Differential Expression Analysis Methods Comparison

Xiyuan Sun

Iowa State University

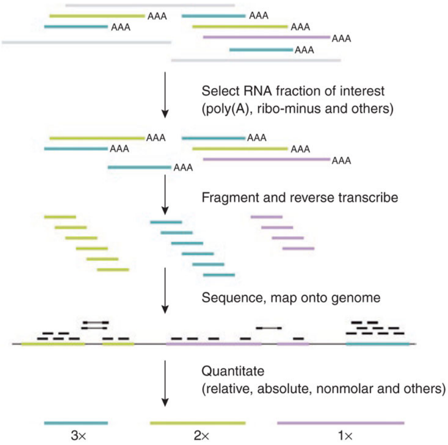
April 30, 2019

This research was built on Niemi et al's approach (Niemi et al., 2015). Their research was supported by National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health and joint National Science Foundation / NIGMS Mathematial Biology Program under award number R01GM109458.

Outline

- Background
  - RNA-seq procedure
  - RNA-seq data
  - Differential Expression (DE) Analysis
- Modeling
  - Negative Binomial Model in Generalized Linear Model
  - Hierarchical overdispersed count regression model
  - Null hypotheis for DE analysis
- Inference
  - Empirical Bayes
  - Alternative Methods
- Simulation studies
  - DE Genes detection via ROC curves
  - Area under ROC curve values

RNA fragmentation, sequencing, and alignment


AAA
AAA
AAA
AAA
AAA
AAA
Select RNA fraction of interest
(poly(A), ribo-minus and others)
Fragment and reverse transcribe
Sequence, map onto genome
Quantitate
(relative, absolute, nonmolar and others)
3×
2×
1×

(Pepke, Wold, and Mortazavi (2009) `http://www.nature.com/nmeth/journal/v6/n11s/fig_tab/nmeth.1371_F5.html`)

RNAseq data

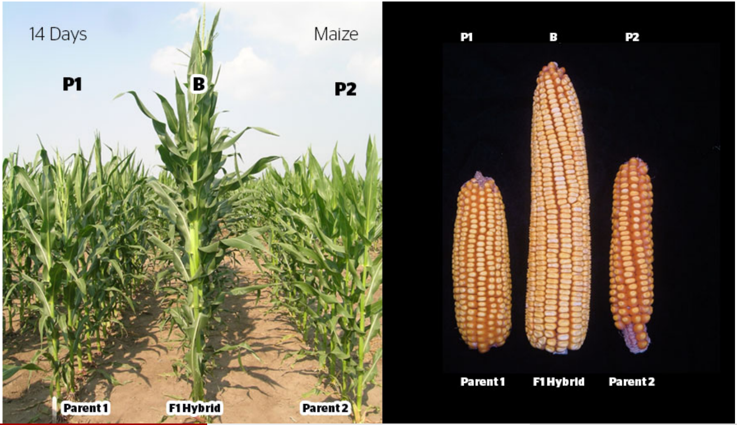
Genes	B73	B73	B73	B73	Mo17	Mo17	Mo17	Mo17
	Rep1	Rep2	Rep3	Rep4	Rep1	Rep2	Rep3	Rep4
AC148152.3_FG001	3	4	6	0	8	17	18	20
AC148152.3_FG008	3	3	4	1	31	40	45	49
AC152495.1_FG002	33	46	18	13	4	0	2	6
AC152495.1_FG017	41	44	16	13	2	2	2	0
AC184130.4_FG012	24	47	18	21	110	144	121	96
AC184133.3_FG001	0	1	1	0	14	13	4	9
AC148152.3_FG005	2323	1533	1932	1945	2070	1582	2196	1882
AC148167.6_FG001	672	598	728	713	743	655	821	824
AC149475.2_FG002	459	438	451	483	467	448	634	532
AC149475.2_FG003	1184	976	1131	1206	891	743	1288	1107
AC149475.2_FG005	551	535	360	353	550	524	492	440
AC149475.2_FG007	245	214	169	159	297	262	210	302

- DE genes: expression in Genotype Variety *B*73 is different from that in another Genotype Variety *Mo*17

Differential Expression Genes

Definition

A gene is regarded as differentially expressed (DE) when the expected count reads of this gene corresponding to one genotype variety differs from that of another genotype variety.



