Normalization

Methods

Affymetrix
- Background correction + expression estimation + summarization
- RMA (Robust Multichip Averaging) uses only PM probes, fits a model to them, and gives out expression values after quantile normalization and median polishing

Agilent
- Background correction + averaging duplicate spots + normalization

Illumina
- Background correction (in GenomeStudio) + normalization

➢ After normalization the expression values are in log2-scale
  - Hence a fold change of 2 means 4-fold up, -2 means 4-fold down, etc
Normalization of Affymetrix data

Preprocessing = background correction + expression estimation + summarisation

Methods: MAS5, Plier, RMA, GCRMA, Li-Wong

- MAS5 is the older Affymetrix method, Plier is a newer one
- RMA is the default, and works rather nicely if you have more than a few chips
- GCRMA is similar to RMA, but takes also GC% content into account
- Li-Wong is the method implemented in dChip

Variance stabilization makes the variance over all the chips similar

- Works only with MAS5 and Plier (the other methods output log2-transformed data, which is thus corrected for the same phenomenon)

Custom chip type

- Because some of the Affymetrix probe-to-transcript mappings are not correct, probes have been remapped in the Bioconductor project. To use these remappings (alt CDF environments), select the matching chip type from the Custom chip type menu.

Normalization of Agilent data

Background treatment often generates negative values, which are coded as missing values after log2-transformation.

- Usual subtract option does this
- Using normexp + offset 50 will not generate negative values, and it gives rather good estimates (the best method reported)

Loess removes curvature from the data (suggested)
Normalization of Agilent data in Chipster

**Background correction**
- **Background treatment**
  - None, Subtract, Edwards, Normexp
- **Background offset**
  - 0 or 50

**Normalize chips**
- None, median, loess

**Normalize genes (not needed for statistical analysis)**
- None, scale (to median), quantile

**Chiptype**
- You must give this information in order to use annotation-based tools later

Normalization tools for Illumina data in Chipster

**Normalization / Illumina**
- **Normalization method**
  - None, scale, quantile, vsn (variance stabilizing normalization)
- **Illumina software version**
  - GenomeStudio or BeadStudio3, BeadStudio2, BeadStudio1
- **Chiptype**
- **Identifier type**
  - Target ID, Probe ID (for BeadStudio version 3 data)

**Normalization / Illumina - Lumi pipeline**
- **Transformation**
  - none, vst (variance stabilizing transformation), log2
- **Normalize chips**
  - none, rsn (robust spline normalization), loess, quantile, vsn
- **Chiptype**
  - human, mouse, rat
- **Background correction**
  - none, bgAdjust.Affy
### Quantile normalization procedure

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sample A</th>
<th>Sample B</th>
<th>Sample C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>20</td>
<td>10</td>
<td>350</td>
</tr>
<tr>
<td>Gene 2</td>
<td>100</td>
<td>500</td>
<td>200</td>
</tr>
<tr>
<td>Gene 3</td>
<td>300</td>
<td>400</td>
<td>30</td>
</tr>
</tbody>
</table>

1. Raw data

<table>
<thead>
<tr>
<th>Quantile 1</th>
<th>Sample A</th>
<th>Sample B</th>
<th>Sample C</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantile 2</td>
<td>100</td>
<td>400</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Quantile 3</td>
<td>300</td>
<td>500</td>
<td>350</td>
<td>350</td>
</tr>
</tbody>
</table>

2. Rank data within sample and calculate median intensity for each row

<table>
<thead>
<tr>
<th>Quantile 1</th>
<th>Sample A</th>
<th>Sample B</th>
<th>Sample C</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantile 2</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Quantile 3</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
</tbody>
</table>

3. Replace the raw data of each row with its median (or mean) intensity

<table>
<thead>
<tr>
<th>Quantile 1</th>
<th>Sample A</th>
<th>Sample B</th>
<th>Sample C</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantile 2</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Quantile 3</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
</tbody>
</table>

4. Restore the original gene order

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Sample A</th>
<th>Sample B</th>
<th>Sample C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 2</td>
<td>20</td>
<td>20</td>
<td>350</td>
</tr>
<tr>
<td>Gene 3</td>
<td>200</td>
<td>350</td>
<td>200</td>
</tr>
</tbody>
</table>

### Checking normalization

[Images of distribution graphs and scatter plots showing normalization check results.]
Exercise 2: Normalize Illumina data

- Import folder IlluminaTeratospermiaHuman6v1_BS1 using the Import tool like in exercise 1.

