Statistical tests for differential expression in RNAseq

Jelle Goeman and Erik van Zwet, LUMC

NBIC Advanced RNA-seq course
16 October 2012
Academic Medical Center, Amsterdam
The analysis of a microarray experiment

- Pre-process
  - image analysis to produce relative abundances
  - take the logarithm
  - remove experimental artifacts (normalize)
- perform 20,000 $t$ tests
- Correct for multiple testing
The model

For any particular gene, the log expression levels are assumed to be normally distributed:

\[ \text{fit} = \text{lm}(y \sim \text{group}) \]

or perhaps

\[ \text{fit} = \text{lm}(y \sim \text{age+group}) \]

- One might further assume that the gene variances come from a common (Gamma) distribution. In R: limma.
- Results in shrinkage: the empirical variances are pulled towards their common mean.
- Favours genes with a large variance.
Microarray data analysis

Assumptions

- Data are normally distributed
- Single error term combines biological and technical variation
- After normalization (log) no relationship between mean and variance

In NGS data
These three assumptions do not hold anymore
Next Generation Sequencing

An NGS experiment produces very many counts.

▶ not well approximated by the normal distribution
▶ or maybe so: voom (later)

Classical way to deal with counts

Binomial and Poisson distribution

Relationship between binomial and Poisson

If \( n \) is large and \( p \) small, \( X \sim \mathcal{B}(n, p) \) is the same as \( X \sim \mathcal{P}(np) \)
Example from Baggerly et al. 2003

<table>
<thead>
<tr>
<th>library</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>pheno</td>
<td>T+</td>
<td>T+</td>
<td>T+</td>
<td>T+</td>
<td>T+</td>
<td>T−</td>
<td>T−</td>
<td>T−</td>
</tr>
<tr>
<td>size</td>
<td>100474</td>
<td>96631</td>
<td>92510</td>
<td>95785</td>
<td>18705</td>
<td>95155</td>
<td>91593</td>
<td>98220</td>
</tr>
<tr>
<td>tag 1</td>
<td>129</td>
<td>167</td>
<td>71</td>
<td>61</td>
<td>6</td>
<td>43</td>
<td>247</td>
<td>509</td>
</tr>
<tr>
<td>tag 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We want to test if tag 1 is more/less abundant in the tumor versus the control group.
In the tumor group we have
\[
\frac{434}{404105} = 0.0011
\]
and in the control group we have
\[
\frac{799}{284968} = 0.0028.
\]
So, tag 1 occurs more than twice as often in the control group.
Is this statistically significant?
Total count for tag 1 in the tumor group has a binomial distribution with $n_1 = 404105$ and some probability $p_1$.

Total count for tag 1 in the control group has a binomial distribution with $n_2 = 284968$ and some probability $p_2$.

We can use R to test the null hypothesis $H_0 : p_1 = p_2$

```r
x=c(434,799)
n=c(404105,284968)
prop.test(x,n)
```

p-value < 2.2e-16.

Giddyup!
We put all subjects in each group together, and assumed binomial distributions for the total counts of tag 1. This is only valid if the proportions for each subject within a group are the same.

<table>
<thead>
<tr>
<th>library</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>pheno</td>
<td>T+</td>
<td>T+</td>
<td>T+</td>
<td>T+</td>
<td>T+</td>
<td>T−</td>
<td>T−</td>
<td>T−</td>
</tr>
<tr>
<td>size</td>
<td>100474</td>
<td>96631</td>
<td>92510</td>
<td>95785</td>
<td>18705</td>
<td>95155</td>
<td>91593</td>
<td>98220</td>
</tr>
<tr>
<td>tag 1</td>
<td>129</td>
<td>167</td>
<td>71</td>
<td>61</td>
<td>6</td>
<td>43</td>
<td>247</td>
<td>509</td>
</tr>
<tr>
<td>prop. (%)</td>
<td>0.13</td>
<td>0.17</td>
<td>0.08</td>
<td>0.06</td>
<td>0.03</td>
<td>0.05</td>
<td>0.27</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Do subjects 1 and 2 have the same proportion? No! \( p=0.012 \).

```r
x=c(129,167)
n=c(100474,96631)
prop.test(x,n)
```
Overdispersion

- There is biological/measurement variation between subjects.
- This will increase the variance of the total count: overdispersion.
- Failure to recognize this, will lead to underestimation of the variance.
- Underestimation of the variance, will lead to many false positives.
Modeling overdispersion

We need to account for extra (biological) variation.

- For every subject, $X_i \sim \text{Bin}(n_i, p_i)$.
- Equivalent ($n_i$ large): $X_i \sim \text{Pois}(n_i p_i)$
Modeling overdispersion

We need to account for extra (biological) variation.

- For every subject, $X_i \sim \text{Bin}(n_i, p_i)$.
- Equivalent ($n_i$ large): $X_i \sim \text{Pois}(n_i p_i)$
- Technical variation: $X / n_i$ varies around $p_i$
Modeling overdispersion

We need to account for extra (biological) variation.

