Detecting Differentially Expressed Genes with RNA-seq Data Using Backward Selection to Account for the Effects of Relevant Covariates

Yet Nguyen and Dan Nettleton, Department of Statistics, Iowa State University
Haibo Liu and Chris Tuggle, Department of Animal Science, Iowa State University

Introduction

RNA-seq datasets often contain several covariates in addition to the factor of primary scientific interest. Either ignoring relevant covariates or accounting for the effects of irrelevant covariates can result in low power for identifying DE genes. To address the challenge of identifying DE genes with RNA-seq datasets that include covariates, we propose a backward selection algorithm for selecting a subset of covariates whose effects are estimated and adjusted for when testing for differential expression.

Generalized Linear Model for RNA-seq Read Count Data

- Let y_{gi} be the read count for gene g from unit i $(g = 1, \ldots, m; i = 1, \ldots, n)$. $y_{gi} \sim \text{NB}(\mu_{gi}, \omega_g)$, i.e., $\text{Var}(y_{gi}) = \mu_{gi} + \omega_g \mu_{gi}^2$.
- Let $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}', \ldots, \boldsymbol{x}_{ik}')'$ denote a vector of known explanatory variable values for the ith unit.
- Letting $\mathcal{S}$ represent a subset of $\{1, \ldots, k\}$ that contains 1, we consider log-linear models of the form

$$\log(\mu_{gi}) = o_i + \beta_{g0|\mathcal{S}} + \sum_{j \in \mathcal{S}} \boldsymbol{x}_{ij}' \boldsymbol{\beta}_{gj|\mathcal{S}} \tag{1}$$

- Using `QuasiSeq` R package (Lund et al (2012)), we test $H_{0g1}^{\hat{\mathcal{S}}^*} : \boldsymbol{\beta}_{g1|\hat{\mathcal{S}}^*} = \mathbf{0}$ to declare DE genes.

The Proposed Backward Selection Algorithm

Let $\mathcal{S} \subset \{1, \ldots, k\}$. For any $j \in \mathcal{S}$, let $\boldsymbol{p}_{j|\mathcal{S}}$ denote the vector of m p-values obtained by testing $H_{0gj}^{\mathcal{S}} : \boldsymbol{\beta}_{gj|\mathcal{S}} = \mathbf{0}$ for $g = 1, \ldots, m$. Let $r(\boldsymbol{p}_{j|\mathcal{S}})$ be a measure of the relevance of $\boldsymbol{x}_j$ in model (1). Let $\mathcal{S}_1 = \{1, \ldots, k\}$:

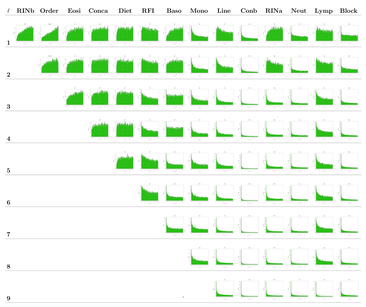
1. Compute $\boldsymbol{p}_{j|\mathcal{S}_\ell}$ for all $j \in \mathcal{S}_\ell$.
2. Let $\boldsymbol{q}_\ell$ be the vector of q-values obtained from $\boldsymbol{p}_{1|\mathcal{S}_\ell}$.
3. Let $R_\ell(\alpha)$ be the number of q-values in $\boldsymbol{q}_\ell$ less than or equal to a user-defined FDR threshold α.
4. If $\mathcal{S}_\ell \neq \{1\}$, find j^* so that $r(\boldsymbol{p}_{j^*|\mathcal{S}}) < r(\boldsymbol{p}_{j|\mathcal{S}})$ for all $j \in \mathcal{S}_\ell \setminus \{j^*\}$.
5. If $j^* = 1$, stop iterating. Otherwise, carry out the $\ell + 1$st iteration with $\mathcal{S}_{\ell+1} = \mathcal{S}_\ell \setminus \{j^*\}$.

Suppose the iterative procedure concludes after L iterations, and let ℓ^* be the smallest element of $\{1, \ldots, L\}$ such that $R_{\ell^*}(\alpha) \geq R_\ell(\alpha)$ for all $\ell \in \{1, \ldots, L\}$. We set $\hat{\mathcal{S}} = \mathcal{S}_{\ell^*}$ and base our inference about differential expression on the fit of model (1).