- In the workflow view, click on the box "13 files" to select all of them.

- Select the tool Normalization / Illumina. Set parameters so that:
  - Illumina software version = BeadStudio1
  - identifier type = TargetID
  - chipType = Human-6v1

- Repeat the run as before, but change:
  - Normalization method = none
  - This "mock-normalization" gives you unnormalized data that can be used as a comparison point when looking at the effect of normalization later.

Describing the experimental setup
Phenodata – describing experiment setup

- Experimental setup is described with a phenodata file, which is created during normalization.

- Fill in the group column with numbers describing your experimental groups:
  - e.g. 1 = healthy control, 2 = cancer sample
  - necessary for the statistical tests to work
  - note that you can sort a column by clicking on its title.

How to describe pairing, replicates, time, etc?

You can add columns to the phenodata file

- Time
  - Use either real time values or recode with group codes

- Replicates
  - All the replicates are coded with the same number

- Pairing
  - Pairs are coded using the same number for each pair
Phenodata for prenormalized data

If you bring in previously created normalized data and phenodata:
- Choose “import directly” in the Import tool
- Right click on normalized data, choose ”Link to phenodata”

If you brought in normalized data and need to create phenodata for it:
- Use Import tool to bring the data in
- Use the tool Normalize / Process prenormalized to create phenodata
  - Remember to give the chiptype
  - Fill in the group column

Exercise 3: Describe the experiment

- Double click the phenodata file
- In the phenodata editor, enter 1 in the group column for the control samples and 2 for the teratospermia samples.
Quality control (inc. clustering theory)

Quality control tools

- **Affymetrix basic (RNA degradation + Affy QC)**
- Affymetrix RLE (relative log expression) and NUSE (normalized unscaled standard error plot) fit a model to expression values
- Illumina (density plot + boxplot)
- Agilent 2-color (MA-plot + density plot + boxplot)
- Agilent 1-color (density plot + boxplot)

**Visualization**
- Dendrogram
- Correlogram

**Statistics**
- Non-metric multidimensional scaling (NMDS)
- Principal components analysis (PCA)
Affymetrix

Quality control information
- Affymetrix QC metrics
- RNA degradation
- Spike-in controls linearity
- RLE (relative log expression)
- NUSE (normalized unscaled standard error plot)

Visualization
- Dendrogram
- Correlogram

Statistics
- Non-metric multidimensional scaling (NMDS)
- Principal component analysis (PCA)

Affymetrix: QC at array level

- QC metrics
  - Average background on the chip
  - % probesets with present flag
  - scaling factors for the chips
  - beta-actin 3'–5' ratio
  - GADPH 3'–5' ratio

Blue area shows where scaling factors are less than 3-fold of the mean.
- If the scaling factors or ratios fall within this region (1.25-fold for GADPH), they are colored blue, otherwise red
Affymetrix: QC at array level

**Spike-in linearity**

**RNA degradation plot**

---

**Affymetrix: QC at array level**

**RLE (relative log expression)**

**NUSE (normalized unscaled standard error)**
Agilent / Illumina: QC at array level

Scatter plot of log intensity ratios $M = \log_2(R/G)$ versus average log intensities $A = \frac{\log_2(R \cdot G)}{2}$, where $R$ and $G$ are the intensities for the sample and control, respectively.

- $M$ is a mnemonic for minus, as $M = \log R - \log G$
- $A$ is mnemonic for add, as $A = (\log R + \log G) / 2$

Agilent QC: MA-plot
All array platforms: QC at experiment level

Visualizations

- Dendrogram
- Correlogram

All platforms: QC at experiment level

Statistics

- Non-metric multidimensional scaling (NMDS)
- Principal component analysis (PCA)
Hierarchical clustering Example, two dimensions

Unsupervised methods Hierarchical clustering

Method parameters

- **Distance method**
  - Euclidean
  - Pearson correlation
  - Spearman correlation
  - Manhattan correlation

- **Drawing method**
  - Single linkage
  - Average linkage
  - Complete linkage
Hierarchical clustering | Distance methods

One can either calculate the *distance* between two pairs of data sets (e.g. samples) or the *similarity* between them.