- For every subject, $X_i \sim \text{Bin}(n_i, p_i)$.
- Equivalent ($n_i$ large): $X_i \sim \text{Pois}(n_ip_i)$
- Technical variation: $X/n_i$ varies around $p_i$
- Biological variation: $p_i$ is different for each subject
- Assumption: the $p_i$ are drawn from a population distribution
Modeling overdispersion

We need to account for extra (biological) variation.

- For every subject, $X_i \sim \text{Bin}(n_i, p_i)$.
- Equivalent ($n_i$ large): $X_i \sim \text{Pois}(n_ip_i)$
- Technical variation: $X/n_i$ varies around $p_i$
- Biological variation: $p_i$ is different for each subject
- Assumption: the $p_i$ are drawn from a population distribution
- Overdispersion: biological variation increases the dispersion (=variance)
- Test: if the $p_i$ differ on average between the two groups.
Many ways to deal with overdispersion:

- \( \log(P_i/1 - P_i) \) normally distributed (GLMM)
  - too slow to do a million tests or more.

- \( P_i \) beta distributed. R package BBSeq
  - seems to be too slow as well.

- If \( n \) is large and \( p \) is small, then we can use Poisson-Gamma or negative binomial. R packages baySeq, DESeq, edgeR.
  - we’ll discuss now.
Negative binomial

Technically
If $X|\lambda \sim \mathcal{P}(\lambda)$ and $\lambda \sim \Gamma(r, p/(1 - p))$, then $X \sim \mathcal{NB}(p, r)$
Recall, Binomial$(n, p)$ distribution has

$$\text{mean} = np \quad \text{and} \quad \text{variance} = np(1 - p).$$

DESeq and edgeR assume Negative-Binomial$(\mu, \phi)$ distribution with

$$\text{mean} = \mu \quad \text{and} \quad \text{variance} = \mu + \phi \mu^2.$$

Much effort goes into estimating the dispersion $\phi$ when very few (biological) replicates are available.
Negative binomial

Technically

If $X|\lambda \sim \mathcal{P}(\lambda)$ and $\lambda \sim \Gamma(r, p/(1 - p))$, then $X \sim \mathcal{NB}(p, r)$

Recall, Binomial($n, p$) distribution has

$$\text{mean} = np \quad \text{and} \quad \text{variance} = np(1 - p).$$

DESeq and edgeR assume Negative-Binomial($\mu, \phi$) distribution with

$$\text{mean} = \mu \quad \text{and} \quad \text{variance} = \mu + \phi \mu^2.$$

Much effort goes into estimating the dispersion $\phi$ when very few (biological) replicates are available.

▶ Here we should point out that it is generally better to increase the sample size, if you can.
Common or “tagwise” overdispersion

How much overdispersion is there?
Must estimate from the data

Common overdispersion
Biological variation is the same for every read/gene
Highly unrealistic assumption that makes overdispersion easy to estimate

Tagwise overdispersion
Biological variation different for every gene
Biological reality, but needs large sample size to estimate reliably

Middle ground
Bit of both (cf. limma). More tagwise with larger sample size
From the edgeR manual
Much more on baySeq, DESeq and edgeR later today.
We condition on the library sizes, and test if the proportion of any given tag relative to all other tags is different between groups.

But, if any tag differs between groups then other tags must do so as well, simply because \( \sum_i p_i = 1 \).

In a two sample binomial experiment with

\[
X_1 \sim \text{Bin}(n_1, p_1) \quad \text{and} \quad X_2 \sim \text{Bin}(n_2, p_2)
\]

it makes no sense to test both

\[
H_0 : p_1 = p_2 \quad \text{and} \quad H_0 : (1 - p_1) = (1 - p_2).
\]
The problem is very noticeable if there are tags that are both very abundant and different between groups.

Anders and Huber (2010) and Robinson and Oshlack (2010) suggest to modify the library sizes so that single tags have little influence.

Exclude the k% most abundant genes when calculating library size.

Alternatively, we might take a set of tags that we know won’t differ between groups and test all tags of interest relative to that set.
VOOM

Alternative to negative binomial modelling
Treat the data the same as continuous (microarray) data
  ▶ Log-transform \(^2\log(X + \frac{1}{2})\)
  ▶ Correct by lowess to suppress mean-variance relationship
  ▶ Put into linear model (limma)

Rationale
If \(n\) is large, biological variance dominates
Ignore technical variation
Reverse of binomial model we started with
VOOM: pros and cons

pros

▶ Easy and fast
▶ Microarray analysis methods become available (gene set tests)
▶ Accurate and powerful for high abundant genes

cons

▶ Problems with low-abundant reads (zero counts problem)
▶ Count nature of data ignored
▶ Constant mean-variance relationship disputable

More on voom when discussing gene sets
Many tests

Unadjusted testing
Reject null hypotheses if p-value $< 0.05$
Means 5% of true null hypotheses is significant
5% of 20,000 tests $\rightarrow$ 1,000 false positives?

Prevent a flood of false positives
Do multiple testing correction
False discovery rate

**Definition**

FDR = proportion of false positives among findings

**Control of FDR at 5%**

Make sure that on average at most 5% of findings is false

**Procedure of Benjamini and Hochberg**

Recommended to control FDR

Use function `p.adjust` in R

**Adjusted P-values**

P-values corrected for multiple testing. Significant if $< 0.05$