Differential Expression Analysis: RNA-Seq

2011 Winter School in Mathematical and Computational Biology
Tuesday, 5 July 2011

Davis McCarthy
Bioinformatics Division
Walter + Eliza Hall Institute of Medical Research
DNA Sequence Ain’t Everything

Study development: same DNA, different form/function

Find genes driving disease

RNA-Seq

- Use of ultra high-throughput sequencing ('next-' or 'second'-generation) technologies for the study of gene expression
- Discover gene activity and function by finding differentially expressed genes
- My focus: statistical methods for differential expression analysis
DE Analysis is Crazy (Statistically Speaking)

- Make inference on DE for ~20,000 genes...simultaneously...from 6 samples
- Ultra high-dimensional data
- Degrees of freedom?
- Analytical methods still open research
The Point

• Good DE analysis of RNA-Seq data requires:
  1. Good statistical models for DE
  2. Understanding and proper modeling of the variability in RNA-seq data
  3. Efficient, useable statistical software

• Negative Binomial GLM methods can be used to analyze complex experimental designs while accounting for biological variation

• edgeR software implements these methods
DE Analysis Happens in R

- R: free, open-source statistical software
- Bioconductor: repository for free, open-source, R packages for bioinformatics
- Several packages currently available for DE analysis of RNA-seq data in R:
  - baySeq [Hardcastle & Kelly, 2010]
  - DEGSeq [Wang et al, 2010]
  - DESeq [Anders & Huber, 2010]
  - edgeR [Robinson, McCarthy & Smyth, 2010]
  - NBPSSeq [Di et al, 2011]
  (+ more methods with less refined code, e.g. TSPM [Auer & Doerge, 2011])
- I work on edgeR, so this is my favourite
- edgeR is only package so far with full GLM capabilities
Outline

1. RNA-Seq and Models for Differential Expression
2. Accounting for Biological Variation
3. NB GLMs in \texttt{edgeR} Handle Complex Designs
4. Case Study: Paired Design Tumour vs Normal
High-throughput sequencing, RNA-Seq, Count Data, Statistical Models, Current State of Play

RNA-SEQ AND DE MODELS
Long pipeline to go from reads to results in an RNA-Seq study

Differential expression (DE) testing

Fig: Oshlack, Robinson & Young, Genome Biology, 2010
**RNA-Seq Data for DE Analysis is a Table of Counts**

<table>
<thead>
<tr>
<th>Tag ID</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000124208</td>
<td>478</td>
<td>619</td>
<td>4830</td>
<td>7165</td>
</tr>
<tr>
<td>ENSG00000182463</td>
<td>27</td>
<td>20</td>
<td>48</td>
<td>55</td>
</tr>
<tr>
<td>ENSG00000125835</td>
<td>132</td>
<td>200</td>
<td>560</td>
<td>408</td>
</tr>
<tr>
<td>ENSG00000125834</td>
<td>42</td>
<td>60</td>
<td>131</td>
<td>99</td>
</tr>
<tr>
<td>ENSG00000197818</td>
<td>21</td>
<td>29</td>
<td>52</td>
<td>44</td>
</tr>
<tr>
<td>ENSG00000125831</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ENSG00000215443</td>
<td>4</td>
<td>4</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>ENSG00000222008</td>
<td>30</td>
<td>23</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ENSG00000101444</td>
<td>46</td>
<td>63</td>
<td>54</td>
<td>53</td>
</tr>
<tr>
<td>ENSG00000101333</td>
<td>2256</td>
<td>2793</td>
<td>2702</td>
<td>2976</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**very high dimensional data**
We Assess DE for Each Gene

- For EACH GENE, is the mean expression level for the gene under one condition significantly different from the mean expression level under a different condition?

<table>
<thead>
<tr>
<th>Tag ID</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000124208</td>
<td>478</td>
<td>619</td>
<td>4830</td>
<td>7165</td>
</tr>
<tr>
<td>ENSG00000182463</td>
<td>27</td>
<td>20</td>
<td>48</td>
<td>55</td>
</tr>
<tr>
<td>ENSG00000125835</td>
<td>132</td>
<td>200</td>
<td>560</td>
<td>408</td>
</tr>
<tr>
<td>ENSG00000125834</td>
<td>42</td>
<td>60</td>
<td>131</td>
<td>99</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

* ... tens of thousands more tags ...*
Model Variability to Assess DE

- Want to assess DE for ~20 000 genes simultaneously
- Which differences are real, which likely to have appeared by chance?
- Answer this question with a statistical model for DE
- Need to understand variability and model it appropriately
Significance of Differences Depends on Variability of Samples

