshirani, 2013), and the empirical Bayes approach introduced by *baySeq* (Hardcastle and Kelly, 2010), which estimates posterior probabilities of a pre-specified set of models. Although also using the negative binomial distribution for the count data, model specification in *baySeq* essentially entails specifying different partitions of samples, where samples within each group share the same set of parameters. For a further introduction to these and other methods for Differential expression analysis of RNA-seq data, see Lorenz et al. (2014).

The most widely used statistical methods for RNA-seq data analysis discussed above have freely accessible software and are much more computationally efficient than fully Bayesian methods. The approach we pursue enjoys the flexibility and information-sharing capabilities of a fully Bayesian approach, while maintaining computational affordability via integrated nested Laplace approximation (INLA). INLA facilitates quick and accurate approximations of the marginal posteriors of latent Gaussian fields with a non-Gaussian response (Rue et al., 2009). The *R* package *ShrinkBayes* leverages the speed of INLA and the potential of parallel computing to facilitate an empirical-Bayes-type analysis of RNA-seq data, approximating the marginal posteriors of interest relatively quickly (van de Wiel et al., 2012). The empirical Bayesian approach provides a natural mechanism for borrowing information across genes for estimation of means and dispersion parameters. A major advantage of *ShrinkBayes* over commonly used frequentist-based methods is its ability to share information across genetic features while accounting for random effects in models for complex experimental designs.

In this paper, we illustrate the use of INLA and *ShrinkBayes* for the analysis of data from a complex experimental design like others common in agricultural studies. We analyze an RNA-seq data set from maize. The data consist of counts associated with the abundance of nearly 30,000 genetic features for replicate plant samples of four different genotypes, each grown under two different treatments. The data collection process gives the data additional split-plot structure. After constructing an appropriate model and estimating the hyperparameters of prior distributions, we illustrate estimation and inference for simple effects, main effects, and interactions.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The remainder of the paper is arranged as follows. Section 2 details the experimental design and structure of the data. Section 3 gives a brief review of INLA, the methods used in *ShrinkBayes*, and the model constructed for the analysis of the maize data. Section 4 reports results from fitting the model to the maize data. Section 5 summarizes a small simulation study, and we conclude with a discussion in Section 6.

2. DATA

Throughout this paper, we consider an RNA-seq data set from maize that includes eight RNA samples from each of two inbred lines (B73 and Mo17) and their hybrids (B73 × Mo17 and Mo17 × B73) formed by reciprocal crosses where the male and female parental genotypes are reversed. Throughout the remainder of the paper, we use BB, MM, BM, and MB as abbreviations for these four genotypes. From each genotype, RNA samples were drawn from each of four different plants subjected to drought stress conditions and from four other plants grown under control conditions. Plants were grown and processed in four blocks, with each combination of treatment and genotype represented in each block.

Although all samples were sequenced simultaneously, the manner in which they were prepared and arranged for sequencing added additional structure to the data that should be accounted for in modeling and analysis. All 32 RNA samples (4 blocks × 4 genotypes × 2 treatments) were sequenced in the eight lanes of a single Illumina flowcell. (See Nettleton (2014) for a general introduction to sequencing on flowcells from a statistical perspective.) The BB, MM, BM, and MB RNA samples corresponding to any single block and treatment combination were sequenced together in a single lane. Each sample within each lane was associated with a different identifying “barcode” so that each sequenced RNA fragment (known as a read) could be attributed to the sample from which it originated. The concept of a Latin square was used to match barcodes with genotypes within each block. The layout of the sequencing design is depicted in Table 1, where C and D are used to designate the control and drought treatment conditions.

Based on the layout in Table 1, the experiment has a structure similar to that of a split-plot design. The whole-plot portion of the experiment is arranged as a randomized complete block design with four blocks, lane as the whole-plot experimental unit, and treatment (C vs. D) as the whole-plot factor. Genotype (BB, MM, BM, or MB) is the split-plot factor, and barcode is an additional blocking factor whose effects, though not expected to be large, will be accounted for in our modeling and analysis.

For each of the 32 samples represented by a cell in Table 1, a read count associated with RNA abundance for each of 29, 985 genetic features was derived from sequencing. The number of bases that compose each feature (length) and the proportion of the bases that are guanine or cytosine (GC content) of each feature were recorded. Our primary objective is to build a model for these count data and use Bayesian methods to identify Differentially expressed features via INLA and *ShrinkBayes*.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

3. METHODS

3.1 Model

For each $i = 1, \ldots, m = 29,985$ and each $j = 1, \ldots, n = 32$, let Y_{ij} denote the observed read count for genetic feature i and experimental unit j, and let LL_i and GC_i be the log length and GC content of feature i, respectively. We consider a generalized linear mixed-effects model for the read count data. Such models are inherently hierarchical. At the data level of the hierarchy, we assume

$$Y_{ij} \sim \text{Negative Binomial}(e^{\eta_{ij}}, e^{\nu_i}),$$

where $E(Y_{ij}) = e^{\eta ij}$ and $\text{Var}(Y_{ij}) = E(Y_{ij}) + e^{\nu i}\{E(Y_{ij})\}^2$. Conditional on all η_{ij} and ν_i values, all the Y_{ij} counts are assumed to be mutually independent. At the next level of the hierarchy, we assume η_{ij} is a linear combination of feature-specific fixed effects (contained in a vector $\boldsymbol{\beta}_i$), feature-specific random effects (contained in a vector $\boldsymbol{u}_i$), a smooth function (h) of feature length and GC content, and a sample-specific normalization factor (T_j) given by

$$\eta_{ij} = \boldsymbol{x}_j' \boldsymbol{\beta}_i + \boldsymbol{z}_j' \boldsymbol{u}_i + h(LL_i, GC_i) + T_j. \quad (1)$$

The terms $h(LL_i, GC_i)$ and T_j are offsets included for normalization purposes as described in Section 3.3. The other terms in equation (1) are defined as follows.