Analysis of RFI RNA-seq Dataset

The dataset contains RNA-seq count from 31 animals of 2 genetic lines which has 12280 genes with average count greater than 8. We test those genes for differential expression between two lines ($\boldsymbol{x}_{\cdot 1}$) in the presence of the thirteen categorical and continuous covariates: RFI ($\boldsymbol{x}_{\cdot 2}$, continous), Diet ($\boldsymbol{x}_{\cdot 3}$, categorical, 2 levels), Baso ($\boldsymbol{x}_{\cdot 4}$, continuous), Eosi ($\boldsymbol{x}_{\cdot 5}$, continuous), Lymp ($\boldsymbol{x}_{\cdot 6}$, continuous), Mono ($\boldsymbol{x}_{\cdot 7}$, continuous), Neut ($\boldsymbol{x}_{\cdot 8}$, continuous), Concb ($\boldsymbol{x}_{\cdot 9}$, continuous), Conca ($\boldsymbol{x}_{\cdot 10}$, continuous), RINb ($\boldsymbol{x}_{\cdot 11}$, continuous), RINa ($\boldsymbol{x}_{\cdot 12}$, continuous), Block ($\boldsymbol{x}_{\cdot 13}$, categorical, 4 levels), and Order ($\boldsymbol{x}_{\cdot 14}$, categorical, 4 levels). The best subset of relevant covariates is obtained at iteration $\ell = 7$, and number of DE genes is 448 when FDR is controlled at 5%.

RFI RNA-seq Data Analysis - pvalue05



RFI RNA-seq Data Analysis - ks

Simulation Study

1. Simulation 1: Using the set of covariates at iteration $\ell = 7$. $\mathcal{S}_\ell = \{1, 4, 5, 7, 8, 9, 12, 13\}$, the count data y_{gi} of gene g $(g = 1, \ldots, 12280)$ and pig i $(i = 1, \ldots, 31)$ has $\text{NB}(\mu_{gi}, \omega_g)$, where $\text{Var}(y_{gi}) = \mu_{gi} + \omega_g \mu_{gi}^2$, and

$$\log(\mu_{gi}) = o_i + \beta_{g0|\mathcal{S}_\ell} + \sum_{j \in \mathcal{S}_\ell} \boldsymbol{x}_{ij}' \boldsymbol{\beta}_{gj|\mathcal{S}_\ell}.$$

Suppose that

$$\hat{\omega}_{g|\mathcal{S}_\ell}, \hat{\boldsymbol{\beta}}_{g0|\mathcal{S}_\ell}, \hat{\boldsymbol{\beta}}_{gj|\mathcal{S}_\ell} \ (j \in \mathcal{S}_\ell) \text{ are the estimates of } \omega_g, \boldsymbol{\beta}_{g0|\mathcal{S}_\ell}, \boldsymbol{\beta}_{gj|\mathcal{S}_\ell},$$

respectively, for $g = 1, \ldots, 12280$. Let $\hat{m}_0$ be the estimate of m_0, the number of true null hypotheses among all $m = 12280$ hypotheses tested (for line effect, i.e., $\boldsymbol{x}_{\cdot 1}$). Let $\boldsymbol{q}_\ell = (q_{1|\mathcal{S}_\ell}, \ldots, q_{m|\mathcal{S}_\ell})$ be the vector of q-values obtained from testing $H_{0g1}^{\mathcal{S}_\ell} : \boldsymbol{\beta}_{g1|\mathcal{S}_\ell} = \mathbf{0}$, $g = 1, \ldots, m$. Suppose that the order statistic of components of $\boldsymbol{q}_\ell$ is $q_{(1)|\mathcal{S}_\ell} \leq \ldots \leq q_{(m)|\mathcal{S}_\ell}$. Set

$$\tilde{\boldsymbol{\beta}}_{g1|\mathcal{S}_\ell} = \begin{cases} \mathbf{0} & \text{if } g \in \mathcal{G}_{0|\mathcal{S}_\ell} := \{g : q_g \geq q_{(m-\hat{m}_0+1)|\mathcal{S}_\ell}\} \\ \hat{\boldsymbol{\beta}}_{g1|\mathcal{S}_\ell} & \text{if } g \in \mathcal{G}_{1|\mathcal{S}_\ell} := \{g : q_g < q_{(m-\hat{m}_0+1)|\mathcal{S}_\ell}\}. \end{cases}$$