Difference in means unlikely due to chance
Significance of Differences Depends on Variability of Samples

Difference in means more likely due to chance

- Mean 1: 61
  - Standard deviation: 62
- Mean 2: 114
  - Standard deviation: 119

High variance samples

Read count

0
50
100
200
300

mean=61
sd=62

mean=114
sd=119
There are Several Options for Creating a Model for DE for Count Data

Transform count data and apply standard methodology:
• Gaussian (normal) model using transformed counts (e.g. log transform)

Analyze using models for count data:
• Poisson
• Negative binomial (NB)
+ other possibilities

Whatever approach we take, we must account for **biological variation**
Test DE by Fitting a Model for Each Gene

- Fit a statistical model for DE to each gene, accounting for biological variability
- Test null hypothesis of zero DE against alternative of some DE
- Tens of thousands of hypothesis tests
- Multiple testing (usu. control FDR)
Importance of variance model, Technological & Biological Variation, Coefficient of Variation, Negative Binominal Model

ACCOUNTING FOR BIOLOGICAL VARIATION
RNA-Seq Data Exhibits Technical and Biological Variation

• Two levels of variation in any RNA-Seq experiment
  1. Relative abundance (expression level) of each gene will vary between RNA samples, due mainly to biological causes.
  2. There is measurement error
     – uncertainty with which the abundance of each gene in each sample is estimated by the sequencing technology.
Counts from a Single RNA Sample Can be Modeled with Poisson Distribution

Short reads

RNA sample

Map reads to genome

Read 1
Read 2
Read 3
Read 4
Read 5
Read 6
...

M = total number of reads ≈ 20 million
λ_g = true proportion of gene g
y_g = number of reads for gene g

Large M, small λ_i → y_g approximately Poisson, μ_g = Mλ_g
A Small RNA-Seq Experiment (Tech Reps)

RNA from stem cells

RNA from luminal cells

\[ \lambda_{g1} \quad \lambda_{g2} \quad \lambda_{g3} \quad \lambda_{g4} \]

\[ y_{g1} \quad y_{g2} \quad y_{g3} \quad y_{g4} \]

Reads \( M_i \approx 20 \text{ million} \)

Genes \( g = 1, \ldots, 30k \)

\[ \mathbb{E}(y_{gi}) = M_i \lambda_{gi} \]
True Technical Reps Show Poisson Variation for Each Gene

Data:
Marioni et al., *Genome Res*, 2008

binned variance, sample variance
A Small RNA-Seq Experiment (Biological Reps)

RNA from stem cells

RNA from luminal cells

Genes $g = 1, \ldots, 30k$

Reads $M_i \approx 20$ million

$E(y_{gi}) = M_i \lambda_{gi}$
Technical replicate counts for a gene vary according to a Poisson law, i.e. *sequencing variation is Poisson*

- **Biological CV (BCV)** is the coefficient of variation with which the (unknown) true abundance of the gene varies between RNA samples.

- Let $BCV^2 = \phi$

- If you can determine BCV then you have a quadratic mean-variance relationship

\[
\text{var}(y_{gi}) = \mu_{gi} + \phi_g \mu_{gi}^2
\]

with $\phi_g =$ Biological CV$^2$
Biological Coefficient of Variation Dominates Technical

Total $CV^2 = \text{Technical } CV^2 + \text{Biological } CV^2$

- Separate biological and technical variation
- Technical $CV$ decreases as size of counts increases. BCV does not.
- BCV likely to be the dominant source of uncertainty for high-count genes
- Reliable estimation of BCV is crucial for realistic assessment of DE in RNA-Seq experiments.

$$CV^2(y_{gi}) = \frac{\text{var}(y_{gi})}{\mu_{gi}^2} = \frac{1}{\mu_{gi}} + \phi_g$$
Biological Replicate Data shows Quadratic Mean-Variance Relationship
(development cycle of slime mould, 2 samples at hr00, & 2 at hr04)

BCV = 0.38

Data: Parikh et al, Genome Biology, 2010

binned variance, sample variance
Quadratic Mean-Variance Relationship Leads to Negative Binomial Model

• (With a couple of assumptions) counts follow a negative binomial distribution

\[
E(y_{gi}) = M_i \lambda_{gi}, \quad y_{gi} \sim \text{NegBin}(\mu_{gi}, \phi_g)
\]

• Reasonable model for DE in RNA-seq data that accounts for biological variation
Generalized Linear Models, NB Model, Estimating BCV, Flexibility

NB GLMS IN EDGER HANDLE
COMPLEX DESIGNS
We Need to Analyse Multifactor RNA-Seq Experiments

