Conditional estimation of local pooled dispersion parameter in small-sample RNA-seq data improves differential expression test

Gim, Jungsoo
Institute of Health and Environment
Seoul National University

Co-work with Sungho Won and Taesung Park
RNA-Seq is useful

RNA-Seq provides many information.
  novel alternative splicing
  different exon usage
  RNA editing

RNA-Seq measures accurate expression profiles.
  Especially, both low and high intensity measure
RNA-Seq is usefull

Trade-off in RNA-Seq

RNA-Seq is still expensive

Trade-off (with the same cost)

<table>
<thead>
<tr>
<th></th>
<th>High coverage</th>
<th>Low coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extra-Info</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Replicate</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- **Different Structure Analysis**
- **Different Expression Analysis**
Count modeling in RNA-Seq

- **Count-valued data has overdispersion property.**
  - Microarray SAGE
  - RNA-Seq
- **Modeling count-valued data**
  - Poisson, Binomial
  - Beta-binomial, over-dispersed log-linear, etc
- **Simulation study**
  - Lu et al (2005): suggested the NB assumption can be reliable even with non-NB sampling, thus provide a more flexible framework.
- **Comparison of methods**
  - F. Rapaport (2013): NB-based models work superior to others in general
- **Log-counts trials**
  - Charity et al (2014): log-count based modeling works better
  - R. Morrison Cassie (1862): log-transformed NB behaves like normal
Limitations

• The number of replicates is small (sometimes non-replicated) due to sequencing costs.

• With a small number of (biological) replicates, the most methods perform poorly (Rapaport, 2013; Soneson, 2013).

• Even with reduced sequencing cost, a study with a small number of replicates happens.
  • When the sample of interest is a rare sample

• Thus,
The Objective

• To develop a method to test DE (differential expression) of RNA-Seqencing data with a small number of replicates
The Objective

• To develop a method to test **DE (differential expression)** of RNA-Sequencing data with a small number of replicates by adopting ideas of

  • **Local Pooled Error (LPE)** idea: to group genes with similar mean intensity (guaranteed in microarray data analysis)

  • **edgeR**: Estimation of dispersion parameter using conditional likelihood based on NB modeling
edgeR

• Developed by Mark D. Robinson and Gordon K Smyth

• Performance is generally good
  • Our own simulation results and other comparison reports
    • Gim et al, 2016; Soneson et al, 2013; Rapaport et al, 2013

• One of the most widely used

• Essential model in edgeR was published in Biostatistics (2008)
  • “Small-sample estimation of NB dispersion, with application to SAGE data”
edgeR (conti’)

- Dispersion estimation (Common vs. Tagwise, or gene-wise)
  - Common dispersion
    - Estimate common dispersion: maximizes the common likelihood
    - $l_C(\phi) = \sum_{g=1}^{G} l_g(\phi)$, where $G$ is the number of genes
  - Tagwise dispersion
    - Empirical Bayes using weighted conditional log-likelihood (WL)
    - $WL(\phi_g) = l_g(\phi_g) + \alpha l_C(\phi_g)$, $\alpha$: weight given to the common likelihood
    - Selecting $\alpha$ as an approximate EB rule
Local-pooled-error

• Developed by Jain et al (2003)
• Analysis for microarray with a small number of replicates
• Evaluate error distribution for each condition
  • For each condition X and Y, MA plot was first evaluated then var(M) across the genes with similar intensities
  • For condition X, three genes in a bin

\[
\begin{align*}
X_1 &\sim N(\mu_1, \sigma_1^2) \rightarrow M_1 = X_{11} - X_{12} \sim N(0,2\sigma_1^2) \\
X_2 &\sim N(\mu_2, \sigma^2) \rightarrow M_2 = X_{21} - X_{22} \sim N(0,2\sigma_2^2) \\
X_3 &\sim N(\mu_3, \sigma^2) \rightarrow M_3 = X_{31} - X_{32} \sim N(0,2\sigma_3^2)
\end{align*}
\]

\[\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2\]
edgeR + Local-pooled-error

• Unlike microarray data, which follows $N(\mu, \sigma^2)$, LPE idea cannot be directly applied to NB-modeled RNA-Seq data

• Variance is largely depends on mean
  • $X_1 \sim NB(\mu, \phi) \rightarrow \text{Var}(X_1) = \mu + \mu^2 \phi$
  • $X_2 \sim NB(\mu + \delta, \phi) \rightarrow \text{Var}(X_2) = \text{Var}(X_1) + \delta(1 + 2\mu\phi + \delta\phi)$

• Variances of the difference between NB variables in a bin are different. However, genes with similar mean value have a similar dispersion value.
edgeR + Local-pooled-error

- **Variances** of the difference between NB variables in a bin are different. However, genes with similar mean value have a similar dispersion value.

