Analysis of Microarray Data

Lecture 1:
Experimental Design and Data Normalization

George Bell, Ph.D.
Senior Bioinformatics Scientist
Bioinformatics and Research Computing
Whitehead Institute
Outline

• Introduction to microarrays
• Experimental design
• Data normalization
• Other data transformation
• Exercises
Expression microarrays: Underlying assumption and concepts

• Measuring relative changes in levels of specific mRNAs provide information about what’s going on in the cells from which the mRNA came.

Samples provide info about Genes

• A gene expression profile is a molecular phenotype of a cell in a specific state
Experimental design: Most important question

• Why are you doing this experiment?

(Be as specific as possible.)

“To learn something interesting about my cells” is usually not the best answer.
Common partial experimental objectives

Comparison: identify differentially expressed genes

Discovery: identify clusters of genes or samples

Prediction: use a gene expression profile to label a cell sample
General experimental issues

• What is the best source of mRNA?
• Reduce variables as much as possible
• Avoid confounding by randomizing remaining variables
• Collect comprehensive information about all potential variables
• Make no more assumptions than necessary
• Does a factor influence your measurements? Collect the data and find out with ANOVA.
Comparisons

- Virtually all array analysis depends on a comparison between samples (on 3+ chips)
- Expression is usually described in relative terms
- What comparison(s) do you plan to make?
- Research in progress: How can one measure absolute (molar) expression levels?

*Spike-in controls*?
**Replication**

- **Biological** replicates: use different cell cultures prepared in parallel
- **Technical** replicates: use one cell culture, first processed and then split just before hybridization
- **Sample** replicates: use one cell culture, first split and then processed

* most informative
How many replicates?

• Replication is needed to have confidence about your results.
• To determine the optimal number using statistics,
  – How large an effect do you want to identify?
  – How confident do you want to be of your conclusions?
  – How variable is gene expression in your system?
  – Perform a test of statistical power (such as ‘power.t.test’ in R)
• Most common practical answer: More than you’ve planned
• If microarray analysis is followed by further confirmation, a high error rate may be tolerated (and may be more efficient)
Designs for 2-color arrays

Given two replicates of samples A and B,

• **Reference** design
  

• **Balanced block** design
  
  A1-B1  B2-A2

• **Loop** design
  
What design to use?

• Best design depends on objective(s) of experiment
• What comparisons are most important?
• Some guidelines:
  – Balanced block is most efficient for 2-way comparison
  – Reference design is often best when making lots of different comparisons
  – Loop design is not very robust
Spike-in controls

• How can you confirm that your experiment and analysis was done correctly?
• Control mRNA added before hybridization (or RNA extraction) can help with quality control
• Some chip manufacturers recommend a control mix of exogenous mRNA
• External RNA Control Consortium (ERCC): determining optimal control mix to evaluate "reproducibility, sensitivity, and robustness in gene expression analysis"
Image analysis

- Map region of the chip to a probe and convert its pixels into foreground and background intensities for the spot

- This is a crucial step in the analysis pipeline – but will not be covered in this course

- What instruments and algorithms are recommended by the chip manufacturer?
Why normalize data?

- The experimental goal is to identify biological variation (expression changes between samples)
- Technical variation can hide the real data
- Unavoidable systematic bias should be recognized and corrected – the process referred to as normalization
- Normalization is necessary to effectively make comparisons between chips – and sometimes within a single chip
Normalization assumptions and approaches

• Some genes exhibit constant mRNA levels:
  – Housekeeping genes

• The level of some mRNAs are known:
  – Spike-in controls

• The total of all mRNA remains constant:
  – Global median and mean; Lowess

• The distribution of expression levels is constant
  – quantile
Within-array normalization

- 2-color arrays may need normalization between red and green channels
- These methods are similar to between-array methods
- What should be normalized?
  - Red intensities vs green intensities?
    - Global mean/median
  - Log ratio vs average intensity?
    - Linear regression or loess
- Within-array may be followed by between-array methods
Normalization by global mean (total intensity)

- Procedure: Multiply/divide all expression values for one color (or chip if one-color) by a factor calculated to produce a constant mean (or total intensity) for every color.
- Example with 2 one-color arrays with a total intensity target of 50,000:

<table>
<thead>
<tr>
<th>Chip</th>
<th>Sample gene expr (raw)</th>
<th>Total expr on chip (raw)</th>
<th>Norm. factor ( \text{tot}<em>{\text{des}} / \text{tot}</em>{\text{obs}} )</th>
<th>Sample gene expr (norm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2.0</td>
<td>100,000</td>
<td>( \frac{50,000}{100,000} = 0.5 )</td>
<td>2.0 x 0.5 = 1.000</td>
</tr>
<tr>
<td>B</td>
<td>2.2</td>
<td>125,000</td>
<td>( \frac{50,000}{125,000} = 0.4 )</td>
<td>2.2 x 0.4 = 0.88</td>
</tr>
</tbody>
</table>

- Similar scheme can be used with a subset of genes such as with spike-in controls or housekeeping genes.
Global median normalization

- Procedure: Transform all expression values to produce a constant median
- More robust than using the mean
Variance normalization

- Different chips may have the same median or mean but still very different standard deviations.
- If we assume the chips should have common standard deviations, they may be transformed in that manner.
Quantile normalization

- Different chips may have the same standard deviation but different distributions.
- If we assume the chips should have common distributions, they may be transformed in that manner.

