Differential Gene Expression II – quantifying differences

When is a difference significant I
  – modest numbers of counts: Fisher's Exact Test
  – means and standard deviations: Student's t-test

The signal and the noise - normalization

When are differences significant II
  – multiple test correction: Bonferroni
  – False discovery rates (FDR, q-value)

To learn more:

1. Pevsner, Chapter 8 pp. 331-373
Differential Gene Expression

- Large quantity of data (>20,000 genes)
  - Affychip data has ≥20 replicates per gene
  - RNAseq has counts (FPKM: Fragments per Kilobase per Million mapped reads)
  - but a small number of biological replicates
- Ideally, identify modest change (1.5x or larger) for modest levels of transcription
  - 10 or fewer transcripts may account for 90% of reads, so 5,000 transcripts for < 10% of reads
  - If technical replicates vary more than 2x, how do you measure 1.5x change?
- Large numbers of tests: how to correct?
  - Family-wide-error-rate (FWER) Bonferroni correction (used for similarity search E()-values)
  - False-discovery-rate (FDR, qvalue)

Cells in different tissues are different

Because they express different proteins from different mRNAs
induction of detoxification gene mRNAs


Microarrays vs RNAseq

* Sequenced at a concentration of 1.5 pM

Figure 1. Graphical representation of the study design. (A) Summary of the experimental design. (B) The lanes in which each sample was sequenced across the two runs. In each run, the control sample was sequenced in lane 5. Samples were sequenced at two concentrations: 1.5 pM (indicated by an asterisk) and 3 pM (no asterisk).

Microarrays vs RNAseq

Figure 3. Comparing counts from Illumina sequencing with normalized intensities from the array: brain (left) and liver (right). In each panel, the average log2 counts for each gene are plotted on the y-axis and the array's normalized intensity on the x-axis. To avoid biasing the log2 fold change of each of the average fold change toward high.

Figure 4. Comparison of estimated log2 fold changes (liver/kidney) from Illumina (Y-axis) and Affymetrix (X-axis). We consider only genes that were interrogated using both platforms and genes where the mean number of counts across lanes was greater than 0 for both the liver and kidney samples. (Red and green dots) Genes called as differentially expressed based on the Illumina sequencing data at an FDR of 0.1% with a mean number of counts greater than (red) or less than (green) 200 reads in both tissues. (Black dots) Genes not called as differentially expressed based on the Illumina sequencing data. The set of differentially expressed genes that show the strongest correlation between the two technologies seems to be those that are mapped to by fewer reads (green), while the correlation is weaker for differentially expressed genes mapped to by fewer reads (green).


Measuring differences – sources of variation

Technical
- RNA isolation
- cDNA synthesis
- hybridization (AffyChip)
- PCR amplification
- G+C content
- sequencing depth
- location on AffyChip/sequencing "lane"

Biological
- genetic background
- sex
- last meal/sleep/exercise
- dividing/quiescent
- cell type within tissue type
- …

fasta.bioch.virginia.edu/biol4230
Biological and technical variation - replicates

The variance of the FPKM varies with abundance (expected)
But large variance for replicates (no biology)
FPKM: fragments per Kbase per million mapped reads

Differential Gene Expression

- Large quantity of data (>20,000 genes)
  - Affychip data has ~20 replicates per gene
  - RNAseq has counts (FPKM: Fragments per Kilobase per Million mapped reads)
  - but a small number of biological replicates
- Ideally, identify modest change (1.5x or larger) for modest levels of transcription
  - 10 or fewer transcripts may account for 90% of reads, so 5,000 – 10,000 transcripts for <10% of reads
  - If technical replicates vary more than 2x, how do you measure 1.5x change?
- Large numbers of tests: how to correct?
  - Family-wide-error-rate (FWER) Bonferroni correction (used for similarity search E()-values)
  - False-discovery-rate (FDR, qvalue)
Biological and technical variation - replicates