- **Pearson**
- **Euclidean**

Hierarchical clustering | Distance methods

Can yield very different results

- the Correlation distance
  - red-blue is 0.006
  - red-gray is 0.768
  - blue-gray is 0.7101
- Euclidean distance:
  - red-blue is 9.45
  - red-gray is 10.26
  - blue-gray is 3.29
NOTE: Correlations are VERY sensitive to outliers! (use spearman)

Hierarchical clustering

Distance methods

Hierarchical clustering

Dendrogram drawing

Single, average, and complete linkage
### Clustering methods

#### Hierarchical clustering (euclidean)

1. **Calculate distance matrix**
   
<table>
<thead>
<tr>
<th></th>
<th>gene 1</th>
<th>gene 2</th>
<th>gene 3</th>
<th>gene 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene 1</td>
<td>0</td>
<td>2</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>gene 2</td>
<td>2</td>
<td>0</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>gene 3</td>
<td>8</td>
<td>7</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>gene 4</td>
<td>10</td>
<td>12</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

2. **Calculate averages of most similar**
   
<table>
<thead>
<tr>
<th></th>
<th>gene 1,2</th>
<th>gene 3</th>
<th>gene 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene 1,2</td>
<td>0</td>
<td>7.5</td>
<td>11</td>
</tr>
<tr>
<td>gene 3</td>
<td>7.5</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>gene 4</td>
<td>11</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

3. **Calculate averages of most similar**

<table>
<thead>
<tr>
<th></th>
<th>gene 1,2</th>
<th>gene 3,4</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene 1,2</td>
<td>0</td>
<td>9.25</td>
</tr>
<tr>
<td>gene 3,4</td>
<td>9.25</td>
<td>0</td>
</tr>
</tbody>
</table>

### Clustering methods

#### Hierarchical clustering (avg. linkage)

1. **Calculate distance matrix**
   
<table>
<thead>
<tr>
<th></th>
<th>gene 1</th>
<th>gene 2</th>
<th>gene 3</th>
<th>gene 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene 1</td>
<td>0</td>
<td>2</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>gene 2</td>
<td>2</td>
<td>0</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>gene 3</td>
<td>8</td>
<td>7</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>gene 4</td>
<td>10</td>
<td>12</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

2. **Calculate averages of most similar**
   
<table>
<thead>
<tr>
<th></th>
<th>gene 1,2</th>
<th>gene 3</th>
<th>gene 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene 1,2</td>
<td>0</td>
<td>7.5</td>
<td>11</td>
</tr>
<tr>
<td>gene 3</td>
<td>7.5</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>gene 4</td>
<td>11</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

3. **Calculate averages of most similar**

<table>
<thead>
<tr>
<th></th>
<th>gene 1,2</th>
<th>gene 3,4</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene 1,2</td>
<td>0</td>
<td>9.25</td>
</tr>
<tr>
<td>gene 3,4</td>
<td>9.25</td>
<td>0</td>
</tr>
</tbody>
</table>
Hierarchical clustering

Isomorphism

Look at branching pattern when assessing similarity, not simply the sample (or gene) order!

Cluster evaluation

Principal component analysis, PCA

Goal

Project a high dimensional space into a lower dimensional space

Method

- Compute a variance-covariance matrix for all variables (genes)
- The first principal component is the linear combination of variables that maximizes the variance
- The linear combination, orthogonal to the first, that maximizes variance is the second principal component
- etc.

Illustration
Explains most of the variability in the shape of the pen.

X is the first principal component of the pen.

Y is the second principal component of the pen.

Explains most of the remaining variability in the shape of the pen.
Cluster evaluation | Multidimensional scaling, MDS

**Goal**
Project a high dimensional space into a lower dimensional space

**Method**
- Compute a distance matrix for all variables (genes)
- Define number of dimensions of reduced space
- Construct the dimensions as to maximise the similarity of distances between the high and lower dimensional space

**Compared to PCA**
- Allows choice of distance metric
  => better agreement with clustering methods

---

Cluster evaluation | Principal components analysis

![3D scatter plot showing the first three principal components](image)
Exercise 4: Illumina quality control

- Run **Quality control / Illumina** for the normalized data
- Repeat this for the "mock-normalized" data and compare the results with those for the normalized data (use the Detach button to view the images side by side). Can you see the effect of normalization?