Differential expression analysis

Definition

For a given gene, we use statistical testing to decide whether an observed difference in read counts is significant, i.e., whether it is greater than what would be expected just due to natural random variation.

- Normalization
  Estimated normalization factors should ensure that a gene with the same expression level in two samples is not detected as DE.
- Assumed distribution
  Negative binomial
- Parameter estimation
  Mean, Dispersion
- Test for DE
  Exact test, Wald test, t-test

Negative Binomial Model in Generalized Linear Model Framework (Part 1)

Let

- g $(g = 1, \ldots, G)$ identify the gene,
- i $(i = 1, 2)$ identify the genotype variety,
- j $(j = 1, 2, 3, 4)$
- Y_{gij} be the RNAseq counts of gene g, genotype variety i, replicate j

We assume

$$Y_{gij} \overset{ind}{\sim} \mathrm{NB}\left(\mu_{gij}, \phi_g\right) \tag{1}$$

where

- μ_{gij} are means of read counts of gene g genotype i replicate j,
- ϕ_g allow for gene-specific overdispersion

Negative Binomial Model in Generalized Linear Model Framework (Part 2)

In the generalized linear model (GLM) setting, the mean response, μ_{gij} is linked to a linear predictor with natural log link:

$$log(\mu_{gij}) = x_i^T \beta_g + log(N_{ij}) \tag{2}$$

where

- x_i is row of the design matrix containing the covariates indicating this sample belongs to variety i,
- $\beta_g = (\beta_{g1}, \beta_{g2})$ is a vector of regression parameters
- N_{ij} is the normalized library size of replicate j in variety i

Hierarchical model for RNA-seq counts

We assume

$$Y_{gij} \overset{ind}{\sim} \text{NB}\left(\mu_{gij}, \phi_g\right)$$

where

- $\mu_{gij} = \exp(x_i^T \beta_g + \log(N_{ij}))$
- $\lambda_{gi} = x_i^T \beta_g, \gamma_{ij} = \log(N_{ij})$, then γ_{ij} are normalization factors
- $\phi_g = \exp(\psi_g)$ allow for gene-specific overdispersion

We reparameterized the mean dispersion structure into the genespecific average β_{g1} and half-variety difference β_{g2}

$$\beta_{g1} = \frac{\lambda_{g1} + \lambda_{g2}}{2}, \beta_{g2} = \frac{\lambda_{g1} - \lambda_{g2}}{2} \tag{3}$$

we also assume

$$\beta_{g1} \overset{ind}{\sim} \text{N}\left(\eta_{\beta_1}, \sigma^2_{\beta_1}\right), \beta_{g2} \overset{ind}{\sim} \text{N}\left(\eta_{\beta_2}, \sigma^2_{\beta_2}\right), \psi_g \overset{ind}{\sim} \text{N}\left(\eta_{\psi}, \sigma^2_{\psi}\right) \tag{4}$$

$\beta_{g1}, \beta_{g2}, \psi_g$ are independent to each other.

Empirical Bayes Method

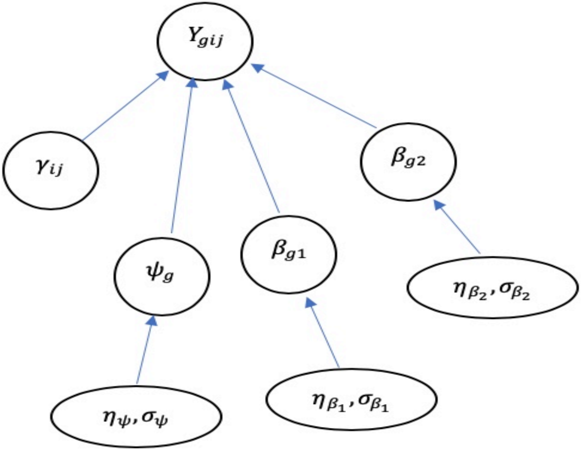
Let

- $\theta = (\theta_1, ..., \theta_G)$ $(g = 1, \ldots, G)$ where $\theta_g = (\beta_{g1}, \beta_{g2}, \psi_g)$,
- γ_{ij} is the normalized facor for replicate j in variety i,
- $\pi = (\eta, \sigma)$,where $\eta = (\eta_{\beta_1}, \eta_{\beta_2}, \eta_\psi), \sigma = (\sigma_{\beta_1}, \sigma_{\beta_2}, \sigma_\psi)$

