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

- In plant breeding, hybrids benefit 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 with 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

\[
Y_{gi} = \text{ASE count of gene } g, \text{ in sub-sample } i \\
Y_{gi} \sim \text{Poisson}(\theta_i + \xi(g, \beta_i)) \\
\xi(g, \beta_i) \sim \text{N}(0, \gamma_i)
\]

- \(h_i\): normalization factor
- \(\beta_i\): regression coefficients (\(p \geq 2\))
- \(\gamma_i\): overdispersion with gene-specific variance

Gene-specific layer

Regression coefficients \(\beta_g \sim \text{N}(\theta_g, \sigma_g^2 \xi(g, \beta_i) \xi(g, \beta_i) \sim p(\eta))\)
- Shrinking distributions: Student-t, Laplace, horseshoe, normal

Overdispersion variances \(\gamma_i \sim IG(\tau, \nu)\)
- \(\gamma_i\) are shrunk around \(\tau\), \(\nu\) controls amount of shrinkage.

Results from Simulation Study

Simulate 24 scenarios
- Sparsity (\(w\)): proportion of genes with NO effect.
- Strength (\(s\)): enlarge factor for DE genes
- Bias (\(p\)): proportion of non-reference allele lost due to bias
- Overdispersion (\(T\)): multiplicative factor of overdispersion

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(w)</td>
<td>Sparsity</td>
<td>(.5, .95)</td>
</tr>
<tr>
<td>(s)</td>
<td>Strength</td>
<td>(1.8)</td>
</tr>
<tr>
<td>(p)</td>
<td>Bias</td>
<td>(1.5)</td>
</tr>
<tr>
<td>(T)</td>
<td>Overdispersion</td>
<td>(0.25, 1, 4)</td>
</tr>
</tbody>
</table>

ROC curves from simulations


Bayesian analysis of maize data

\(\beta_{10} \sim \text{Cauchy}(\theta_{10}, \sigma_{10})\) is based on the results from simulation study.

- |\(\Delta_i > c\) 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: \(E(\beta_{10}|y) = 0.126\), suggest 1 out 5 reads from Mo17 is lost
- results suggest \(\sigma_{10} \approx 10\nu_{10}\)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\nu)</td>
<td>3.6</td>
</tr>
<tr>
<td>(\tau)</td>
<td>0.0023</td>
</tr>
<tr>
<td>(\theta_{10})</td>
<td>(3, 4.3)</td>
</tr>
<tr>
<td>(\sigma_{10})</td>
<td>(0.00094, 0.0012)</td>
</tr>
<tr>
<td>(\sigma^2)</td>
<td>(0.000015, 0.000023)</td>
</tr>
</tbody>
</table>

Non-Diff Exp. Diff Exp.