Fully Bayesian analysis of allele-specific RNA-seq data using a hierarchical, overdispersed, count regression model

Ignacio Alvarez Jarad Niemi Dan Nettleton

Department of Statistics, Iowa State University

Allele-specific gene expression

Diploid orgnisms have two copies of each genes (alleles) that can be separately transcribed. The RNA abundance of any particular allele is known as allele-specific expression (ASE).

- In plant breeding, hybrids benefits from heterosis (hybrid vigor).
- ASE is relevant for the study of this phenomenon at the molecular level.

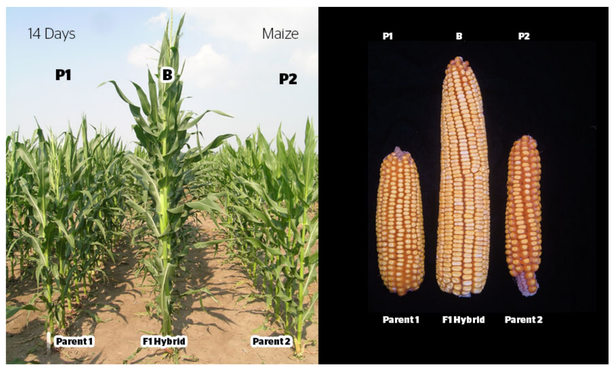
Goals of the study:

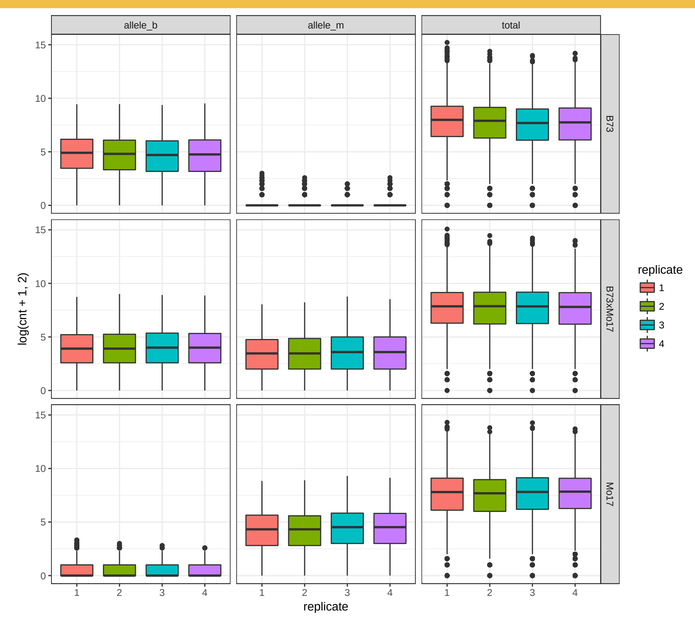
We present statistical methods for modelling ASE and detecting genes where differential allele expression. We propose a hierarchical overdispersed Poisson model to deal with ASE counts.

Maize experimental dataset

Dataset : Data from Paschold et al. (2012).

Design : Hybrid genotype (B73xMo17), 2 flow cell blocks, 4 replicate plants per block, 3 measures per plant





Poisson-lognormal hierarchical model

Data model

Y_{gn} : ASE count of gene g, in sub-sample n

$$\begin{aligned} Y_{gn} &\overset{\text{ind}}{\sim} Po(e^{h_n + x_n^\top \beta_g + \epsilon_{gn}}) \\ \epsilon_{gn} &\overset{\text{ind}}{\sim} N(0, \gamma_g) \end{aligned} \tag{1}$$

- h_n: normalization factor
- p regression coefficients ($p \geq 2$):
- ϵ_{gn}: overdispersion with gene-specific variance

Gene-specific layer

Regression coefficients $\beta_{gk} \overset{\text{ind}}{\sim} N(\theta_k,\ \sigma_k^2 \xi_{gk})\ \xi_{gk} \sim p(\eta)$

- Shrinkage distributions: Student-t, Laplace, horseshoe, normal

Overdispersion variances $\gamma_g \overset{\text{ind}}{\sim} IG(\frac{\nu}{2}, \frac{\nu\tau}{2})$

- γ_g are shrunk around τ, ν controls amount of shrinkage.

Allele effect: Δ_g

we set β_{g2} as the half allele difference

$$\Delta_g = \beta_{g2} - \theta_2.$$

Bias correction θ_2: systematic difference among alleles.