Then,

- $\hat{\gamma}$ was obtained from trimmed mean of M values (TMM)
- $\hat{\psi}_g$ was got through the adjusted profile likelihood (APL)
- $\hat{\beta}_{g1}, \hat{\beta}_{g2}$ was retrieved by fitting the generalized linear model with log link function
- Using $\hat{\theta_g} = (\hat{\beta}_{g1}, \hat{\beta}_{g2}, \hat{\psi}_g)$, we estimated hyperparameters for the location and scale parameters in the hierarchical model using a central method of moments approach, as $\hat{\pi} = (\hat{\eta}, \hat{\sigma})$, where $\hat{\eta} = \sum_{g=1}^{G} \hat{\beta}/G, \hat{\sigma}^2 = \sum_{g=1}^{G} (\hat{\beta} - \hat{\eta})^2/(G-1)$

Empirical Bayes Method (cont)


Y_{gij}
γ_{ij}
β_{g2}
ψ_g
β_{g1}
$\eta_{\beta_2}, \sigma_{\beta_2}$
$\eta_{\psi}, \sigma_{\psi}$
$\eta_{\beta_1}, \sigma_{\beta_1}$

Empirical Bayes Method Posterior Probability of Parameters

Condition on the estimated normalization factor $\hat{\gamma}$ and hyperparameters $\hat{\pi}$, we perform a Bayesian analysis to re-estimate the θ as:

$$p(\theta|y,\hat{\pi},\hat{\gamma}) \propto \prod_{g=1}^{G}\prod_{i=1}^{2}\prod_{j=1}^{n_i} p(y_{gij}|\hat{\mu}_{gij},\hat{\phi}_g)p(\hat{\beta}_{g1}|\hat{\eta}_{\beta_1},\hat{\sigma}^2_{\beta_1})p(\hat{\beta}_{g2}|\hat{\eta}_{\beta_2},\hat{\sigma}_{\beta_2})p(\hat{\psi}_g|\hat{\eta}_{\psi},\hat{\sigma}^2_{\psi}) \tag{5}$$

where $\hat{\mu}_{gij} = \exp(\lambda_{gi} + \hat{\gamma}_{ij}), \hat{\phi}_g = \exp(\hat{\psi}_g)$, and
$\hat{\beta}_{g1} \overset{ind}{\sim} N(\hat{\eta}_{\beta_1},\hat{\sigma}^2_{\beta_1}), \hat{\beta}_{g2} \overset{ind}{\sim} N(\hat{\eta}_{\beta_2},\hat{\sigma}^2_{\beta_2}), \psi_g \overset{ind}{\sim} N(\hat{\eta}_{\psi},\hat{\sigma}^2_{\psi})$.

Null Hypothesis for DE Analysis

$$H_0 : \beta_{g2} = 0$$

which is equivalent to $\lambda_{g1} = \lambda_{g2}$
Statistics used to do the DE analysis is based on the posterior probabilities of β_{g2} as

$$P(DE_g | y, \hat{\pi}, \hat{\gamma}) = \min(P(\beta_{g2} < 0 | y, \hat{\pi}, \hat{\gamma}), P(\beta_{g2} > 0 | y, \hat{\pi}, \hat{\gamma}))$$

where $P(\beta_{g2} < 0 | y, \hat{\pi}, \hat{\gamma}) = \frac{1}{M} \sum_{m=1}^{M} I(\beta_{g2}^{(m)} < 0)$, and
$P(\beta_{g2} > 0 | y, \hat{\pi}, \hat{\gamma}) = \frac{1}{M} \sum_{m=1}^{M} I(\beta_{g2}^{(m)} > 0)$