- Run **Statistics / NMDS** for the normalized data
- Run **Visualization / Dendrogram** for the normalized data
- Run **Statistics / PCA** (change the parameter do.pca.on to chips) on the normalized data. View the result as "3D scatter plot for PCA". Can you see 2 groups?
- Save the analysis session with name sessionTeratospermia.zip
Filtering

Gene filtering

Removing probes for genes that are
  • Of low quality
  • Not expressed
  • Not changing

Often a good idea, and reduces the severity of multiple testing correction
  Some controversy on whether filtering should be used or not...

Non-specific filtering
  • Expression, flags, SD, CV, ...

Specific filtering
  • Statistical testing
Non-specific filtering tools in Chipster

**Preprocessing category**
- Filter by standard deviation (SD)
  - Select the percentage of genes to be filtered out
- Filter by coefficient of variation (CV = SD / mean)
  - Select the percentage of genes to be filtered out
- Filter by flag
  - Flag value and number of arrays
- Filter by expression
  - Select the upper and lower cut-offs
  - Select the number of chips required to fulfil this rule
  - Select whether to return genes inside or outside the range
- Filter by interquartile range (IQR)
  - Select the IQR

**Other possibilities:**
1. Utilities / Calculate descriptive statistics
2. Preprocessing / Filter using a column value

Specific filtering

Selecting genes that are associated with some phenotype
(involves statistical testing)

Biologists typically concentrate on fold change (magnitude of effect), statisticians on p-value.
- Both tell a slightly different story. Fold change ignores variability, p-value ignores the size of the effect.
- Take both into account by combining the filters.
  - Filter on expression value (what is biologically significant) and test for differences (what is statistically significant)
Venn diagram

Select 2-3 datasets and the visualization method Venn diagram
- Can use Venn also for filtering with your own gene identifier list

Exercise 5: Filtering

- Select the normalized data and play with different filters. In order to compare the results, set the cutoffs so that you get approximately the same amount of filtered genes (for example 0.9 for SD and CV, and 1.1 for IQR)
  - Preprocessing / Filter by SD
  - Preprocessing / Filter by CV
  - Preprocessing / Filter by IQR

- Select the result files and compare them using the Venn diagram visualization
  - Save the intersection of the three lists as a new dataset
Statistical testing

Statistical analysis of microarray data: Why?

Distinguish measurement variation from treatment effect under study
- replication
- estimation of uncertainty (variability)

Generalisation of results
- representative sample
- statistical inference
**Parametric methods**

Comparing means (1-2 groups)
- student’s t

Comparing means (>2 groups)
- 1-way ANOVA

Comparing means (multifactor)
- 2-way ANOVA

**Non-parametric methods**

Comparing ranks (2 groups)
- Mann-Whitney

Comparing ranks (>2 groups)
- Kruskal-Wallis

<table>
<thead>
<tr>
<th>Ranks</th>
<th>group A</th>
<th>group B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

**Parametric statistics**

\[ t = \frac{X_1 - X_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \]

\[ H_0 : \mu_A = \mu_B, \mu_A - \mu_B = 0 \]
\[ H_1 : \mu_A \neq \mu_B \]

Type 1 error, \( \alpha \)
Type 2 error, \( \beta \)
Power = 1 - \( \beta \)
Non-parametric statistics

Comparing ranks (2 groups)
- Mann-Whitney

Comparing ranks (>2 groups)
- Kruskal-Wallis

Expression

<table>
<thead>
<tr>
<th>Ranks</th>
<th>group A</th>
<th>group B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

\[
U_1 = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1
\]

\[
U_2 = n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} - R_2
\]

Non-parametric compared to parametric tests

Benefits

Do not make any assumptions on data distribution
- robust to outliers
- allows for cross-experiment comparisons

Drawbacks

- Lower power than parametric counterpart
- Granular distribution of calculated statistic
  - many genes get the same rank
  - requires at least 6 samples / group
Multiple testing problem

Problem
1 gene, $\alpha = 0.05$
$\Rightarrow$ false positive incidence = $1 / 20$

30 000 genes, $\alpha = 0.05$
$\Rightarrow$ false positive incidence = 1500

Solution
- Bonferroni
- Holm (step down)
- Westfall & Young
- Benjamini & Hochberg

More false negatives
More false positives

Multiple testing correction methods

Bonferroni
- corrected p-value = $p \times n$
$\Rightarrow \alpha = \alpha / n$

Holm (step down)
- rank p-values
- highest ranked p-value = $p \times n$
- next lower ranked p-value = $p \times (n-1)$
  .
- lowest ranked p-value = $p (n-(n-1)) = p$
Multiple testing correction methods II