For $k = 1, \ldots, 8$, the kth component of $\boldsymbol{\beta}_i$ (β_{ik}) is a fixed effect for the kth combination of treatment and genotype as indicated in Table 2. If the experimental unit j is associated with the kth combination of treatment and genotype, then $\boldsymbol{x}_j'$ is the kth row of the 8×8 identity matrix ($\boldsymbol{I}_{8\times8}$) so that $\boldsymbol{x}_j' \boldsymbol{\beta}_i = \beta_{ik}$. The feature-specific vector of random effects $\boldsymbol{u}_i$ contains eight random effects for lanes, four random effects for blocks, and four random effects for barcodes and is assumed to follow a multivariate Gaussian distribution with mean $\mathbf{0}$ and diagonal variance with blocks $\sigma^2_{Li} \boldsymbol{I}_{8\times8}$, $\sigma^2_{BLi} \boldsymbol{I}_{4\times4}$, and $\sigma^2_{BCi} \boldsymbol{I}_{4\times4}$. The vector $\boldsymbol{z}_j$ is a vector of length 16 indicating the lane, block, and barcode of experimental unit j. For example,

$$\boldsymbol{z}_1' = [\ 1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 1\ \ 0\ \ 0\ \ 0\ \ 1\ \ 0\ \ 0\ \ 0\ ]$$

signifies that experimental unit 1 was sequenced in lane 1, was in block 1, and was associated with barcode 1.

At the final stage of our hierarchical model are priors for the feature-specific parameters:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\begin{aligned}
\nu_1, \ldots, \nu_m &\overset{iid}{\sim} N(\mu_\nu, \sigma^2_\nu), \\
\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m &\overset{iid}{\sim} N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \textstyle\sum_{\boldsymbol{\beta}}), \\
\sigma^{-2}_{L1}, \ldots, \sigma^{-2}_{Lm} &\overset{iid}{\sim} \text{Gamma}(\omega_L, \phi_L), \\
\sigma^{-2}_{BL1}, \ldots, \sigma^{-2}_{BLm} &\overset{iid}{\sim} \text{Gamma}(1, 10^{-5}), \text{ and} \\
\sigma^{-2}_{BC1}, \ldots, \sigma^{-2}_{BCm} &\overset{iid}{\sim} \text{Gamma}(1, 10^{-5}).
\end{aligned}$$

The unspecified hyperparameters $\mu_\nu, \sigma^2_\nu, \boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}, \omega_L$, and φ_L, which we represent collectively by $\boldsymbol{\varpi}$, are estimated from the data through the empirical Bayes procedure described in Section 3.2. We specify relatively diffuse priors for the precisions of the blocking factors (block and barcode), but we choose to estimate the parameters of the prior for the lane variance components because, as the whole-plot experimental units, lanes play an important role in inferences involving the whole-plot treatment factor.

3.2 INLA and *ShrinkBayes*

INLA is an alternative to Markov chain Monte Carlo methods for latent Gaussian models, with the advantage of greater computational speed without sacrificing accuracy. INLA provides a deterministic approximation to marginal posterior distributions, as well as an approximation of the marginal likelihood. Because it is common in RNA-seq analyses to assign a negative binomial likelihood to the observed counts, and to model some function of the mean using an additive linear predictor, we can readily apply INLA to RNA-seq data by assigning Gaussian priors to the coefficients in our linear predictors.

The methods introduced by van de Wiel et al. (2012), and implemented in the *R* package *ShrinkBayes*, utilize INLA to facilitate an empirical-Bayes-type analysis of RNA-seq data, making use of the high dimensionality of the data to shrink both dispersion and regression parameter estimates. *ShrinkBayes* aims to allow for flexibility in the count model and in experimental design, while facilitating shrinkage of multiple parameters and addressing multiple testing. We achieve shrinkage of the parameters of interest by estimating the hyperparameters of the distributions according to the following paradigm.

As an example, consider estimation of $\varpi_\nu = (\mu_\nu, \sigma^2_\nu)$, the hyperparameters of the Gaussian prior for $\nu_1, \ldots, \nu_m$. For simplicity, initially suppose the hyperparameters in $\boldsymbol{\varpi}$ other than $\boldsymbol{\varpi}_\nu$ are known. Let $\mathbf{Y}_i$ be a vector containing the counts for genetic feature i, with distribution $F_{\boldsymbol{\varpi}}(\mathbf{Y}_i)$ defined by the model in Section 3.1. We can express the Gaussian prior for $\nu_1, \ldots, \nu_m$ as

$$\pi_{\varpi_\nu}(\nu) = \int \pi_\varpi(\nu|\mathbf{y}) \mathrm{d}F_\varpi(\mathbf{y}), \quad (2)$$

where $\pi_{\boldsymbol{\varpi}}(\nu|\mathbf{Y}_i)$ is the posterior of ν_i given $\mathbf{Y}_i$. Assuming $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ are draws from the distribution $F_{\boldsymbol{\varpi}}$, the above integral can be approximated by $\frac{1}{m}\sum_{i=1}^{m} \pi_\varpi(\nu|\mathbf{Y}_i)$. van de Wiel et al. (2012) showed that finding $\boldsymbol{\varpi}_\nu$ such that

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\pi_{\varpi_\nu}(\nu) \approx \frac{1}{m}\sum_{i=1}^{m}\pi_{\varpi}(\nu|\mathbf{Y}_i)=\pi_{\varpi}^{Emp}(\nu)$$

is approximately equivalent to the conventional empirical Bayes approach of choosing hyperparameters that maximize the marginal likelihood. *ShrinkBayes* finds such an $\boldsymbol{\varpi}_\nu$ through an iterative algorithm, first using initial values for $\boldsymbol{\varpi}_\nu$ to approximate $\pi_{\boldsymbol{\varpi}}(\nu|\mathbf{Y}_1)$, …, $\pi_{\boldsymbol{\varpi}}(\nu|\mathbf{Y}_m)$ via INLA, then drawing a large sample from the distribution defined by $\pi_{\varpi}^{Emp}(\nu)$, finding the value of $\boldsymbol{\varpi}_\nu$ that maximizes the likelihood of the sample according to $\pi_{\boldsymbol{\varpi}_\nu}$, and repeating until convergence. In practice, all the elements of $\boldsymbol{\varpi}$ are unknown, and the remaining elements of $\boldsymbol{\varpi}$ are estimated concurrently using an analagous approach. See van de Wiel et al. (2012) for further details on updating the estimate of $\boldsymbol{\varpi}$, theoretical properties of the iterative procedure, simultaneous shrinkage of parameters, and other features of *ShrinkBayes*. Upon convergence, INLA is again used to approximate marginal posterior distributions of interest for use in testing, which is explained in detail in Section 4.

