Statistical analysis of miRNA-seq and RNA-seq data

Luisa Cutillo
cutillo@tigem.it

SYSTEM BIOLOGY TRAINING COURSE
Bioinformatics core
http://bioinformatics.tigem.it
Outline

• Methods for differential expression analysis of SEQ data
• Presence call: finding a count threshold
• Classification and Clustering of Seq data
a number of software packages have been developed especially for differential expression analysis of RNA-Seq data.
Which test to use for counts?

- An important summary statistic in SEQ experiments is the number of reads in a class (i.e. target transcript, binding region etc.)
- For RNA-Seq, this read count has been found to be linearly related to the abundance of the target transcript.
- Interest lies in comparing read counts between different biological conditions. In the simplest case, the comparison is done separately, class by class.
- We will use the term gene synonymously to class.
- We would like to use statistical testing to decide whether, for a given gene, an observed difference in read counts is significant, that is, whether it is greater than what would be expected just due to natural random variation.
Methods for differential expression analysis of RNA-seq data

- DESeq
- edgeR
- NBPSeq
- TSPM
- baySeq
- EBSeq
- NOISeq
- SAMseq
- ShrinkSeq

Work on the counts directly

Return nominal p-values

+ voom(+limma)
+ vst(+limma)

Combine a data transformation with limma
A comparison of methods for differential expression analysis of RNA-seq data

Charlotte Soneson\textsuperscript{1}\textsuperscript{*} and Mauro Delorenzi\textsuperscript{1,2}

Soneson and Delorenzi \textit{BMC Bioinformatics} 2013, 14:91
http://www.biomedcentral.com/1471-2105/14/91

En extensive comparison of these 11 methods can be found in the recent paper of \textit{Soneson\&Delorenzi, BMC Bioinformatics 2013}!
**edgeR versus DESeq: Commons and Differences**

**COMMONS: NB model**

- If reads were independently sampled from a population with given, fixed fractions of genes, the read counts would follow a multinomial distribution, which can be approximated by the Poisson distribution.

- The Poisson distribution has a single parameter, which is uniquely determined by its mean; However, it has been noted that the resulting statistical test does not control type-I error (the probability of false discoveries) and does not account for variation in relative abundance of different genes between different samples.

- To address this so-called *overdispersion* problem, edgeR and DESeq model count data with negative binomial (NB) distributions: $K_{ij} = NB(\mu_{ij}, \sigma_{ij})$
edgeR versus DESeq: Commons and Differences

Differences: dispersion estimation

- **edgeR:** \( \sigma^2 = \mu + \alpha \mu^2 \)
  - a single constant need to be estimated for each gene
  - statistical test based on \( glm \)

- **DESeq:**
  - a more general data driven mean-variance model
  - statistical test based on fisher exact test like
One of the main challenges in interpreting deep sequencing data is the large number of transcripts with small read counts in the range of 1 to 10. Therefore there is a need to determine a threshold value for saving out significant ones!

This issue can be addressed by first examining the overall distribution of transcripts in our samples which are biological replicates.

BMC Genomics
Distributions of the expression signal count derived from biological replicates of intracellular samples show a high degree of similarity.

In general, the distribution shows **three distinct phases**.

1) A large number of unique transcripts have low count number and are distributed unevenly across the low count number range.
2) This is followed closely by a subsequent phase where the numbers of unique transcripts are distributed evenly across a large range of count number magnitude.
3) The last phase follows the previous with transcripts reverting to an uneven distribution of high count number.

This gives the possibility to distinguish the relative proportions of biologically significant transcripts from potential noise: **we need a threshold for counts!**
Distributions of peak magnitudes after mapping shows the same trend after an initial uneven distribution suggesting the initial phase to be noise! KS statistics is used as distance to estimate where the two distributions begin to differ! HERE COUNT of 32 reads was found to be the THRESHOLD!
32 counts is an optimum threshold for achieving similar cumulative frequency distribution amongst replicates. KS statistics was applied iteratively to the biological replicates (red and black lines), each graph depicts a gradual change in the threshold value. The KS test statistics can be thought of as a cost function that we seek to minimize to ensure that the distributions between the two replicates are similar.
Presence call: an application to Human Retina

Annamaria Carissimo, Luisa Cutillo

from BMC Genomics 2010, 11(Suppl 1):S6
Classification and clustering of sequencing data

The Annals of Applied Statistics
2011, Vol. 5, No. 4, 2493–2518
DOI: 10.1214/11-AOAS493
© Institute of Mathematical Statistics, 2011

CLASSIFICATION AND CLUSTERING OF SEQUENCING DATA USING A POISSON MODEL

BY DANIELA M. WITTEN
University of Washington

Package ‘PoiClaClu’

February 15, 2013

Type Package

Title Classification and clustering of sequencing data based on a Poisson model

Version 1.0.1

Date 2013-01-02

Author Daniela Witten

Maintainer Daniela Witten, <dwitten@u.washington.edu>

Description Implements the methods described in the paper, Witten
(2011) Classification and Clustering of Sequencing Data using a
In her paper Daniela Witten:

• Makes a power transformation of the data

• Proposes new approaches for performing classification and clustering of observations on the basis of sequencing data.