Benjamini & Hochberg

- rank p-values from smallest to largest
- largest p-value remains unaltered
- second largest p-value = \( p \times n / (n-1) \)
- third largest p-value = \( p \times n / (n-2) \)

- smallest p-value = \( p \times n / (n-n+1) = p \times n \)

Comparison of multiple testing correction methods

- No correction
- Bonferroni
- FDR
Exercise 6: Statistical testing

- Run different two group tests
  - Select the file sd-filter.tsv
  - Run Statistics / Two group test with the default parameter setting
  - Repeat the run but change the parameter test to t-test and next to Mann-Whitney. Rename the t-test result to t.tsv and the Mann-Whitney-test result to MW.tsv

- Compare the results with a Venn diagram
  - Select the three result files (by keeping the control key down) and select Venn diagram as a visualization method from the pull-down menu
  - Which method seems most powerful?
  - Select the genes common to all three datasets and create a new dataset

- View the Empirical Bayes result as a volcano plot
  - Select the two-sample.tsv and Visualization method volcano plot

- Focus on the prominent changes
  - Select the file sd-filter.tsv and run the tool Preprocessing / Filter using a column value so that you keep genes that have fold change higher than +/-3 (select FC column, cutoff = 3, and "outside"). View the result file as expression profile.

How to improve statistical power?

More accurate estimates of effect size and variability

Experimentally

- Increase number of replicates (biological)
- Randomization
- Blocking

Analytically

- Variance shrinking (Ebayes)
- Partitioning variability (ANOVA, linear modeling)
Improving power with variance shrinking

Concept
Borrow information from other genes with similar expression level and form a pooled error estimate

How?
- model the error-intensity dependence based on replicate to replicate comparisons
- use a smoothing function to estimate the error for any given intensity
- calculate a weighted average between observed gene specific variance and model-derived variance (pooling)
- incorporate the pooled variance estimate in the statistical test (usually t-or F-test)

Linear modeling: Pairing

<table>
<thead>
<tr>
<th>Unpaired</th>
<th>Experimental Group 1</th>
<th>Experimental Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Paired</th>
<th>Experimental Group 1</th>
<th>Experimental Group 2</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

One sample T-test
### Linear modeling: Multiple factors

#### 1 factor

<table>
<thead>
<tr>
<th>Experimental Group 1</th>
<th>Experimental Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>5</strong></td>
</tr>
</tbody>
</table>

#### 2 factors

<table>
<thead>
<tr>
<th></th>
<th>Experimental Group 1</th>
<th>Experimental Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Males</strong></td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td><strong>Females</strong></td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>8</td>
<td>4</td>
</tr>
</tbody>
</table>

### Linear modeling: Interaction effect

![Graph showing interaction effect]
Linear modeling in Chipster

- $Y = a + bx_1 + cx_2 + dx_1x_2$
  - Like a normal multiple regression
  - Intercept (a) is included by default
  - Can contain both main effects (b, c) and interaction effects (d)
- **Linear modeling in Chipster can take into account at most three main effects, their interactions, one technical replication level, and one level of pairing**
  - This is enough for all the experiments I’ve encountered in GEO so far.
  - Technical replication: one biological sample is hybridized on more than one array
  - Pairing: before-after –type of setting. Measurements available prior to treatment and after it from exactly the same cell culture flasks.

Linear modeling: Setting up the model

- **All columns (max. three) in the phenodata can be either tested as linear (is there a trend towards higher numbers?) or as a factor (are there differences between the groups?).**
  - With 2 groups there’s no difference in these settings.
Gene annotation = information about biological function, pathway involvement, chromosomal location etc

Annotation information is collected from different biological databases to a single database by the Bioconductor project

Annotation information is required by certain analysis tools (annotation, GO/KEGG enrichment, promoter analysis, chromosomal plots)

- These tools don’t work for those chiptypes which don’t have Bioconductor annotation packages
Alternative CDF environments for Affymetrix

- CDF is a file that links individual probes to gene transcripts (probesets)
- Affymetrix default annotation uses old CDF files that map many probes to wrong genes
- Alternative CDFs fix this problem
- In Chipster
  - selecting "custom chiptype" in Affymetrix normalization takes altCDFs to use
- For more information see
  - http://brainarray.mbnl.med.umich.edu/Brainarray/Database/CustomCDF/genomic_curated_CDF.asp
Also a problem with Illumina