For the maize data discused in Section 2, our interest is in identifying genetic features that substantially change expression level across combinations of treatment and genotype. In the context of the model specified in Section 3.1, we seek features for which $|\mathbf{c}'\boldsymbol{\beta}_i|$ is large for some contrast vector $\mathbf{c}$ that defines a comparison of interest. As an example, with $\mathbf{c}' = [1, -1, 0, 0, 0, 0, 0, 0]$, the magnitude of $\mathbf{c}'\boldsymbol{\beta}_i = \beta_{i1} - \beta_{i2}$ measures the extent of Differential expression for between the parental genotypes BB and MM under control conditions for the ith feature. A contrast like $\beta_{i1} - \beta_{i2}$ is often referred to as a log "fold change" because it represents a log ratio of means, appropriately adjusted for random effects and normalization factors. In addition to approximating marginal posteriors for individual feature-specific parameters, *ShrinkBayes* is able to estimate marginal posteriors for linear combinations of feature-specific parameters, including log fold changes and differences in log fold changes. This allows estimation and inference for a variety of contrasts that may be of interest. In Section 4, we show how to use the marginal posteriors estimated by *ShrinkBayes* to draw conclusions about three specific example contrasts in an analysis of the maize data.

3.3 Normalization

Normalization can account for differences in the total number of reads per sample and RNA composition of samples, and has been shown to be necessary for comparison across samples (Dillies et al., 2013; Robinson, Oshlack et al., 2010). Furthermore, biases introduced by the GC content and length of each feature have been well documented, but are not typically consistent across data sets (Oshlack et al., 2009; Benjamini and Speed, 2012). A common approach to normalization is including an offset in the linear predictor, as we have done in Section 3.1 by use of the $h(LL_i, GC_i)$ and T_j terms. We use the log of the trimmed mean of M values (TMM) for T_j to normalize between samples (Robinson, Oshlack et al., 2010). However, we also include a gene-specific term $h(LL_i, GC_i)$. Using the counts from all experimental units, we fit a smoothing spline to response log(count+1), with GC content and log feature length as explanatory variables, using the *mgcv* package in *R*. Some characteristics of the estimated function, displayed in Figure 1, show the nontrivial

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

relationship that exists between read count abundance and the length and GC content of genetic features. For each feature, the fitted value of the estimated function at the feature's GC content and length is included in the linear predictor as $h(LL_i, GC_i)$ in equation (1).

3.4 Prior Specification for $\boldsymbol{\beta}_i$

In Section 3.1, we assumed $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m \overset{iid}{\sim} N(\boldsymbol{\mu}_\beta, \sum_\beta)$. Riebler et al. (2014) described techniques for using *ShrinkBayes* to estimate joint priors, but estimation of an unstructured 8 × 8 covariance matrix Σ_β is currently intractable using these techniques. One natural simplification would be to assume the Σ_β is diagonal and to proceed with empirical Bayes hyperparameter estimation and approximate posterior inference under independent priors. We executed that strategy for the maize data using *ShrinkBayes* to obtain (for all $i = 1, \ldots, m$ and $k = 1, \ldots, 8$) a posterior median $\hat{\beta_{ik}}$ for β_{ik}. Figure 2 shows a scatterplot of the points $\{(\hat{\beta_{ik}}, \hat{\beta_{ik^*}}): i = 1, \ldots, m\}$ for each $k < k^*$ with $k, k^* \in \{1, \ldots, 8\}$.

All the scatterplots show strong correlations between posterior medians. Although correlation between posterior medians does not, in general, imply a need for dependent priors, we would expect much less correlation in the scatterplots if Σ_β were truly diagonal. Instead, the scatterplots are consistent with the idea that variation in expression level across genetic features is a dominant source of variation in transcript abundance levels as measured by read counts. Lund et al. (2012) discussed this phenomenon for microarry-based measures of transcript abundance. In the maize RNA-seq data, some genetic features have many thousands of reads across all eight combinations of treatment and genotype. Other genetic features tend to have single-digit read counts regardless of treatment and genotype. Variations in expression level within genetic feature are often relatively small compared to differences in expression level across genetic features, even after accounting for variations due to gene length and GC content as discussed in Section 3.3. This suggests that Σ_β should have relatively large diagonal elements and positive off-diagonal elements that are non-negligible in magnitude.