• Develops an analog of diagonal linear discriminant analysis that is appropriate for sequencing data, based on a Poisson log linear model (sparse PLDA).

• Proposes an approach for clustering sequencing data using a new dissimilarity measure that is based upon the Poisson model.
Competitors

1) Classification:
   - Nearest Shrunken Centroid (NSC)
   - Nearest Shrunken Centroid with sqrt error transformation

2) Clustering:
   - edgeR (Robinson, McCarthy, and Smyth (2010))
   - Variance Stabilizing Transformation (VST) according to Anders and Huber (2010)
   - Euclidean distance
# Simulation: Classification results

Simulation results: nine classification methods. NSC, NSC on $\sqrt{X_{ij} + 3/8}$, and sPLDA were performed, using three different size factor estimates: total count (TC), quantile (Q), and median ratio (MR). Cross-validation was performed on a training set of n observations, and error rates were computed on n test observations. We report the mean numbers of test errors and nonzero features over 50 simulated data sets. Standard errors are in parentheses.

<table>
<thead>
<tr>
<th>n</th>
<th>(\phi)</th>
<th>(\sigma)</th>
<th>Method</th>
<th>NSC err.</th>
<th>NSC sqrt err.</th>
<th>sPLDA err.</th>
<th>NSC nonzero</th>
<th>NSC sqrt nonzero</th>
<th>sPLDA nonzero</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>0.01</td>
<td>0.05</td>
<td>TC</td>
<td>4.18 (0.34)</td>
<td>5.74 (0.28)</td>
<td>2.24 (0.26)</td>
<td>1947.6 (441.3)</td>
<td>2217.9 (509.0)</td>
<td>791.4 (111.7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Q</td>
<td>4.38 (0.34)</td>
<td>5.82 (0.26)</td>
<td>2.26 (0.25)</td>
<td>1670.6 (394.5)</td>
<td>2010.1 (478.2)</td>
<td>782.3 (110.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MR</td>
<td>4.28 (0.34)</td>
<td>5.78 (0.27)</td>
<td>2.20 (0.24)</td>
<td>1731.8 (402.6)</td>
<td>2327.8 (517.9)</td>
<td>795.4 (110.8)</td>
</tr>
<tr>
<td>50</td>
<td>0.01</td>
<td>0.025</td>
<td>TC</td>
<td>19.14 (0.67)</td>
<td>24.06 (0.70)</td>
<td>16.84 (0.55)</td>
<td>2316.6 (398.8)</td>
<td>3122.5 (516.6)</td>
<td>1830.7 (217.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Q</td>
<td>20.32 (0.71)</td>
<td>24.82 (0.70)</td>
<td>17.14 (0.56)</td>
<td>1870.7 (335.2)</td>
<td>3380.9 (519.4)</td>
<td>1860.2 (229.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MR</td>
<td>19.66 (0.69)</td>
<td>24.48 (0.69)</td>
<td>16.88 (0.60)</td>
<td>2488.7 (437.8)</td>
<td>2698.7 (513.4)</td>
<td>1934.9 (224.5)</td>
</tr>
<tr>
<td>12</td>
<td>0.1</td>
<td>0.1</td>
<td>TC</td>
<td>2.52 (0.31)</td>
<td>2.66 (0.26)</td>
<td>1.58 (0.25)</td>
<td>5143.2 (527.2)</td>
<td>2738.5 (461.2)</td>
<td>3878.2 (369.4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Q</td>
<td>2.12 (0.27)</td>
<td>2.68 (0.26)</td>
<td>1.62 (0.26)</td>
<td>5207.0 (536.8)</td>
<td>2879.8 (456.7)</td>
<td>3927.2 (371.7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MR</td>
<td>2.28 (0.29)</td>
<td>2.88 (0.28)</td>
<td>1.60 (0.26)</td>
<td>4849.4 (531.2)</td>
<td>2932.0 (477.3)</td>
<td>3889.2 (368.6)</td>
</tr>
<tr>
<td>50</td>
<td>0.1</td>
<td>0.05</td>
<td>TC</td>
<td>16.80 (0.54)</td>
<td>17.76 (0.61)</td>
<td>17.94 (0.70)</td>
<td>3802.5 (408.1)</td>
<td>3785.5 (418.1)</td>
<td>3308.5 (355.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Q</td>
<td>17.08 (0.64)</td>
<td>17.16 (0.60)</td>
<td>17.88 (0.65)</td>
<td>4293.2 (471.9)</td>
<td>3921.5 (371.8)</td>
<td>3284.0 (352.8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MR</td>
<td>16.78 (0.59)</td>
<td>17.34 (0.66)</td>
<td>17.96 (0.71)</td>
<td>3475.3 (392.4)</td>
<td>4398.3 (457.0)</td>
<td>3489.3 (371.9)</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0.2</td>
<td>TC</td>
<td>3.24 (0.25)</td>
<td>4.28 (0.35)</td>
<td>4.26 (0.32)</td>
<td>8846.2 (380.7)</td>
<td>6127.8 (524.6)</td>
<td>4502.0 (509.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Q</td>
<td>3.20 (0.23)</td>
<td>4.04 (0.33)</td>
<td>4.08 (0.29)</td>
<td>8991.8 (318.8)</td>
<td>6342.1 (557.6)</td>
<td>4551.7 (512.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MR</td>
<td>3.22 (0.26)</td>
<td>3.60 (0.30)</td>
<td>4.00 (0.31)</td>
<td>8389.0 (396.4)</td>
<td>7082.7 (515.5)</td>
<td>4518.9 (514.6)</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>0.1</td>
<td>TC</td>
<td>25.56 (0.61)</td>
<td>25.80 (0.55)</td>
<td>25.66 (0.50)</td>
<td>4237.8 (503.5)</td>
<td>4293.5 (495.9)</td>
<td>3150.5 (433.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Q</td>
<td>25.82 (0.61)</td>
<td>25.90 (0.64)</td>
<td>26.02 (0.55)</td>
<td>4629.1 (516.2)</td>
<td>4170.5 (491.7)</td>
<td>3131.2 (406.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MR</td>
<td>25.92 (0.68)</td>
<td>25.86 (0.59)</td>
<td>25.52 (0.51)</td>
<td>4427.5 (524.0)</td>
<td>4362.6 (498.0)</td>
<td>3156.8 (410.4)</td>
</tr>
</tbody>
</table>
Simulation: Clustering results