- Genes with a similar intensity
  - Case of $X_1, X_2 \sim NB(\mu, \phi)$ (set up $\mu = 10, \phi = 1$) with $n=100$
    - $Var(X_1) = 110.4; Var(X_2) = 108.3; Var(X_1 - X_2) = 217.1$
    - $\hat{\phi} (using \ X_1) = 0.9736; \hat{\phi} (using \ X_2) = 0.9521; \hat{\phi} (using \ X_1 & X_2) = 0.9627$
  - Case of $X_1 \sim NB(\mu = 10, \phi = 1), X_2 \sim NB(\mu = 12, \phi = 1)$ with $n=100$
    - $Var(X_1) = 111.2; Var(X_2) = 156.2; Var(X_1 - X_2) = 265.9$
    - $\hat{\phi} (using \ X_1) = 0.9736; \hat{\phi} (using \ X_2) = 0.9826; \hat{\phi} (using \ X_1 & X_2) = 1.0038$
Genewise or Binwise: An Example

• 10 genes within a same bin (three replicates in each class)

• Estimate dispersion ($\phi$)
  • Bin-wise: estimate dispersion using whole data in a bin
  • Gene-wise: average of dispersions over bin (MLE or MME)
  • Repeat 100 times
An Example - result

<table>
<thead>
<tr>
<th>The number of samples</th>
<th>MSE (Bin-wise estimation)</th>
<th>MSE (Average of gene-wise estimation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.0419</td>
<td>0.2092</td>
</tr>
<tr>
<td>6</td>
<td>0.0192</td>
<td>0.0351</td>
</tr>
<tr>
<td>8</td>
<td>0.0124</td>
<td>0.0177</td>
</tr>
<tr>
<td>10</td>
<td>0.0091</td>
<td>0.0137</td>
</tr>
<tr>
<td>20</td>
<td>0.0048</td>
<td>0.006</td>
</tr>
<tr>
<td>40</td>
<td>0.0024</td>
<td>0.0031</td>
</tr>
</tbody>
</table>
A Proposed Idea

1. Similar to the dispersion estimation method in edgeR
2. For each group, evaluate mean of all genes
3. Order genes according their mean value
   - Median might be alternative
4. Group the genes into bins based on their mean value
   - Equal distance vs. percentile
   - Should determine proper # of bins: start with 100 as was done in LPE
5. For each bin, estimate common dispersion
   - Using edgeR
6. Predict dispersion based on estimates
   - Curve smoothing along the bins
   - Estimates itself
7. DE test using the exact test scheme in edgeR
A Proposed Idea

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>Y1</th>
<th>Y2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.1</td>
<td>1.2</td>
<td>1.4</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1.2</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1.8</td>
<td>3.0</td>
</tr>
<tr>
<td>5</td>
<td>1.5</td>
<td>1.7</td>
<td>2.3</td>
<td>2.4</td>
</tr>
<tr>
<td>6</td>
<td>2.0</td>
<td>2.5</td>
<td>1.8</td>
<td>1.9</td>
</tr>
<tr>
<td>7</td>
<td>2.1</td>
<td>2.3</td>
<td>1.2</td>
<td>1.0</td>
</tr>
<tr>
<td>8</td>
<td>0.3</td>
<td>0.5</td>
<td>1.9</td>
<td>2.0</td>
</tr>
<tr>
<td>9</td>
<td>0.8</td>
<td>0.7</td>
<td>1.2</td>
<td>1.3</td>
</tr>
<tr>
<td>10</td>
<td>1.3</td>
<td>1.6</td>
<td>1.8</td>
<td>1.7</td>
</tr>
</tbody>
</table>
A Proposed Idea