The variance of the FPKM varies with abundance (expected)
But large variance for replicates (no biology)

Goal: to identify differential expression
Separate between sample differences from within sample differences

The significance of differences:
Fisher's Exact Test

1. Around 1930, Muriel Bristol claimed, in a conversation with R. A. Fisher, that she could tell when milk was poured into tea, which was much preferable to tea being poured into milk.
2. Fisher choose to test this hypothesis by preparing 8 cups of tea, 4 tea first, 4 milk first, and asking Ms. Bristol to identify the 4 cups with tea first.
3. If she has no ability to identify milk first/tea first, then one expects her to be right 50% of the time (4 cups). But what if she was right for 6 of the 8 cups?

\[
\text{fisher.test(matrix(c(4,0,0,4), nrow=2),}
+ \text{alternative='greater')}\]

Fisher's Exact Test for Count Data

data: matrix(c(4, 0, 0, 4), nrow = 2)
p-value = 0.01427
alternative hypothesis: true odds ratio is not equal to 1
Fisher's Exact Test

If she has no ability to identify milk first/tea first, then one expects her to be right 50% of the time (2 cups). But what if she was right for 3 of the 4 cups?

```r
> fisher.test(matrix(c(4,0,0,4),nrow=2),alternative='greater')
Fisher's Exact Test for Count Data
p-value = 0.01427
alternative hypothesis: true odds ratio is not equal to 1

> fisher.test(matrix(c(4,0,1,3),nrow=2),alternative='greater')
p-value = 0.07143

> fisher.test(matrix(c(4,1,1,4),nrow=2),alternative='greater')
p-value = 0.1032

> fisher.test(matrix(c(5,1,1,5),nrow=2),alternative='greater')
p-value = 0.04004

> fisher.test(matrix(c(8,2,2,8),nrow=2),alternative='greater')
p-value = 0.01151
```

1. Perfect is significant in 8 correct assignments
2. 1 mistake is almost significant (4 mistakes seems random)
3. 2 mistakes is ALMOST significant in 10 choices
4. 2 mistakes IS significant in 12 choices
5. 4 mistakes IS significant in 20 choices

Fisher's Exact Test when?

- Categorical data:
  - is/is not a eukaryote
  - is/is not in multiple domains
  - is/is not an enzyme
- 2x2 contingency table
- one table per protein
  - for many proteins, multiple tests
Differential gene expression

- mRNA levels affect protein levels
  - no mRNA, no protein
  - little mRNA, sometimes lots of protein (long half-life)
  - lots of mRNA, often lots of protein
- RNA abundance:
  - most RNA is ribosomal RNA (rRNA)
  - 10 – 50 mRNA species account for >90% of mRNA abundance
  - sensitive methods detect < 1 molecule/cell (but not with single cells)
- which changes matter?
  - fold differences
    - 100X, from 1:100 molecules/cell?
    - 5X, from 50,000 to 250,000 molecules/cell?
  - mostly high abundance? mostly low abundance?

The significance of differences: Differences of means: Student's 't'-test

data: rn3 and rn3b
\[ t = -4.6426, \text{df} = 2.283, p\text{-value} = 0.0335 \]
alt hyp: true diff in means is not equal to 0
sample est: mean(x) mean(y) 0.1886128 2.8588774

data: rn3.1 and rn3b.1
\[ t = 0.4594, \text{df} = 2.536, p\text{-value} = 0.6824 \]
alt hyp: true diff in means is not equal to 0
sample est: mean(x) mean(y) 1.518745 1.069586

data: rn3.2 and rn3b.2
\[ t = -0.3909, \text{df} = 3.342, p\text{-value} = 0.7195 \]
alt hyp: true diff in means is not equal to 0
sample est: mean(x) mean(y) 0.8793091 1.1442473