<table>
<thead>
<tr>
<th>$\phi$</th>
<th>$\sigma$</th>
<th>Method</th>
<th>Clustering error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.15</td>
<td>Cai</td>
<td>0.3592 (0.0071)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Berninger</td>
<td>0.5704 (0.0173)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EdgeR</td>
<td>0.0000 (0.0000)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VST</td>
<td>0.6201 (0.0029)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Squared Euclidean total count</td>
<td>0.5675 (0.0191)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Squared Euclidean quantile</td>
<td>0.5662 (0.0215)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Squared Euclidean median ratio</td>
<td>0.5755 (0.0178)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poisson total count</td>
<td>0.0045 (0.0045)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poisson quantile</td>
<td>0.0057 (0.0047)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poisson median ratio</td>
<td>0.0045 (0.0045)</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2</td>
<td>Cai</td>
<td>0.3803 (0.0058)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Berninger</td>
<td>0.1905 (0.0258)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EdgeR</td>
<td>0.0000 (0.0000)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VST</td>
<td>0.6204 (0.0029)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Squared Euclidean total count</td>
<td>0.3051 (0.0327)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Squared Euclidean quantile</td>
<td>0.2875 (0.0325)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Squared Euclidean median ratio</td>
<td>0.3297 (0.0350)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poisson total count</td>
<td>0.2053 (0.0225)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poisson quantile</td>
<td>0.2067 (0.0228)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poisson median ratio</td>
<td>0.2006 (0.0219)</td>
</tr>
</tbody>
</table>
Discussion

• Performance degrades in the presence of severe overdispersion.

• Most sequencing data seem to be somewhat overdispersed relative to the Poisson model.

• It may be that extending the proposed approaches to the negative binomial model could result in improved performance in the presence of overdispersion.
Open Questions

• Independence assumption?
• Transformation into non-integer values?
• Classification: no clear superiority over NSC in overdispersed simulated or in real data
• Clustering: edgeR performs best
Some bibliography...