Screening

Methods Course: Gene Expression Data Analysis

-Day Three –

Rainer Spang
Comparing Conditions

Two cell/tissue/disease types:
- wild-type / mutant
- control / treated
- disease A / disease B
- responding / non responding
- etc. etc....

For every sample (cell line/patient) we have the expression levels of thousands of genes and the information whether it is A or B.
Which genes are differentially expressed in the two tissue type populations?
A cost efficient (cheap) experiment:

We observe a gene with a two-fold higher expression in profile A than in profile B.

Is two-fold trust worthy?

Well, by how much can this gene change in group A and in group B?

By no more than 10% than the answer is yes, by up to 500% then the answer is no.
Is a three-fold induced gene more trustworthy than a two-fold induced gene?

Actually this depends on the within class variability of the two genes again, it can be the other way round.
The information in the variability is crucial

In addition to the differences in gene expression you also have a vital interest in its variability ... This information is needed to obtain meaningful lists of genes

Therefore: Invest in repeated experiments!
Standard Deviation and Standard Error

Standard Deviation (SD): Variability of the measurement

Standard Error (SE): Variability of the mean of several measurements

n Replications

Normal Distributed Data:

\[ SE = \frac{1}{\sqrt{n}} SD \]
Precision by Repetition

Repetitions lead to a more precise measurement of gene expression. Single expression measurements are very noisy, average expression across several repetitions is much less noisy.

Therefore: **Invest in repeated experiments!**
The Additive Scale

Most statistics works on an additive scale

Biology works on a multiplicative scale

Conclusion: *Transform your data to the additive scale*

- Simple way: take logs
- Better way: use variance stabilization
Questions:

Which genes are differentially expressed?

→ Ranking

Are these results „significant“

→ Statistical Analysis
Problem: Produce an ordered list of differentially expressed genes starting with the most up regulated gene and ending with the most down regulated gene

*Ranking means finding the right genes... drawing our attention to them*
Ranking is not Testing

**Ranking**: Finding the right genes

**Testing**: Deciding whether genes are significant

The criteria for which ranking is best is different from the criteria which test is best … power is often no argument
Which gene is more differentially expressed?
Ranking is **Scoring**

You need to score differential gene expression

Different scores lead to different rankings

Which scores are there?
Fold Change & Log Ratios

You have transformed your data to additive scale!

Factors become differences:

$$\log(a/b) = \log(a) - \log(b)$$

If you want to rank by fold change you compute the average expression in both groups and subtract them.

$$LR = \bar{X}_1 - \bar{X}_2$$
**T-Score**

**Idea:** Take variances into account

**Change:** low  
**Variance:** high

**Change:** high  
**Variance:** low

**Change:** high  
**Variance:** high

\[
T = \frac{\bar{X}_1 - \bar{X}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}
\]
Fudge Factors:

You need to estimate the variance from data.

You might underestimate an already small variance.

The denominator in T becomes really small.

Constantly expressed genes show up on top of the list.

Correction: Add a constant fudge factor $s_0$.

→ Regularized T-score

$$T_r = \frac{\bar{X}_1 - \bar{X}_2}{c(s + s_0)}$$

→ Limma
→ SAM
→ Twilight
**More Scores:**

- Wilcoxon Score (robust)
- PAUc Score (separation)
- paired t-Score (paired Data)
- F-Score (more than 2 conditions)
- Correlation to a reference gene
  