To estimate the hyperparameters in Σ_β in a more suitable way, we consider a reparameterization. Let the spectral decomposition of Σ_β be $\Sigma_\beta = \boldsymbol{Q}\Lambda\boldsymbol{Q}'$, where $\boldsymbol{Q}$ is an orthogonal 8 × 8 matrix and Λ is a diagonal 8 × 8 matrix. Then $\boldsymbol{Q}'\boldsymbol{\beta}_i$ has mean $\boldsymbol{Q}'\nu_\beta$ and diagonal variance Λ. Because Σ_β is unknown, we use $\hat{\Sigma_\beta}$, defined as the sample variance-covariance matrix of $\hat{\boldsymbol{\beta}_1}, \ldots, \hat{\boldsymbol{\beta}_m}$ from Figure 2, as an empirical approximation of Σ_β. We then compute the spectral decomposition $\hat{\Sigma_\beta} = \boldsymbol{Q}\hat{\Lambda}\hat{\boldsymbol{Q}}'$, and define a new parameter $\boldsymbol{\theta}_i = \boldsymbol{Q}'\hat{\boldsymbol{\beta}}_i$ for all $i = 1, \ldots, m$. We can readily use *ShrinkBayes* to estimate hyperparameters and perform posterior inference for the maize data by specifying

$$\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m \overset{iid}{\sim} N(\boldsymbol{\mu}_\theta, \sum_\theta),$$

where Σ_θ is a positive definite, diagonal matrix. The implied prior for $\boldsymbol{\beta}_i = \boldsymbol{Q}\hat{\boldsymbol{\theta}}_i$ is then multivariate Gaussian with mean $\boldsymbol{\mu}_\beta = \boldsymbol{Q}\hat{\boldsymbol{\mu}}_\theta$ and variance $\Sigma_\beta = \boldsymbol{Q}\hat{\Sigma}_\theta\boldsymbol{Q}'$, a non-diagonal positive definite matrix. Whereas model (1) has a single component of $\boldsymbol{\beta}_i$ for each treatment

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and genotype, the elements of θ_i in the alternative parameterization are orthogonal linear combinations of the β_i parameter vectors. For the given data,

$$\hat{Q}' = \begin{bmatrix} 0.382 & 0.346 & 0.350 & 0.346 & 0.374 & 0.344 & 0.341 & 0.343 \\ 0.498 & -0.504 & -0.012 & -0.017 & 0.489 & -0.507 & -0.026 & -0.016 \\ -0.361 & -0.351 & 0.236 & 0.377 & -0.303 & -0.350 & 0.406 & 0.411 \\ 0.330 & 0.197 & 0.412 & 0.431 & -0.412 & -0.264 & -0.408 & -0.301 \\ -0.515 & 0.107 & -0.067 & 0.454 & 0.535 & -0.087 & -0.467 & 0.042 \\ -0.304 & 0.111 & 0.484 & -0.178 & 0.266 & -0.157 & 0.396 & -0.615 \\ 0.072 & 0.571 & -0.523 & 0.131 & -0.015 & -0.518 & 0.329 & -0.045 \\ 0.075 & -0.339 & -0.373 & 0.545 & -0.041 & 0.363 & 0.264 & -0.490 \end{bmatrix}$$

Note that, as defined by the loadings in the first row of $\hat{Q}'$, θ_{i1} is approximately a constant times the average of the elements of β_i and, hence, is proportional to a general log expression level for gene i. Likewise, for gene i on the log scale, θ_{i2} corresponds roughly to the difference between the parents (BB minus MM) averaged over treatments, θ_{i3} may be interpreted as an approximate difference between hybrids and parents averaged over treatments, and θ_{i4} approximates the difference between treatments averaged over genotypes. According to the corresponding eigenvalues, the first linear combination accounts for 94.3% of the total variance in $\hat{\Sigma}_{\hat{\beta}}$, and the first four linear combinations together account for over 99.5% of the total variance in $\hat{\Sigma}_{\hat{\beta}}$. Figure 3 shows the analog of Figure 2 for the alternative parameterization and prior specification. The scatterplots of posterior medians $\hat{\theta}_1, \ldots, \hat{\theta}_m$ show very little correlation, indicating that the use of independent priors for the elements of θ_i may be considerably more reasonable than using independent priors for the elements of β_i.

As another benefit of reparameterization, note that the "V" pattern of the $\theta_3 \times \theta_2$ scatterplot in Figure 3 clearly differs from the remaining plots, and points us towards a possible set of DE genes where the expression level of one parent may differ from a common level of expression shared by the other parent and the hybrids. Since genes with large $|\theta_2|$ have a large difference between parents, and genes with large θ_3 have hybrids expressed more highly than parents, on average, the genes found at the top of the "V" may consist of one parent with low expression and one parent and both hybrids with high expression. Although this plot may miss features whose expression patterns differ across treatments, the intersection of genes with large $|\theta_2|$ and genes with large θ_3 may contain many features of interest.

4. ESTIMATION AND TESTING

Estimates of ϖ under both the original and alternative parameterization are reported in Table 3.

After estimating ϖ, we are able to approximate the marginal posterior distribution for each parameter and any desired linear combinations of parameters. To demonstrate testing for differential expression, we consider the three comparisons defined in Table 4, representing simple effects, main effects, and interactions, respectively. The simple effect T_1 represents a

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

log fold change between the two parents under control conditions. The main effect T_2 examines the log fold change between treatments averaged over all four genotypes. The interaction effect T_3 represents the change, across treatments, in the log fold change between hybrids. T_1, T_2, and T_3 can be viewed as tests involving the split-plot factor, the whole-plot factor, and split-plot factor by whole-plot factor interaction, respectively. Although Table 4 lists each test in terms of $\beta_1, \ldots, \beta_8$, getting each contrast $\boldsymbol{c}'\boldsymbol{\beta}_i$ in terms of the alternative parameterization $\theta_1, \ldots, \theta_8$ is straightforward, since $\boldsymbol{c}'\boldsymbol{\beta}_i = \boldsymbol{c}'(\boldsymbol{Q}\hat{\boldsymbol{\theta}}_i) = (\boldsymbol{c}'\boldsymbol{Q})\hat{\boldsymbol{\theta}}_i$.