1. Calculate the mean intensity of each gene in two conditions

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>Y1</th>
<th>Y2</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.1</td>
<td>1.2</td>
<td>1.4</td>
<td>0.6</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1.2</td>
<td>0.1</td>
<td>0.2</td>
<td>0.55</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1.8</td>
<td>3.0</td>
<td>0.9</td>
</tr>
<tr>
<td>5</td>
<td>1.5</td>
<td>1.7</td>
<td>2.3</td>
<td>2.4</td>
<td>1.9</td>
</tr>
<tr>
<td>6</td>
<td>2.0</td>
<td>2.5</td>
<td>1.8</td>
<td>1.9</td>
<td>1.9</td>
</tr>
<tr>
<td>7</td>
<td>2.1</td>
<td>2.3</td>
<td>1.2</td>
<td>1.0</td>
<td>1.65</td>
</tr>
<tr>
<td>8</td>
<td>0.3</td>
<td>0.5</td>
<td>1.9</td>
<td>2.0</td>
<td>1.1</td>
</tr>
<tr>
<td>9</td>
<td>0.8</td>
<td>0.7</td>
<td>1.2</td>
<td>1.3</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1.3</td>
<td>1.6</td>
<td>1.8</td>
<td>1.7</td>
<td>1.55</td>
</tr>
</tbody>
</table>
A Proposed Idea

1. Calculate the mean intensity of each gene in two conditions
2. Calculate quantiles with mean intensities and define ‘intensity bins’ using adjacent quantile (or equal-distance) values, then place the genes in the bins where their mean intensities are belonged

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>Y1</th>
<th>Y2</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.1</td>
<td>1.2</td>
<td>1.4</td>
<td>0.6</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1.2</td>
<td>0.1</td>
<td>0.2</td>
<td>0.55</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1.8</td>
<td>3.0</td>
<td>0.9</td>
</tr>
<tr>
<td>5</td>
<td>1.5</td>
<td>1.7</td>
<td>2.3</td>
<td>2.4</td>
<td>1.9</td>
</tr>
<tr>
<td>6</td>
<td>2.0</td>
<td>2.5</td>
<td>1.8</td>
<td>1.9</td>
<td>1.9</td>
</tr>
<tr>
<td>7</td>
<td>2.1</td>
<td>2.3</td>
<td>1.2</td>
<td>1.0</td>
<td>1.65</td>
</tr>
<tr>
<td>8</td>
<td>0.3</td>
<td>0.5</td>
<td>1.9</td>
<td>2.0</td>
<td>1.1</td>
</tr>
<tr>
<td>9</td>
<td>0.8</td>
<td>0.7</td>
<td>1.2</td>
<td>1.3</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1.3</td>
<td>1.6</td>
<td>1.8</td>
<td>1.7</td>
<td>1.55</td>
</tr>
</tbody>
</table>
A Proposed Idea

1. Calculate the mean intensity of each gene in two conditions
2. Calculate quantiles with mean intensities and define ‘intensity bins’ using adjacent quantile (or equal-distance) values, Then, place the genes in the bins where their mean intensities are belonged
3. Estimate common dispersion parameter which maximizes the conditional likelihood of predetermined bin q under each condition

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>Y1</th>
<th>Y2</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.1</td>
<td>1.2</td>
<td>1.4</td>
<td>0.6</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1.2</td>
<td>0.1</td>
<td>0.2</td>
<td>0.55</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1.8</td>
<td>3.0</td>
<td>0.9</td>
</tr>
<tr>
<td>5</td>
<td>1.5</td>
<td>1.7</td>
<td>2.3</td>
<td>2.4</td>
<td>1.9</td>
</tr>
<tr>
<td>6</td>
<td>2.0</td>
<td>2.5</td>
<td>1.8</td>
<td>1.9</td>
<td>1.9</td>
</tr>
<tr>
<td>7</td>
<td>2.1</td>
<td>2.3</td>
<td>1.2</td>
<td>1.0</td>
<td>1.65</td>
</tr>
<tr>
<td>8</td>
<td>0.3</td>
<td>0.5</td>
<td>1.9</td>
<td>2.0</td>
<td>1.1</td>
</tr>
<tr>
<td>9</td>
<td>0.8</td>
<td>0.7</td>
<td>1.2</td>
<td>1.3</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1.3</td>
<td>1.6</td>
<td>1.8</td>
<td>1.7</td>
<td>1.55</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>var</th>
<th>0.5</th>
<th>1.05</th>
<th>1.625</th>
<th>1.9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.675</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A Proposed Idea