Differential expression $\{|\Delta_g| > c\}$

Credible interval Use posterior mean and variance for normal approximation.

$$\begin{aligned} E(\Delta_g|y) &= E(\beta_{g2}|y) - E(\theta_2|y) \\ Var(\Delta_g|y) &= Var(\beta_{g2}|y) + Var(\theta_2|y) - 2Cov(\beta_{g2}, \theta_2|y) \\ &\approx Var(\beta_{g2}|y) \end{aligned}$$

Full Bayesian inference using GPU

- Hyperparameter prior will not have a large impact in the gene-specific parameters (Ghosh et al. 2006)

$$\begin{aligned} \theta_k &\overset{\text{ind}}{\sim} \text{N}(0, c_k) \quad \nu \sim \text{Unif}(0, d) \\ \sigma_k &\overset{\text{ind}}{\sim} \text{Unif}(0, s_k) \quad \tau \sim Ga(a, b) \end{aligned}$$

Obtain fully Bayesian inference with `fbseq` package (Landau and Niemi 2016)

- Parallel MCMC algorithm using GPU
- 2 hours for single hybrid data, 10 when we include 3 varieties.
- `fbseq` output: posterior summaries for gene-specific parameters

Results from Simulation Study

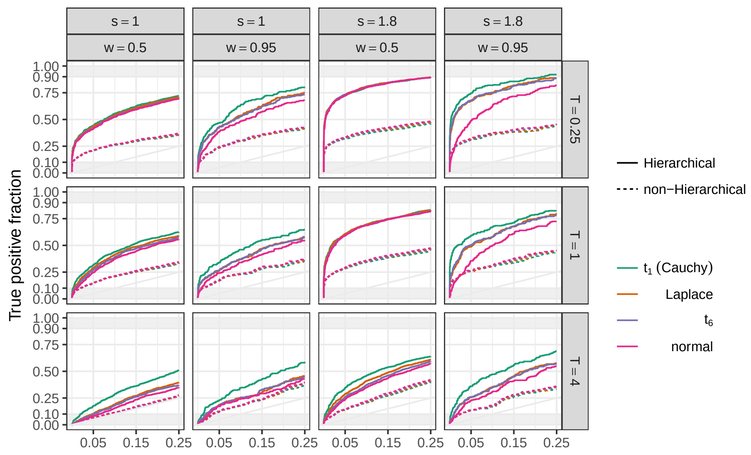
Simulate 24 scenarios

- Sparsity (w): proportion of genes with NO effect.
- Strengh (s): enlarge factor for DE genes
- Bias (p): proportion of non-reference allele lost due to bias
- Overdispersion (T): multiplicative factor of overdispersion

Table 1: Simulation study design parameter values

Description	Sparsity	Strength	Bias	Overdispersion
Parameter	w	s	p	T
Values	$(.5, .95)$	$(1, 1.8)$	$(1, .5)$	$(0.25, 1, 4)$

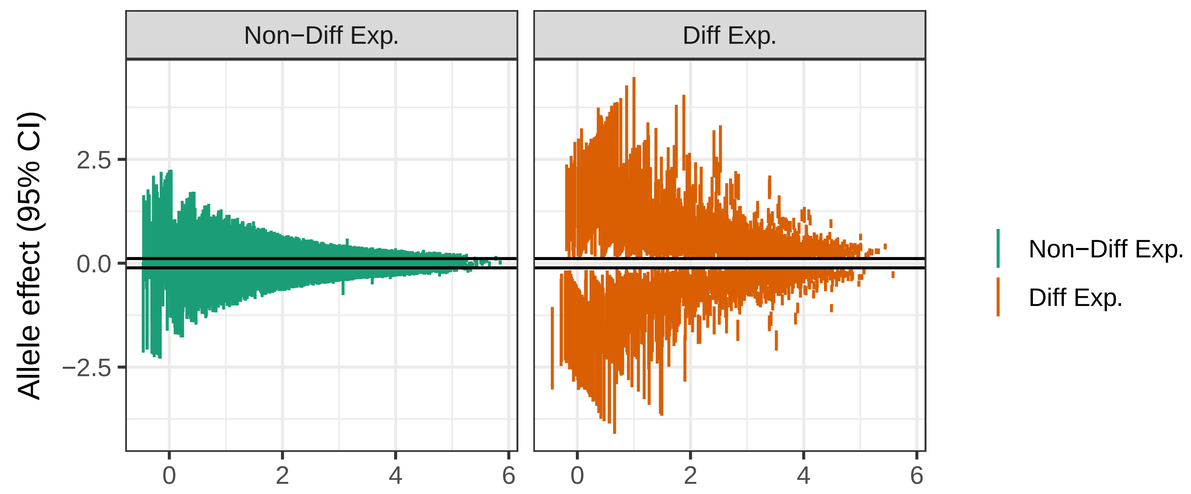
ROC curves from simulations



- non-hierarchical model does not capture bias
- Cauchy slightly better when $w = 0.95$ or $T = 4$

Bayesian analysis of maize data

- $\beta_{kg} \sim Cauchy(\theta_k, \sigma_k)$ is based on the results from simulation study.



- $|\Delta_g| > c$ in 17% of genes, $c = \log(1.25)/2$ (25% fold change)
- Higher expression = narrower CI
- Some DE genes with low expression
- Few genes with high overdispersion variances
- Bias: $E(\theta_2|y) = 0.126$, suggest 1 out 5 reads from Mo17 is lost
- results suggest $\sigma_4 \approx 10\sigma_5$

	mean	CI 95%
ν	3.6	(3, 4.3)
τ	0.0023	(0.0019, 0.0028)
θ_2	0.12	(0.12, 0.13)
σ_4^2	0.0011	(0.00094, 0.0012)
σ_5^2	0.000015	(0.0000095, 0.000023)