The marginal posterior distributions of the linear combinations T_1, T_2, and T_3 were approximated for each feature using *ShrinkBayes* and the model defined in Section 3.1 with the alternative parameterization discussed in Section 3.4. Posterior medians were computed to serve as point estimates of T_1, T_2, and T_3. In addition to point estimates, we also calculated the posterior probability of Differential expression for each feature and each linear combination. As an example, we define a feature to be DE for T_1 if $|T_1| \geq \log(1.25)$ for that feature. This definition of Differential expression corresponds to an increase of at least 25% in the expression level of one parent relative to the other. The threshold 1.25 is an arbitrary choice that we have made here simply for the sake of illustration. Depending on the goals of an investigator, smaller or larger thresholds could be selected. Based on the 1.25 threshold, the null hypothesis of equivalent expression is then H_0: $|T_1| < \log(1.25)$. van de Wiel et al. (2012) recommended using a conservative adjustment to the posterior probability of the null hypothesis, $P(H_0|\mathbf{Y})$, given by

$$P^{II}(H_0|\boldsymbol{Y}) = \min\{P(T_1 < \log(1.25)|\boldsymbol{Y}), P(T_1 > -\log(1.25)|\boldsymbol{Y})\}$$

to avoid the case of an extremely vague posterior returning a small posterior probability of the null hypothesis. We denote this conservative estimate of posterior probability of equivalent expression, $P^{II}(H_0|\mathbf{Y})$, as the local false discovery rate, *lfdr*. The posterior probability of Differential expression, $1 - P^{II}(H_0|\mathbf{Y}) = 1 - lfdr$, was calculated for every feature. This process was repeated with T_2 and T_3, using the same definition of Differential expression ($|T_2| \geq \log(1.25)$ and $|T_3| \geq \log(1.25)$). Figure 4 shows a plot of the posterior probability of Differential expression vs. posterior median for each linear combination, with vertical lines representing a 1.25-fold change in either direction.

van de Wiel et al. (2012) recommended the use of Bayesian false discovery rate (BFDR) to control the experiment-wise false discovery rate (Lewin et al., 2007; Ventrucci et al., 2010). Making use of the local false discovery rate for the ith feauture ($lfdr_i$), we define $BFDR_i$ to be the average of all *lfdr* values for features with *lfdr* less than or equal to $lfdr_i$. If we wish to maintain a 0.05 FDR, we simply declare all features with $BFDR \leq 0.05$ to be DE.

5. SIMULATION STUDY

To evaluate the properties of our approximated posterior probabilities and investigate the value of reparameterization in data similar to our motivating case, we conducted a sequence of brief simulation studies, differing only in how the expected counts and dispersion parameter of each genetic feature were determined. In each simulation, we generated data

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

from the negative binomial model, with a constant dispersion parameter within each genetic feature. For Simulation 1, we generated data using the model of Section 3.1 and the corresponding estimated hyperparameters from Section 4 as the true values. For Simulation 2, we took the posterior means for each parameter obtained in Section 4 as the truth. For Simulation 3, we used the estimated means and dispersions from a standard *edgeR* analysis of the maize data as the truth. The *edgeR* analysis used TMM normalization (Robinson, Oshlack et al., 2010), Cox-Reid profile-adjusted likelihood to estimate dispersion parameters (McCarthy et al., 2012), and treated block and barcode as fixed effects, but omitted lane. Including both lane and genotype by treatment effects would result in a rank defficient design matrix, because each lane contains samples from only one of the treatments. So the column of the design matrix corresponding to the effect of drought, for example, would be equal to the sum of the columns corresponding to the effects of the lanes containing drought-stressed samples (lanes two, four, six, and eight), and therefore the design matrix would not be of full column rank. Thus, for the given experimental design, lane cannot be modeled using fixed effects alongside genotype and treatment effects. Treating lane effects as random (as we have done in our model defined in Section 3.1) is not possible in the current version of *edgeR*. While Simulation 1 presents ideal conditions, with the model exactly matching the data generating mechanism, Simulations 2 and 3 represent progressively greater departures from our model in order to test the robustness of our methods.

In each setting, we simulated 10 data sets of identical dimensions and repeated the analysis of Section 4 under both the original parameterization and prior specification with diagonal Σ_β and the alternative parameterization and prior specification where Σ_β is non-diagonal. For each simulated data set, we estimated the smooth function $h(LL_i, GC_i)$ and calculated the TMM normalization factors from the simulated data in the same manner as before. Then, for each parameterization/prior specification, we estimated new hyperparameters based on the simulated data, and used the hyperparameters to compute *lfdr* and *BFDR* values for all features. We evaluated performance using two measures: empirically estimated FDR when setting the nominal FDR at 0.05, and the partial area under the receiver operating characteristic curve (pAUC) for false positive rate ranging from 0 to 0.1.

Figure 5 depicts the mean pAUC of the test of each contrast (T_1, T_2, and T_3) of interest under each simulation setting, accompanied by the corresponding standard error bars. For the first two simulation settings, we observe the ordering of genetic features from the analysis based on the alternative parameterization outperformed that of original parameterization. This relation does not hold for T_2 and T_3 in Simulation 3. The analogous plots of FDR in Figure 6 show a general tendency towards liberal testing under the original parameterization. However, under the alternative parameterization, we see adequate control of FDR, albeit erring towards lower than specified FDR, with the exception of T_1 under Simulation 3.

To illustrate why the original parameterization leads to liberal testing and does not permit control of FDR, we consider the implied priors from the observed data on T_2 under each parameterization. Using the left half of Table 3, it is straightforward to find that the implied prior on T_2 under the original parameterization is Gaussian with mean −0.013 and standard deviation 1.315. This corresponds to a prior probability of Differential expression of 0.865.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Under the alternative parameterization, however, the implied prior on T_2 is Gaussian with mean –0.011 and standard deviation 0.167. With a similar mean but much smaller standard deviation, the prior probability of Differential expression under the alternative parameterization is only 0.187. Given a prior probability of Differential expression almost five times greater for the original parameterization than for the alternative, it is not surprising to observe high false discovery rates for T_2 under the original parameterization, but accurate or low false discovery rates under the alternative parameterization.