Ratio's are accurate, one significant
The significance of differences:
Differences of means: Student's 't'-test

```r
t.test(rn35, rn35b)
Welch Two Sample t-test
data: rn35 and rn35b
t = -3.0229, df = 2.379, p-value = 0.07604
alt hyp: true diff in means is not equal to 0
samp est: mean of x mean of y: 0.9889457 1.9788296

> t.test(rn35.1, rn35b.1)
Welch Two Sample t-test
data: rn35.1 and rn35b.1
t = -2.7326, df = 3.539, p-value = 0.05982
alt hyp: true diff in means is not equal to 0
samp est: mean of x mean of y: 1.353749  2.370543

> t.test(rn35.2, rn35b.2)
Welch Two Sample t-test
t = -2.7434, df = 2.444, p-value = 0.08929
alt hyp: true diff in means is not equal to 0
samp est: mean of x mean of y: 1.147306  1.875439
```

Ratio's are accurate, but not significant
Combined, data is very significant

Biological and technical variation - replicates

The variance of the FPKM varies with abundance (expected)
But large variance for replicates (no biology)
The significance of differences: normalization

Normal vs Normal

Normal vs Downs

Why are the replicates different?
Should the bulk properties differ?

log intensity

log10(abundance)

log2(norm/norm)

log10(norm)

1.0

2.0

3.0

4.0

5.0

6.0

7.0

8.0

9.0

10.0

11.0

12.0

13.0

14.0

15.0

16.0

17.0

18.0

19.0

20.0

Why are the replicates different?
Should the bulk properties differ?
Should individual genes differ?
Should blue (normal) and red (Downs) differ?
Differential Gene expression

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Chromosome</th>
<th>average_DS</th>
<th>average_normal</th>
<th>ttest</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP5O</td>
<td>21</td>
<td>10.48200008</td>
<td>9.78274141</td>
<td>5.95402E-07</td>
</tr>
<tr>
<td>CRYBB2</td>
<td>21</td>
<td>5.852878671</td>
<td>6.71121298</td>
<td>3.54121E-06</td>
</tr>
<tr>
<td>C21orf33</td>
<td>21</td>
<td>8.912057195</td>
<td>8.288735662</td>
<td>8.7199E-06</td>
</tr>
<tr>
<td>WRB</td>
<td>21</td>
<td>9.570755686</td>
<td>8.695134299</td>
<td>9.16733E-06</td>
</tr>
<tr>
<td>ALOX5</td>
<td>10</td>
<td>4.433471475</td>
<td>4.660059997</td>
<td>1.23042E-05</td>
</tr>
<tr>
<td>HRMT1L1</td>
<td>21</td>
<td>9.113649813</td>
<td>8.542185783</td>
<td>1.6958E-05</td>
</tr>
<tr>
<td>PTPN1</td>
<td>20</td>
<td>6.189080034</td>
<td>6.46273814</td>
<td>2.7762E-05</td>
</tr>
<tr>
<td>SBF1</td>
<td>22</td>
<td>4.951511451</td>
<td>5.277542864</td>
<td>4.8516E-05</td>
</tr>
<tr>
<td>ATP5J</td>
<td>21</td>
<td>9.24962725</td>
<td>8.48201437</td>
<td>7.20322E-05</td>
</tr>
<tr>
<td>CAMKK2</td>
<td>12</td>
<td>8.113556363</td>
<td>8.760118621</td>
<td>0.000114723</td>
</tr>
<tr>
<td>NRTN</td>
<td>19</td>
<td>3.380282845</td>
<td>3.599555714</td>
<td>0.000120734</td>
</tr>
<tr>
<td>CTD5PL</td>
<td>3</td>
<td>5.812481403</td>
<td>6.093701363</td>
<td>0.000126665</td>
</tr>
<tr>
<td>USP16</td>
<td>21</td>
<td>7.617121492</td>
<td>6.912594318</td>
<td>0.000127859</td>
</tr>
<tr>
<td>RUNX1</td>
<td>21</td>
<td>3.51090011</td>
<td>3.668377161</td>
<td>0.000129409</td>
</tr>
<tr>
<td>DONSON</td>
<td>21</td>
<td>5.219522885</td>
<td>4.656537056</td>
<td>0.000142897</td>
</tr>
<tr>
<td>FLOT1</td>
<td>6</td>
<td>9.422081402</td>
<td>9.199481419</td>
<td>0.000154443</td>
</tr>
<tr>
<td>USP25</td>
<td>21</td>
<td>7.085999667</td>
<td>6.708867141</td>
<td>0.000203888</td>
</tr>
<tr>
<td>SOD1</td>
<td>21</td>
<td>10.49014262</td>
<td>9.69604866</td>
<td>0.000208907</td>
</tr>
<tr>
<td>ATP5O</td>
<td>21</td>
<td>7.646301474</td>
<td>7.22661437</td>
<td>0.000212335</td>
</tr>
</tbody>
</table>