![Box plots comparing raw data and normalized data for Chip A and Chip B.](image)
Lowess normalization

- Some 2-color arrays exhibit a systematic intensity-dependent bias
- As a result, the normalization factor needs to change with spot intensity
- Lowess (locally weighted scatterplot smoothing) uses local regression to address this

Quackenbush, 2002
Local normalization

• Sometimes global within-array normalization may not correct all systematic unwanted variation

• Examples: print tip differences, degradation in chip regions, thumbprints

• Local normalization adjusts intensities according to chip geography

• It’s best to avoid technologies that require these “excessive” transformations
Normalization - summary

- Normalization removes technical variation and improves power of comparisons
- The assumption(s) you make determine the normalization technique to use
- Always look at all the data before and after normalization
- Spike-in controls can help show which method may be best
Handling low-level values

- What is the background intensity of the chip?
- What expression values are just noise?
- Filtering / flagging low values
- Settings floors and ceilings
- Effects on fold changes and determination of differential expression
Affymetrix preprocessing

- Some oligo chip designs (like Affymetrix) represent each gene (“probeset”) with a set of oligos (“probes”).
- Affymetrix software (MAS) uses a special algorithm to convert measurements for a set of probes into one probeset value.
- Other algorithms (RMA, GCRMA, MBEI) have been developed by people who want to improve this calculation.
- These other algorithms appear to increase precision but decrease dynamic range.
Why use logarithms?

• Produce similar scales for fold changes in both the up and down directions

• Since $\log{(a*b)} = \log{(a)} + \log{(b)}$
  – Multiplicative effects are converted to additive effects, which simplifies statistical analysis

• Produce data with
  – a more normal distribution
  – variability that’s not correlated with intensity
Summary

• Why are you doing a microarray experiment?
• What design will best help address your goal(s)?
• Normalize based on the biology and technology of the experiment
• Other transformations: preprocessing, dealing with low level values; logarithms
• Does your analysis pipeline make sense biologically and statistically?
References

• Li C and Wong WH. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol. 2(8), 2001 [MBEI]
• Wu Z and Irizarry RA. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. Proceedings of RECOMB ’04. [GCRMA]
Microarray tools

• Bioconductor (R statistics package)
  – http://www.bioconductor.org/

• BaRC analysis tools:

• Excel

• TIGR MultiExperiment Viewer (MeV)

• Many commercial and open source packages
## Exercise 1 - Excel syntax

<table>
<thead>
<tr>
<th>Reference</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2</td>
<td>Cell reference</td>
</tr>
<tr>
<td>A2:A100</td>
<td>Series of cells</td>
</tr>
<tr>
<td>=B5</td>
<td>Formula</td>
</tr>
<tr>
<td>=$B$5</td>
<td>Absolute link (‘$’)</td>
</tr>
<tr>
<td>=data!B4</td>
<td>Reference other sheet</td>
</tr>
<tr>
<td>=[otherFile.xls]data!B4</td>
<td>Reference other file</td>
</tr>
</tbody>
</table>
Exercise 1: Excel functions

• MEDIAN
• SUM
• AVERAGE
• TRIMMEAN
• LOG
• IF
• TTEST
• VLOOKUP
The R Project for Statistical Computing

Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.
- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News:

- **R version 2.4.0** has been released on 2006-10-03.
- **R News 6/4** has been published on 2006-10-31.
- The **R Wiki** provides an online forum where users can help other users.
Introduction to R

# Read a data file
dat = read.delim("Data1.txt")
dim(dat)  # Get dimension of matrix
colnames(dat)  # Get names of columns
# Print rows 1-5, columns 2-4
dat[1:5, 2:4]  # or use column and row names
mean(dat[, "my.col.1"])}  # get the mean of a column
# Combine data by columns
all.data = cbind(data1, data2)
# Print a tab-delimited text file
write.table(all.data, "myFile.txt", sep="\t", quote=F)
q()  # quit [or use pull-down menu]
Exercise 1 - To do

Goal: Discovery of human developmentally-regulated genes

- Fetal vs adult; liver vs brain; assayed with Affymetrix chips
- Normalize data - 8 chips (replicates)
  - Normalization by trimmed means
  - \( k = \frac{\text{expression signal}}{\text{chip trimmed mean}} \times 100 \)
- Calculate ratios
  - Reduce data (replicates)
  - Use AVERAGE function
  - Ratio of fetal tissue/adult tissue
- Calculate \( \log_2 \) of expression values and ratios