While Simulation 3 is intended to represent a greater departure from our model than the first two settings, it may in fact represent a systematically difficult case for methods such as the alternative parameterization that effectuate significant shrinkage of parameters. Simulation 3 uses point estimates from an *edgeR* analysis to set true parameter values, but does not take into account the standard error of those point estimates. In negative binomial regression, maximum likelihood estimates of linear combinations of fixed effects (like those produced by *edgeR*) tend to have higher variances for low-count data. All else being equal, it is more likely that a higher variance point estimate will be far from zero. Therefore, many of the genes simulated in Simulation 3 as Differentially expressed are low-count genes. The analysis under the alternative parameterization shrinks the corresponding fold change estimates towards the prior mean, but under the original parameterization's more variable priors seen in Table 3, less shrinkage occurs. Since we also observe more shrinkage in low-count genes than in high-count genes, in a scenario such as Simulation 3 where many low-count genes are DE, the lack of borrowing information across genes in the original parameterization actually works as an advantage. For high-count genes in Simulation 3, performance under the alternative parameterization is similar, if not superior, to that of the original parameterization.

6. DISCUSSION

We have carried out an empirical-Bayes-type analysis of RNA-seq data in order to identify differentially expressed genetic features. The computational efficiency of INLA and the additional tools of *ShrinkBayes* make this possible to do quickly and without advanced programming by the user, while still providing uncommon levels of modeling flexibility. We discussed how careful parameterization can lead to more appropriate model specification, and also demonstrated a simple method to control for variation arising from GC content and feature length by estimating a smooth function and including the fitted value as an offset in the linear predictor. Finally, we demonstrated how to use the marginal posterior distributions computed by *ShrinkBayes* to test whether a feature is DE, and conducted a simple simulation experiment to show the importance of parameterization and that we can adequately control for FDR in a conservative manner, assuming a reasonable model specification.

The methods of *ShrinkBayes* allow for a fast Bayesian analysis of high-dimensional data via simplified functions and pre-compiled routines. While models commonly used for RNA-seq data readily fit into the INLA framework, INLA's requirement of a latent Gaussian field does somewhat limit modeling choices, and its inability to compute marginal posterior for nonlinear combinations of parameters limits the number of types of testable hypotheses. We

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

furthermore find that performance varies both under different tests and under departures from the model, and further work is required to increase the robustness of these methods.

Acknowledgments

Research reported in this chapter was supported by the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health and the joint National Science Foundation/NIGMS Mathematical Biology Program under award number R01GM109458. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

References

Anders S, Huber W. Differential expression analysis for sequence count data. Genome biol. 2010; 11(10):R106. [PubMed: 20979621]

Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic acids research. 2012:gks001.

Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in bioinformatics. 2013; 14(6):671–683. [PubMed: 22988256]

Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying Differential expression in sequence count data. BMC bioinformatics. 2010; 11(1):422. [PubMed: 20698981]

Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014; 15(2):R29. [PubMed: 24485249]

Lewin A, Bochkina N, Richardson S. Fully Bayesian mixture model for differential gene expression: simulations and model checks. Statistical applications in genetics and molecular biology. 2007; 6(1)

Li J, Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying Differential expression in RNA-Seq data. Statistical methods in medical research. 2013; 22(5):519–536. [PubMed: 22127579]

Lorenz, DJ.; Gill, RS.; Mitra, R.; Datta, S. Statistical Analysis of Next Generation Sequencing Data. Springer; 2014. Using RNA-seq Data to Detect Differentially Expressed Genes; p. 25-49.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15(12):550. [PubMed: 25516281]

Lund SP, Nettleton D, et al. The importance of distinct modeling strategies for gene and gene-specific treatment effects in hierarchical models for microarray data. The Annals of Applied Statistics. 2012; 6(3):1118–1133.

McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic acids research. 2012:gks042.

Nettleton, D. Statistical Analysis of Next Generation Sequencing Data. Springer; 2014. Design of RNA Sequencing Experiments; p. 93-113.

Oshlack A, Wakefield MJ, et al. Transcript length bias in RNA-seq data confounds systems biology. Biol Direct. 2009; 4(1):14. [PubMed: 19371405]

Riebler, A.; Robinson, MD.; van de Wiel, MA. Statistical Analysis of Next Generation Sequencing Data. Springer; 2014. Analysis of Next Generation Sequencing Data Using Integrated Nested Laplace Approximation (INLA); p. 75-91.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for Differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26(1):139–140. [PubMed: 19910308]

Robinson MD, Oshlack A, et al. A scaling normalization method for Differential expression analysis of RNA-seq data. Genome Biol. 2010; 11(3):R25. [PubMed: 20196867]

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(2):319–392.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Smyth, GK. Bioinformatics and computational biology solutions using R and Bioconductor. Springer; 2005. Limma: linear models for microarray data; p. 397-420.

van de Wiel MA, Leday GG, Pardo L, Rue H, Van Der Vaart AW, Van Wieringen WN. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics. 2012:kxs031.

Ventrucci M, Scott EM, Cocchi D. Multiple testing on standardized mortality ratios: a Bayesian hierarchical model for FDR estimation. Biostatistics. 2010:kxq040.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

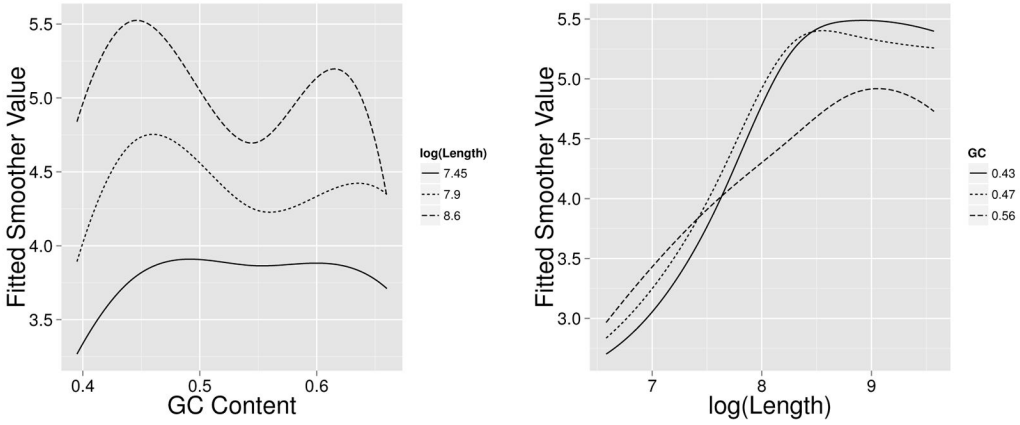

Figure 1.