- Probes are remapped in the R/Bioconductor project
- Chipster uses remapped probes
- For more information see
  - https://prod.bioinformatics.northwestern.edu/nulD/

Exercise 7: Annotation

- Annotate genes
  - Select the file column-value-filter.tsv
  - Run Annotation / Illumina gene list
  - Open the result file annotations.html and click the links in the gene and pathway columns to read more about one of the genes
  - Open the result file annotations.tsv and detach it. Sort it by the pathway column. Slide the pathway column next to the description column and make it wider.
Pathway analysis

Pathway analysis: traditional method
Pathway analysis: traditional method

- Identify differentially expressed genes
- Categorize into functional groups
  - Gene ontologies
  - KEGG pathways
  - Reactome pathways
  - Protein families
- Categorize into other meaningful groups
  - Genomic location: chromosome, cytoband, bp window
  - Disease specific
  - Tissue specific
- Assess degree of enrichment
  - Hypergeometric test
  - Fisher’s exact test
  - Chi-square test
Enrichment analysis:

Differentially expressed genes (50)

Genome, array (30000)

Apoptosis (200)

Differentially expressed and apoptosis (10)

\[ H_0: \frac{10}{50} \frac{200}{30000} = 30 \gg 1 \]

Pathway analysis: Gene set test
Pathway analysis: Gene set test

- Categorize into functional groups
  - Gene ontologies
  - KEGG pathways
  - Reactome pathways
  - Protein families
- Or categorize into other meaningful groups
  - Genomic location: chromosome, cytoband, bp window
  - Disease specific
  - Tissue specific

- Assess differential expression for gene sets
  - Parametric tests
  - Rank-based tests
  - Co-regulation vs. regulation
Pathway analysis: Gene set test

**Benefits**
- Significantly improved sensitivity over single gene tests
- Relative insensitivity to outliers -> no or little filtering
- Reduced number of tests -> less multiple testing correction -> increased power
- Potentially more meaningful interpretation of results

**Downsides**
- Difficult to assess the importance of each individual gene to the overall pathway behavior
- Quality of results limited by quality of gene sets
- Gene ontologies present a few challenges to the analysis and the interpretation of results due to hierarchical acyclic structure

---

Exercise 8: Mining for biological significance

**Identify over-represented GO terms**
- Select the `two-sample.tsv` file
- Select Pathways / Hypergeometric test for GO and make sure that `biological_process`, 0.05, 5, yes, none and over are specified as parameters.

**Extract genes for specific GO term**
- Open the `hypergeo.tsv` file and copy the GO id number for the top term.
- Select `two-sample.tsv` and run tool Utilities / Extract genes from GO, pasting the GO id into the parameter field.
- Open the `extracted-from-GO.tsv` in Spreadsheet and Expression profile view.

**Identify over-represented ConsensusPathDB pathways**
- Select `two-sample.tsv` and run tool and Pathways / Hypergeometric test for ConsensusPathDB. Make sure that 0.05 and genes are specified as parameters.
- Click on the hyperlinks in the `cpdb.html` tile to get more info on a particular pathway.
Exercise 9: Gene set test

- Identify differentially expressed KEGG pathways
  - Select the normalized.tsv file and Pathways / Gene set test. Make sure that group, KEGG, yes, 4, 600 and 600 are specified as parameters.
  - Explore the results both in tabular format and graphically in the global-test-result-table.tsv and multtest.png files, respectively.

Automatic analysis workflows
Workflow – reusing and sharing your analysis pipeline

- Chipster allows you to save your analysis workflow as a "macro", which can be applied to another normalized dataset.
- All the analysis steps and their parameters are saved as a script file to your computer.
- You can share the workflow file with a collaborator.
- In addition to user-made workflows, Chipster contains ready-made workflows for finding and analyzing differentially expressed genes, miRNAs or proteins.

Saving and using workflows

- After completing your analysis, select the starting point for your workflow and click "Workflow/ Save starting from selected".
- You can save the workflow file anywhere on your computer and change its name, but the ending must be .bsh.
- You can apply the workflow to another normalized dataset by selecting:
  - Workflow->Open and run
  - Workflow->Run recent (if you saved the workflow recently).
Changing the workflow file

You can change parameters directly to the workflow file

Exercise 10: Saving a workflow

- **Save a workflow**
  - Prune your workflow if necessary (remove cyclic structures)
  - Select the file normalized.tsv and click on the Workflow / Save starting from selected. Give your workflow a meaningful name and save it.