- Existing software could only handle single-factor (one-way layout) experiments ("classic" edgeR, DESeq, baySeq, etc.)
- Cannot account for paired samples
- Cannot account for multiple factors
- Cannot account for batch effects
- Cannot account for covariates
Inference on Differentially Expressed Genes May Be Wrong If Information Ignored

Difference between cancer and normal samples looks \textbf{not significant}
Differentially Expression Found When Accounting for Patient Effect

2-fold difference between cancer and normal within each patient
GLM Methods Are Flexible

- GLM (generalized linear model) approach handles complicated designs – any design that can be expressed as a linear model
- Fit full model and a null (smaller) model to the data to each gene (i.e. fit 20,000 models)
- Use a likelihood-ratio test to determine DE
- GLM methods apply to NB distribution
Sharing Information Across Genes Improves BCV Estimation

- Variance structure (BCV) estimated from dataset as a whole
- **Stabilize estimates & inference** – important for small sample sizes
- **Common BCV** for all genes – estimated from all of the data
- Common conditional likelihood
- Acts like a **Bayesian prior** distribution
edgeR Overcomes Issues with Estimating BCV

• In edgeR we allow BCV that:
  – varies between genes (“tagwise”), and
  – shows a systematic trend with respect to gene expression

• Weighted likelihood (cf empirical Bayes) to obtain gene-wise BCVs squeezed towards common BCV

• Cox-Reid APL (approx. conditional inference) used to estimate common/trended/gene-wise BCV for GLMs
NB GLM methods can be used to analyse multifactor experiments while accounting for biological variation
CASE STUDY: PAIRED DESIGN
TUMOUR VS NORMAL

NB GLMs improve the analysis of an RNA-Seq experiment studying differential expression from paired oral squamous cell carcinoma and normal oral tissue samples from 3 patients.
We Aim to Find Genes DE between Normal and Tumour, Accounting for Patient Effects

- Comparing oral squamous cell carcinoma tissue to matched healthy oral tissue
- 6 samples, paired design

<table>
<thead>
<tr>
<th>Normal</th>
<th>Tumour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 8</td>
<td>N8</td>
</tr>
<tr>
<td>Patient 33</td>
<td>N33</td>
</tr>
<tr>
<td>Patient 51</td>
<td>N51</td>
</tr>
</tbody>
</table>
Include Patient and Tissue Effects in Additive GLM Approach

- A log-linear model fitted to the counts $y_{gi}$ for each gene
- Model includes a patient factor (three levels) and a tissue factor (tumour & normal)
- We can estimate the baseline patient expression levels for each gene.
- Likelihood ratio tests are used to test the null hypothesis that the tumour vs normal log-fold change for each gene is zero (after accounting for patient effects)
Including Patient Effect Reduces BCV Estimate

• Cox-Reid takes experimental design into account when estimating BCV
  – corrects for baseline differences between patients
• Estimated (common) BCV = 40%
• True expression levels show substantial biological variability, even accounting for patient effects
• Treating the three tumour samples as independent replicates would yield a higher BCV of 52%
• Paired design is successful in correcting for some patient to patient variation – increases power
Common BCV is Too Simple: Substantial subset of genes shows strong evidence of greater variability than implied by common BCV

Data:
Tuch et al., 2008
Tagwise BCV Gives Best Overall Fit to These Data

Deviance goodness of fit statistics transformed to normality for QQ-plot

- Common BCV rejected for 39 genes at a family-wise error rate of 0.05
- Allowing an abundance trend on the BCV does not reduce the number of outlier genes for which the BCV is rejected, but tagwise BCV drastically does
- Poisson model rejected for 72% of genes (not shown)
GLM Approach Compares Favourably to Original Analysis

• A more formal analysis that assesses statistical significance relative to biological variation
• GLM approach finds more DE genes with statistical evidence
• GLM approach yields biologically relevant DE genes
CONCLUSION

Software Implementation, Applications, Final Points
NB GLMs Show Great Utility for DE Analysis of RNA-Seq Data

- NB GLMs can be used to analyse differential expression in multifactorial RNA-Seq experiments while accounting for biological variation.
- Case studies show these methods to be useful on small and large datasets with very different characteristics.
- Implementation in the edgeR package offers flexible, highly efficient statistical tools.
Special Acknowledgement

• Yunshun (Andy) Chen: PhD student, WEHI Bioinformatics
  – Cox-Reid APL (estimating BCV for GLMs)
  – making the GLM fit secure (line search etc.)
  + much more
Acknowledgements

WEHI Bioinformatics

Gordon Smyth
Yunshun Chen
Mark Robinson
Terry Speed
edgeR Users

Thanks!