Estimated mean log(count +1) as a function of GC content for selected log lengths (left), and as a function of log length for selected GC contents (right).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

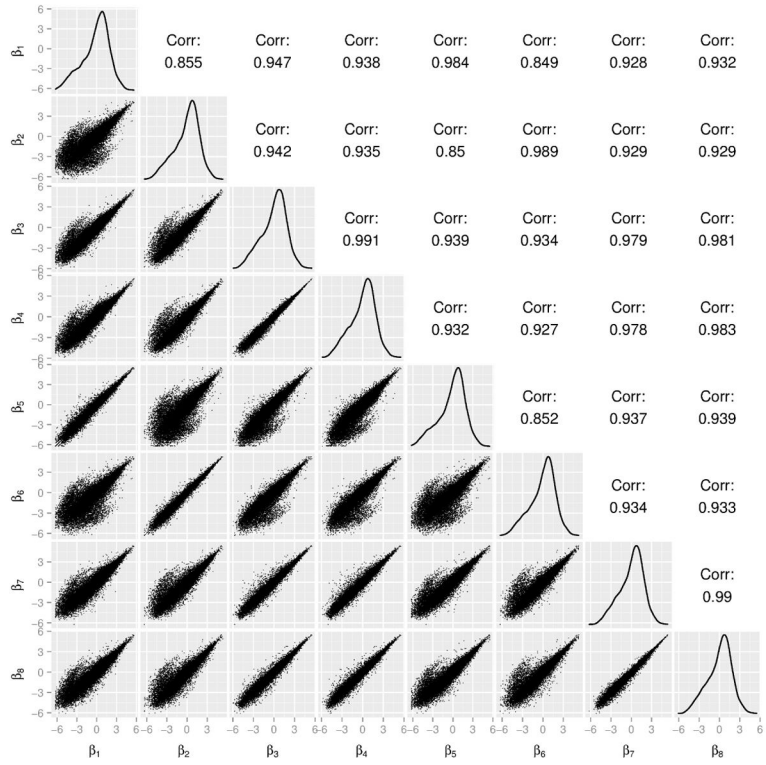

Figure 2.

Scatterplot matrix of posterior medians of each β_k (the parameter for the kth combination of genotype and treatment as defined in Table 2) for every gene when assuming a diagonal $\Sigma_{\boldsymbol{\beta}}$ under the original parameterization.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

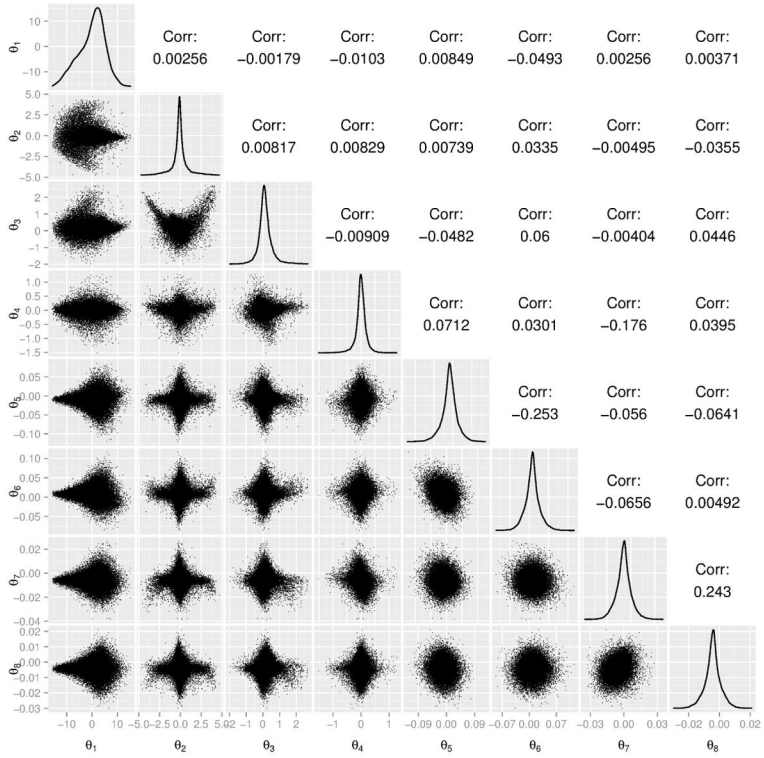

Figure 3.

Scatterplot matrix of posterior medians of each θ_k (the kth orthogonal combination of genotype-treatment parameters) for every gene using the alternative parameterization of Table 2. Note the reduced correlations and the "V" pattern of the $\theta_3 \times \theta_2$ cell.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

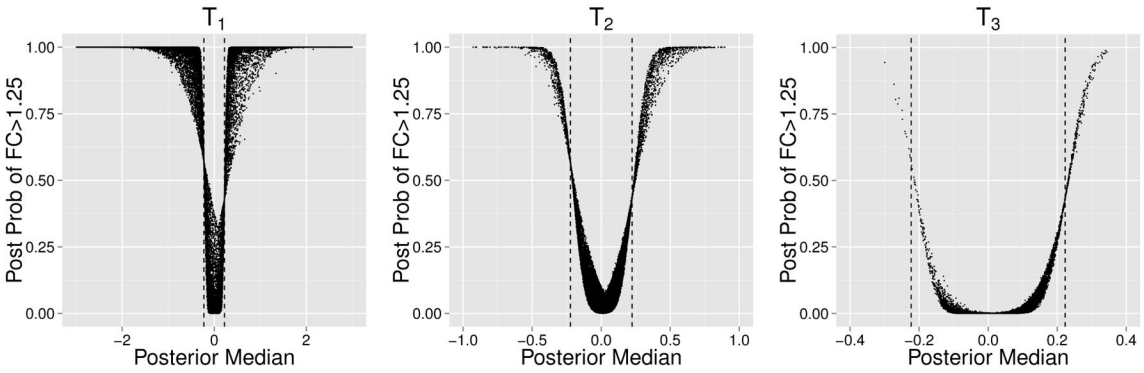

Figure 4.