1. Calculate the mean intensity of each gene in two conditions
2. Calculate quantiles with mean intensities and define ‘intensity bins’ using adjacent quantile (or equal-distance) values, Then, place the genes in the bins where their mean intensities are belonged
3. Estimate common dispersion parameter which maximizes the conditional likelihood of predetermined bin q under each condition
4. Fit a local regression curve
Simulation Setting

- Following DESeq2 setting
  - Mean and dispersions were drawn from Pickrell et al
  - # of genes = 10000
  - Per each gene, generate m samples based on NB(mean, dispersion) randomly selected from Pickrell’s
  - Where m, the sample size varying 6, 8, 10, 20
  - 80% of total genes had no DE, while 20% of total genes had FC of 2, 3, and 4.
  - Direction of FC randomly assigned

- Measure (DESeq2 case)
  - TPR (sensitivity): the fraction of genes with adjusted p-value < 0.1 among the genes with true differences between group means.
  - PPV (precision): the fraction of genes with true differences between group means among those with adjusted p-value < 0.1. (1-PPV = FDR)
  - FPR (specificity): the fraction of genes with adjusted p-value < 0.1 among the genes with no true differences between groups means. (originally p < 0.01)
Simulation Result
Real Data Analysis

- **Hapmap dataset (60 samples in normal condition):** Montgomery et al. performed sequencing the mRNA fraction of the transcriptome of lymphoblastoid cell lines (LCLs) from 60 CEU (HapMap individuals of European descent) individuals to understand the quantitative difference in gene expression within a human population.

- **Gilad group dataset (6 samples):** Gilad group performed comparative studies to assess intra- and interspecies variation in gene regulatory processes. They used RNA-seq to study transcript levels in humans, chimpanzees, and rhesus macaques, using liver RNA samples from three males and three females from each species.

- **Fly dataset (10 samples):** Graveley et al. studied 30 distinct developmental stages in *Drosophila melanogaster* using RNA-Seq, tiling microarrays and cDNA sequencing. Here we selected a part of the RNA-Seq experimental results with 10 samples in two different developmental stages.
Result: Null Dataset
Result: Reproducibility

Graveley dataset: Overlaps of Total DEGs

Graveley dataset: Overlaps of Top 200 DEGs
Discussion

• Two kinds of modeling RNA-seq data
  • (NB) Count-based
  • (Normal) non count-based

• Review papers reported that NB based method generally performed well in many situations

• cLPDseq performed superior to others with an extremely small number of samples, and similar otherwise.
Thank you for paying attention
Variance Stabilization Transformation

• Variance of NB variable X depends on its mean
• VST can be applied to make \( \text{var}(X) \sim \text{Constant} \)
• To find proper VST \( Z = T(X) \), such that \( \text{Var}(X) \sim \text{Const} \),
  \[
  T(X) = T(\mu) + T'(\mu)(X - \mu) + O(X - \mu)
  \]
• Solving differential equation \( T'(\mu) \sim CV\text{ar}(X)^{-1/2} \) might give an answer
• Previous works
  • Anscombe (1948): \( T(X) = \left( r - \frac{1}{2} \right)^{1/2} \sinh^{-1} \sqrt{\frac{X + a}{r - 2a}}, a = \frac{3}{8} \)
  • Laubscher (1961): \( T(X) = r^{1/2} \sinh^{-1} \sqrt{\frac{X}{r}} + (r - 1)^{1/2} \sinh^{-1} \frac{X + 3/4}{r - 3/2} \)
  • Yu Guan (2009): \( T(X) = \left( r - \frac{1}{2} \right)^{1/2} \sinh^{-1} \frac{X + 0.385}{r - 0.77} \)
  • Anderson (2010): \( T(X) = \int_X \frac{dq}{\sqrt{w(q)}} w(q) \) is variance-mean dependence estimated by DESeq
VST: An example

- Generate a matrix of size 100 X 30
  - $X_{ij} \sim NB(\mu_i, s = 1)$
    - $i = 1, \ldots, 100$ rows represent 100 genes
    - $j = 1, \ldots, 30$ columns per each row represent 30 replicates
    - $\mu_i = 10 \times i$ for each $i$

Conclusion: Seems not stable