$$\log(\tilde{\mu}_{gi|\mathcal{S}_\ell}) = o_i + \hat{\beta}_{g0|\mathcal{S}_\ell} + \boldsymbol{x}_{i1}' \tilde{\boldsymbol{\beta}}_{g1|\mathcal{S}_\ell} + \sum_{j \in \mathcal{S}_\ell \setminus \{1\}} \boldsymbol{x}_{ij}' \hat{\boldsymbol{\beta}}_{gj|\mathcal{S}_\ell}.$$

We simulate 100 datasets, each dataset consists of $m^* = 5000$ genes with different π, the proportion of EE genes, $\pi \in \{0.6, 0.7, 0.8, 0.9\}$. First, we randomly sample $\mathcal{G}_0^* = \{g_1, \ldots, g_{m^*\pi}\} \subset \mathcal{G}_{0|\mathcal{S}_\ell}$, $\mathcal{G}_1^* = \{g_{m^*\pi+1}, \ldots, g_{m^*}\} \subset \mathcal{G}_{1|\mathcal{S}_\ell}$. Finally, for each $h \in \{1, \ldots, m^*\}, i \in \{1, \ldots, 31\}$, we simulate $\tilde{y}_{hi}$ from $\text{NB}(\tilde{\mu}_{g_h i|\mathcal{S}_\ell}, \hat{\omega}_{g_h|\mathcal{S}_\ell})$. The collection of $\tilde{y}_{hi}$ is the simulated count data we need.

- Simulation 2: Using the Set of Covariates from Iteration $\ell \in \{6, 7, 8, 9\}$.
- Simulation 3: Using the set of Covariates from Iteration $\ell \in \{6, 7, 8, 9\}$ with modified RFI values ($\boldsymbol{x}_{\cdot 2}$).

Simulation Results

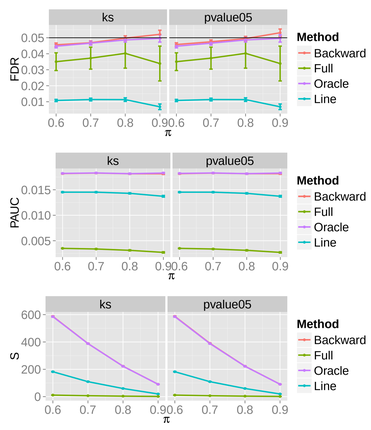


Figure 1: Simulation 1.

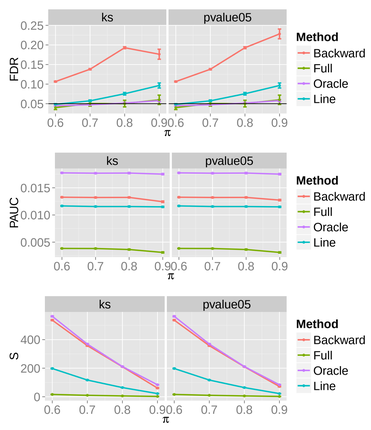


Figure 2: Simulation 2.

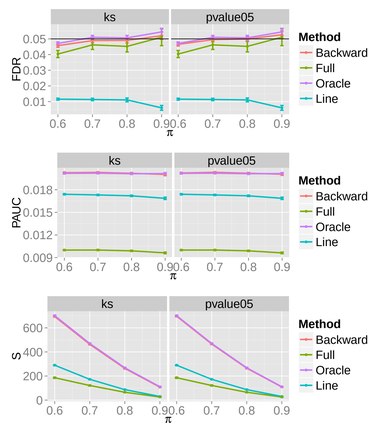


Figure 3: Simulaton 3.

Acknowledgements

This material is based upon work supported by Agriculture and Food Research Initiative Competitive Grant No. 2011-68004-30336 from the USDA National Institute of Food and Agriculture, and the National Science Foundation (NSF) under Grant No. 0922746. This research also was supported by 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 authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.