Posterior probabilities of fold change greater than 1.25 against posterior medians for contrasts T_1, T_2, and T_3. We have little power for contrast T_3 and declare very few genes DE.

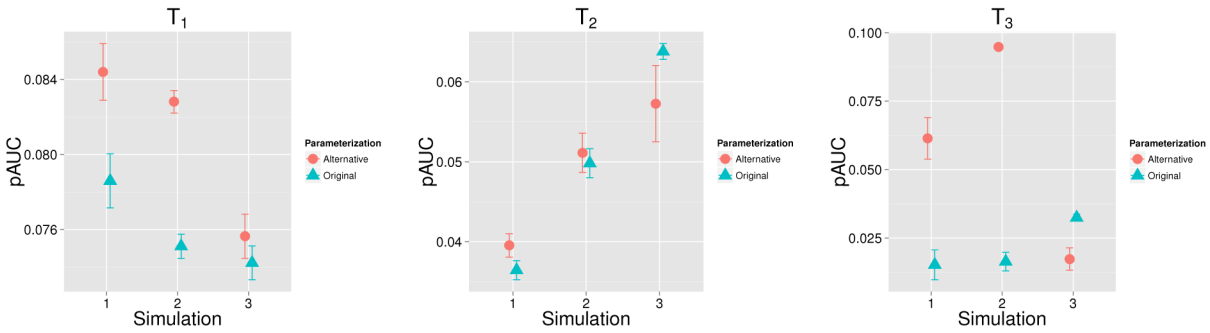

Figure 5.

Mean partial area under the ROC curve (pAUC) using *ShrinkBayes* over 10 simulated data sets for each of three contrasts (T_1, T_2, and T_3) in Simulations 1, 2, and 3.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

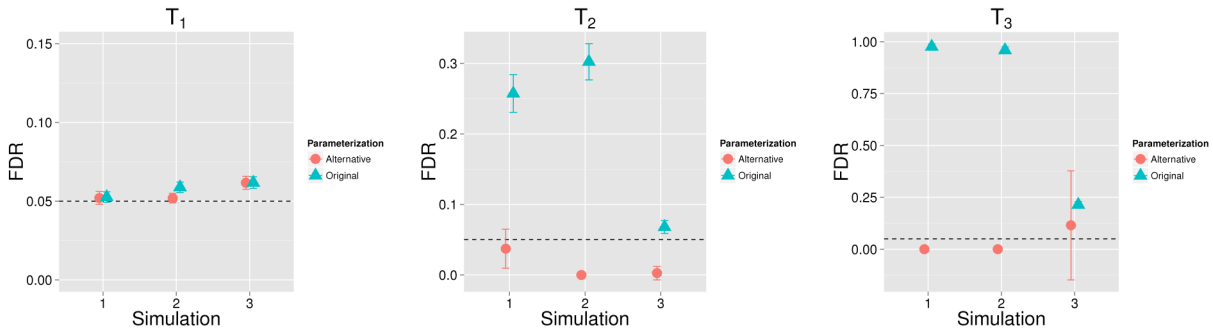

Figure 6.

Mean false discovery rates using *ShrinkBayes* over 10 simulated data sets while attempting to control FDR at .05 for each of three contrasts (T_1, T_2, and T_3) in Simulations 1, 2, and 3.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

RNA sequencing design

Lane	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Block	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
Treatment	**C**	**D**	**C**	**D**	**C**	**D**	**C**	**D**
Barcode 1	BB	BB	MM	MM	BM	BM	MB	MB
Barcode 2	MM	MM	BM	BM	MB	MB	BB	BB
Barcode 3	BM	BM	MB	MB	BB	BB	MM	MM
Barcode 4	MB	MB	BB	BB	MM	MM	BM	BM

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Model parameters for each treatment and genotype combination

Treatment	Genotype	Model (1) Parameter
C	BB	β_1
C	MM	β_2
C	BM	β_3
C	MB	β_4
D	BB	β_5
D	MM	β_6
D	BM	β_7
D	MB	β_8

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3

Hyperparameter estimates based on the original (left half) and alternative (right half) parameterizations

Parameter	Mean	Standard Deviation	Parameter	Mean	Standard Deviation
β_1	0.148	2.014	θ_1	0.200	3.633
β_2	0.049	1.856	θ_2	0.096	0.951
β_3	0.004	1.808	θ_3	0.136	0.436
β_4	0.001	1.795	θ_4	0.004	0.237
β_5	0.124	1.979	θ_5	0.009	0.042
β_6	0.046	1.847	θ_6	0.008	0.040
β_7	0.021	1.774	θ_7	0.006	0.024
β_8	0.002	1.785	θ_8	0.004	0.021
ν	3.952	0.820	ν	3.803	0.941
Parameter	Shape	Rate	Parameter	Shape	Rate
σ_L^{-2}	55.338	0.999	σ_L^{-2}	62.477	0.997

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4

Example comparisons of interest

Label	Comparison	Linear Combination
T_1	Control BB vs. Control MM Simple Effect	$\beta_1 - \beta_2$
T_2	Control vs. Drought Main Effect	$\frac{\beta_1+\beta_2+\beta_3+\beta_4}{4} - \frac{\beta_5+\beta_6+\beta_7+\beta_8}{4}$
T_3	Treatment × Hybrid Interaction	$\beta_3 - \beta_4 - \beta_7 + \beta_8$