Alternative Methods

Normalization
edgeR used gene-wise trimmed median of means (TMM), while DESeq, DESeq2, sSeq, EBSeq used sample-wise size factor.
Dispersion estimation
edgeR used Cox-Reid approximate conditional inderence (CRACI) moderate towards the mean while DESeq, DESeq2 used CRACI with focus on maximum individual dispersion estimate; sSeq estimated dispersion by pooling all the samples using the method of moments(MM), and then shrinking the gene-wise estimates through minimizing the mean-square error; EBSeq also estimated the gene-specific varainces via MM.
Test for DE
edgeR, sSeq used exact test for 2 factors; DESeq, DESeq2 used Wald test for 2 factors;

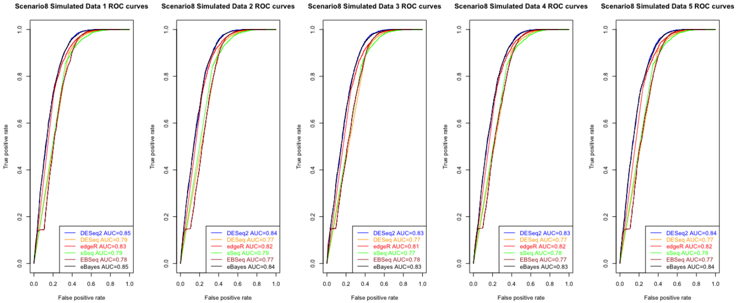
Simulation studies

Parameter estimation use `edgeR`: $\hat{\beta}_{g1}, \hat{\beta}_{g2}, \hat{\phi}_g$, and the normalized library sizes N_{ij}

Simulation scenario set up: nGenes, nSamples, pDiff

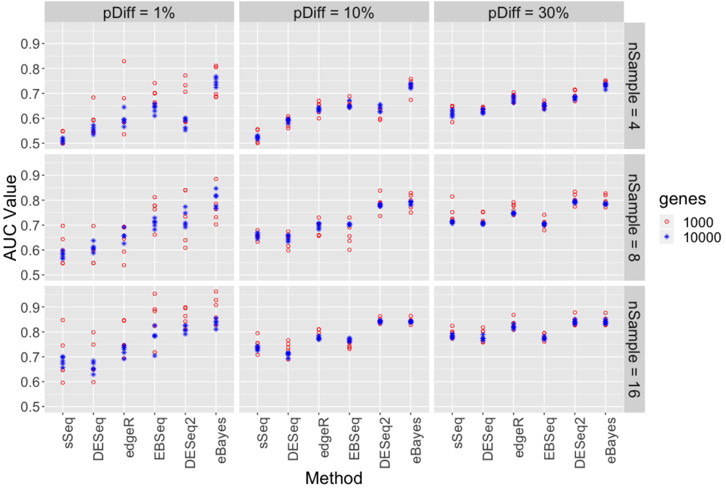
Simulation model: $Y_{gij} \overset{ind}{\sim} \text{NB}\left(\mu_{gij}, \phi_g\right)$ with $\mu_{gij} = \exp(x_i^T \beta_g + \log(N_{ij}))$ where N_{ij} is the normalized library size. For non-DE genes, we set $\mu_{g1} = \mu_{g2}$.

ROC Curves of One Scenario



nGenes=10000, nSample=16, pDiff=30% Scenario ROC Curves

AUC Plot


pDiff = 1%
pDiff = 10%
pDiff = 30%
nSample = 4
nSample = 8
nSample = 16
AUC Value
0.5
0.6
0.7
0.8
0.9
Method
sSeq
DESeq
edgeR
EBSeq
DESeq2
eBayes
genes
1000
10000

Summary of the Results

Effect of nGenes: not obvious

Effect of pDiff: smaller pDiff -> larger differences between eBayes and other methods

Effect of nSample: smaller nSample -> larger differences between eBayes and other mehods

Discussion

For the future research, we could:
(1) add more methods: baySeq, ShrinkSeq, NOISeq, SAMseq;
(2) include more varieties;
(3) consider the flow cell effects;
(4) improve the eBayes by refining the hierarchical model