  etc etc
Different scores give different rankings

<table>
<thead>
<tr>
<th>Gene</th>
<th>t-score</th>
<th>Limma</th>
<th>Fudge</th>
<th>Log ratio</th>
<th>Wilcoxon</th>
<th>pAUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MGST1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>21</td>
<td>5</td>
<td>27</td>
</tr>
<tr>
<td>DF</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>CD33</td>
<td>3</td>
<td>3</td>
<td>8</td>
<td>87</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>CST3</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>TCF3</td>
<td>5</td>
<td>5</td>
<td>11</td>
<td>58</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>MLP</td>
<td>6</td>
<td>7</td>
<td>22</td>
<td>118</td>
<td>8</td>
<td>28</td>
</tr>
<tr>
<td>CSTA</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>18</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>CTSD</td>
<td>8</td>
<td>8</td>
<td>27</td>
<td>144</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>SPTAN1</td>
<td>9</td>
<td>9</td>
<td>19</td>
<td>62</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>CCND3</td>
<td>10</td>
<td>11</td>
<td>17</td>
<td>51</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>PSMA6</td>
<td>20</td>
<td>18</td>
<td>24</td>
<td>63</td>
<td>21</td>
<td>30</td>
</tr>
<tr>
<td>CD63</td>
<td>30</td>
<td>30</td>
<td>46</td>
<td>120</td>
<td>29</td>
<td>158</td>
</tr>
<tr>
<td>FCER1G</td>
<td>40</td>
<td>38</td>
<td>23</td>
<td>29</td>
<td>49</td>
<td>164</td>
</tr>
<tr>
<td>SPI1</td>
<td>50</td>
<td>48</td>
<td>20</td>
<td>10</td>
<td>46</td>
<td>64</td>
</tr>
<tr>
<td>LTC4S</td>
<td>60</td>
<td>63</td>
<td>150</td>
<td>359</td>
<td>105</td>
<td>45</td>
</tr>
</tbody>
</table>

ALL vs AML (Golub et al.)
Which Score is the best one?

That depends on your problem...

Measurement noise of expression differences is Gaussian for all genes …

\[ LR = \bar{X}_1 - \bar{X}_2 \]

Measurements are Gaussian

The average auf Gaussians is Gaussian

The difference of Gaussians is Gaussian

Some fold changes are over estimated and some are underestimated
... but this changes after sorting the fold changes!

<table>
<thead>
<tr>
<th>Gene</th>
<th>Fold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>2.10</td>
</tr>
<tr>
<td>Gene 2</td>
<td>2.08</td>
</tr>
<tr>
<td>Gene 3</td>
<td>1.37</td>
</tr>
<tr>
<td>Gene 4</td>
<td>5.91</td>
</tr>
<tr>
<td>Gene 5</td>
<td>0.92</td>
</tr>
<tr>
<td>Gene 6</td>
<td>2.85</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Fold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank 1</td>
<td>5.91</td>
</tr>
<tr>
<td>Rank 2</td>
<td>2.85</td>
</tr>
<tr>
<td>Rank 3</td>
<td>2.10</td>
</tr>
<tr>
<td>Rank 4</td>
<td>2.08</td>
</tr>
<tr>
<td>Rank 5</td>
<td>1.37</td>
</tr>
<tr>
<td>Rank 6</td>
<td>0.92</td>
</tr>
</tbody>
</table>
Estimation Errors

Genes, for which we **overestimate** the fold change ... *move up* in the ranking

Genes, for which we **underestimate** the fold change ... *go down* in the ranking
Vice Versa

Genes high up in the ranking have most likely overestimated fold changes

Genes far down in the ranking have most likely underestimated fold changes
The noise in rank 1

The noise for a randomly selected gene is centered around zero.

The noise for the top ranking gene is centered around a positive offset.
Extreme Value Distribution

\[ \phi(t) = \theta^{-1} e^{-(t-\xi)/\theta} \exp \left( -e^{\frac{t-\xi}{\theta}} \right) \]

The noise distribution is not only shifted to the right, it also changes its shape from a Gaussian to a Extreme Value Distribution.

Outliers are much more frequent for this type of distribution.
**Screening Noise**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Fold Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank 1</td>
<td>5.91 fold</td>
</tr>
<tr>
<td>Rank 2</td>
<td>2.85 fold</td>
</tr>
<tr>
<td>Rank 3</td>
<td>2.10 fold</td>
</tr>
<tr>
<td>Rank 4</td>
<td>2.08 fold</td>
</tr>
<tr>
<td>Rank 6</td>
<td>1.37 fold</td>
</tr>
<tr>
<td>Rank 7</td>
<td>0.92 fold</td>
</tr>
</tbody>
</table>

Screening for differentially expressed genes:

- Increases noise
- Yields biased fold changes
- Increase the number of noise related outliers
Reproducibility of Rankings

This reproducibility of absolute values translates to …

… this reproducibility of expression differences, which translates to …

… this reproducibility of the ranking of genes
Next Question:

Ok, I chose a score and found a set of candidate genes.

Can I trust the observed expression differences?