Differential Gene Expression

- Large quantity of data (>20,000 genes)
  - Affychip data has >20 replicates per gene
  - RNAseq has counts (FPKM: Fragments per Kilobase per Million mapped reads)
  - but a small number of biological replicates
- Ideally, identify modest change (1.5x or larger) for modest levels of transcription
  - 10 or fewer transcripts may account for 90% of reads, so 5,000 transcripts for < 10% of reads
  - If technical replicates vary more than 2x, how do you measure 1.5x change?
- Large numbers of tests: how to correct?
  - Family-wide-error-rate (FWER) Bonferroni correction (used for similarity search E()-values)
  - False-discovery-rate (FDR, qvalue)
So many tests, what is significant?

![Image of a cartoon illustrating the concept of statistical significance in research. The cartoon shows a sequence of panels where scientists investigate the supposed link between jelly beans and acne. The final panel concludes that there's no significant link, with subplots showing unrelated links.]

fasta.bioch.virginia.edu/biol4230
So many tests, what is significant?

**Why Most Published Research Findings Are False**

*John P.A. Ioannidis*

**Summary**

There is increasing concern that most currently published research findings are false. The fundamental reason is simple: a true finding will produce a true result, whereas a false result will produce a false finding with a high probability. The main determinant of the false finding rate is the study power (1 minus the type II error rate), which is usually low. Illusory research findings are also influential, due to the neglect of the research planning process and the flawed conduct and interpretation of research.

---

**How many tests?**

- **Conditions**
  - Genes = N (20,000)
  - At least N (~20,000) simultaneous tests

---

In a typical scientific field, many researchers conduct numerous studies, each of which has a high probability of producing false positive results. The combined effect of these studies leads to the publication of many false findings. This phenomenon is exacerbated by the fact that many researchers have a vested interest in obtaining statistically significant results, which can lead to biased selection of data and analysis methods. The result is a proliferation of false positives in the scientific literature, which can have serious implications for the reliability and reproducibility of research findings.
How many tests?

20,000 simultaneous t-tests on random normal data from the same distribution. There are 1,009 green points (false positives), making up 0.05 of the comparisons (at $\alpha = 0.05$).

Correcting for multiple tests:

- **Bonferroni:**
  - $E() = P \cdot D$ (similarity search)
  - calculate expectation as probability of result $x$ number of tests
  - Family Wide Error Rate (FWER)
  - Ensures $< 1.0$ false positive among all results ($<1.0$ false positive after 20 studies with $E<0.05$)

- **Q-value (False discovery rate, FDR)**
  - sets a rate of false positives AMONG the set found to be significant
  - $q$-value $< 0.01$ says that one of the 100 "significant" results will occur by chance (10 of the 1000 significant)
  - which one?
    - One with least signal?
    - One with least fold change?
True positives and false positives