➔ Statistical Analysis
Everyone knows that the $p$-value must be below 0.05

0.05 is a holy number both in medicine and biology

... what else should you know about $p$-values
We observe a score $s=1.27$

Can this be just a random fluctuation?

**Assume:** It is a random fluctuation

  = The gene is not differentially expressed

  = The null hypothesis holds

**Theory** gives us the distribution of the score under this assumption

**P-Value:** Probability that a random score is equal or higher to $s=1.27$ in absolute value (two sided test)
Permutations and empirical $p$-values

Target class labels

| 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |

Permuted class labels

| 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| : |
| 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
Rumors

If the gene is not differentially expressed the p-value is high
If the gene is differentially expressed the p-values is low

Both these statements are wrong!
If a gene is not differentially expressed:

The p-value is a random number between 0 and 1!

It is unlikely that such a number is below 0.05 (5% probability)
If a gene is differentially expressed:

The p-value has no meaning, since it was computed under the assumption that the gene is not differentially expressed.

We hope that it is small since the score is high, but there is no theoretical support for this.
Testing only one gene:

If the gene is not differentially expressed a small p-value is unlikely, hence we should be surprised by this observation.

*If we make it a rule that we discard the gene if the p-values is above 0.05, it is unlikely that a random score will pass this filter*

Nevertheless it can still happen and we call this event an *error of the first kind*
Two ways to get 5 Nature publications?

1. Good science and a little bit of luck
2. Fantasy and the error of the first kind

Make up 100 scientific hypotheses all of which should be incorrect but spectacular enough such that Nature would publish them, if one produced significant data to back them up

→ You can expect that about 5 projects will produce p-values < 0.05
Submit those to Nature
Limits of Statistical Significance

Significance tests cannot control the percentage of false published results

The proportion of false claims tested is driving this percentage

Choosing claims to test is not statistics but scientific practice

Corollary 5: The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true.

Corollary 6: The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true.

Ioannidis, PLoS Medicine 2005
Multiple testing with only non-induced genes

1 gene

10 genes

30,000 genes

~1500 p-values
P-values are random numbers between 0 and 1. For only one such number it is unlikely to fall in this small interval, but if we have 30,000 such numbers many will be in there.
If we want to avoid random numbers in this interval we need to make it smaller. The more numbers, the smaller. For 30,000 numbers very small.
How to control the FWER?

Note, that adjusting the interval border can also be done by adjusting the p-values and leaving the cut off at 0.05.

There are many ways to adjust p-values for multiple testing:

Bonferoni: \[ p_{\text{adj}} = p \cdot N \]

Better: Westfall and Young → Exercises
No good idea

In microarray studies controlling the FWER is not a good idea ... It is too conservative.

A different type of error measure became more popular

The False Discovery Rate

What is the idea?
The FDR

1. Score genes and rank them

2. Choose a cutoff

3. Loosely speaking: The FDR is the best guess for the number of false positive genes that score above the cutoff
The FDR refers to a list of genes. The p-value refers to a single gene.

The p-value is based on the assumption that the gene is not differentially expressed, the FDR makes no such assumption.

P-values need to be corrected for multiplicity, FDRs not!
Another difference in concept:

If a 4x change has a small p-value, this means that 4x change is too high to be a random fluctuation

**Conclusion:** 4x change is significant

If a list of 150 genes with 4x change or more has a small estimated FDR this means that we have more genes on this level than would be expected by chance.

**Conclusion:** 4x change can be noise, but 150 genes on that level are too many to be explained just by random fluctuation.

In *p*-value analysis the fold change 4x is significant, in FDR analysis it is the number 150 that is significant.
Histograms of the p-values of all genes on the array
The mixture interpretation of the FDR
**Horizontal vs. Vertical cutoffs**

**FWER:** Vertical cutoff

**FDR:** Horizontal cutoff
The typical plots

Expected random score vs observed scores: Deviations from the main diagonal are evidence for differentially expressed genes
What you typically observe

No differential gene expression
A lot of differential gene expression
Global changes in gene expression
Replications are useful, not only for statistical reasons (5-8 per leg)

Low FWER p-values will lead to many missed genes

FDR (SAM) seems more appropriate

Often there are many induced genes

There are many open questions related to this type of intensive multiple tests
Questions