Mixed change, p < 0.05

- 500 100X
- 1,500 10X
- 3,000 1.5X
- 15,000 negative

Correcting for multiple tests

<table>
<thead>
<tr>
<th></th>
<th>Null True (H₀)</th>
<th>Alternate True (H₁)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Significant</td>
<td>V False Pos</td>
<td>S True Pos</td>
<td>R discoveries?</td>
</tr>
<tr>
<td>Test Not Significant</td>
<td>U True Neg</td>
<td>T False Neg</td>
<td>m-R</td>
</tr>
<tr>
<td>Total</td>
<td>m₀</td>
<td>m−m₀ true altern.</td>
<td>m</td>
</tr>
</tbody>
</table>

FWER (family wide error rate) = p(V>1.0) = 0.05 = 1-p(V=0)
p' = p₀/N (number of tests) false positives per analysis

very conservative
True positives and false positives

Mixed change, $p < 0.05$

Mixed change, $p < 0.05/20K$ (Bonferroni)

5,000 $>1.5X$

FWER (family wide error rate) = $p(V>1.0)$

$0.05 = 1-p(V=0)$

$p' = p_0/N$ (number of tests)

*very conservative*

Correcting for multiple tests

<table>
<thead>
<tr>
<th></th>
<th>Null True ($H_0$)</th>
<th>Alternate True ($H_1$)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Significant</td>
<td>V</td>
<td>S</td>
<td>R</td>
</tr>
<tr>
<td>Test Not Significant</td>
<td>U</td>
<td>T</td>
<td>m-R</td>
</tr>
<tr>
<td>Total</td>
<td>$m_0$</td>
<td>m–m0 true altern.</td>
<td>m</td>
</tr>
</tbody>
</table>

FDR (false discovery rate) = $p(V/R)$

Approx FDR *False* discoveries among all discoveries false positives per *discovery/true positive*
A density histogram of the 3,170 $p$ values from the Hedenfalk et al. (14) data. The dashed line is the density histogram we would expect if all genes were null (not differentially expressed). The dotted line is at the height of our estimate of the proportion of null $p$ values.

Storey (2003) PNAS 100:9440, Fig. 1
True positives and false positives

- No change, p < 0.05
- Mixed change, q < 0.05

Summary:

```
qvalue(p = no_change_pvals)
Cumulative number of significant calls:
<1e-04 <0.001 <0.01 <0.025 <0.05 <0.1 <1
p-value  3  17  138  368  821 1737 20000
q-value  0  0   0   0   0    0   0

qvalue(p = mix_pvals)
Cumulative number of significant calls:
<1e-04 <0.001 <0.01 <0.025 <0.05 <0.1 <1
p-value  204 713 1859 2715 3617 4884 20000
q-value  3   7  375  779 1191 2171 20000
```
Reducing variance improves detection

mixture (1.5X, 5X, 25X), p < 0.05

mixture sqrt(var), p < 0.05

mixture, q < 0.05

500
1500
1039
640

summary(mix_pvals_a_qv)
Cumm <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1
p-value 937 1582 2372 2915 3679 4945
q-value 86 708 1597 1952 2250 2664

mixture, q < 0.05

500
1500
698

qvalue(mix_pvals_b)
<1e-04 <0.001 <0.01 <0.025 <0.05 <0.1
p 1853 2121 2426 3529 4354 5599
q 1381 1906 2176 2420 2809 3496

Differential Gene Expression

• Large quantity of data (>20,000 genes)
  – Affychip data has ≥20 replicates per gene
  – RNAseq has counts (FPKM: Fragments per Kilobase per Million mapped reads)
  – but a small number of biological replicates
• Ideally, identify modest change (1.5x or larger) for modest levels of transcription
  – 10 or fewer transcripts may account for 90% of reads, so 5,000 transcripts for < 10% of reads
  – If technical replicates vary more than 2x, how do you measure 1.5x change?
• Large numbers of tests: how to correct?
  – Family-wide-error-rate (FWER) Bonferroni correction (used for similarity search E()-values)
  – False-discovery-rate (FDR, qvalue)