

Analysis of Microarray Data

Lecture 1: Experimental Design and Data Normalization

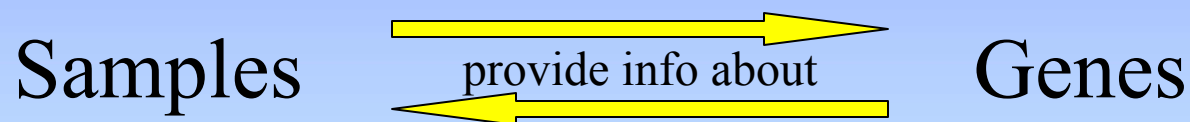
George Bell, Ph.D.
Bioinformatics Scientist
Bioinformatics and Research Computing
Whitehead Institute

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.



- A gene expression profile is a molecular phenotype of a cell in a specific state

Most important design 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.

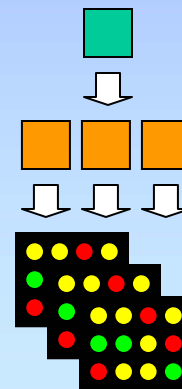
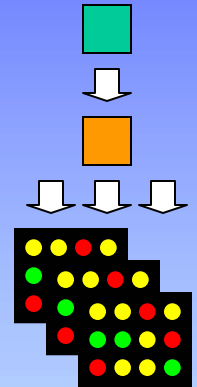
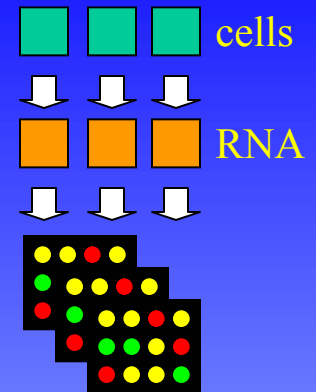
Comparisons

- Virtually all array analysis depends on a comparison between samples (on 2+ chips)
- Expression is usually described in relative terms
- What comparison(s) do you plan to make?
- Ongoing research: How can one measure absolute expression levels (molar)?

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



How many replicates?

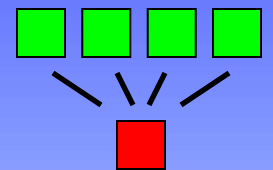
- Most common answer: More than you've planned
- To determine the optimal number using statistics,
 - consider the False Discovery Rate (FDR)
 - What proportion of false positives can you tolerate?
- If microarray analysis is followed by further confirmation, a high FDR may be tolerated (and may be more efficient)

Designs for 2-color arrays

Given two replicates of samples A and B,

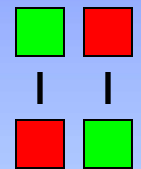
- **Reference design**

A1-R A2-R B1-R B2-R



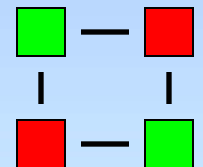
- **Balanced block design**

A1-B1 B2-A2



- **Loop design**

A1-B1 B1-A2 A2-B2 B2-A1



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

Normalization by global mean (total intensity)

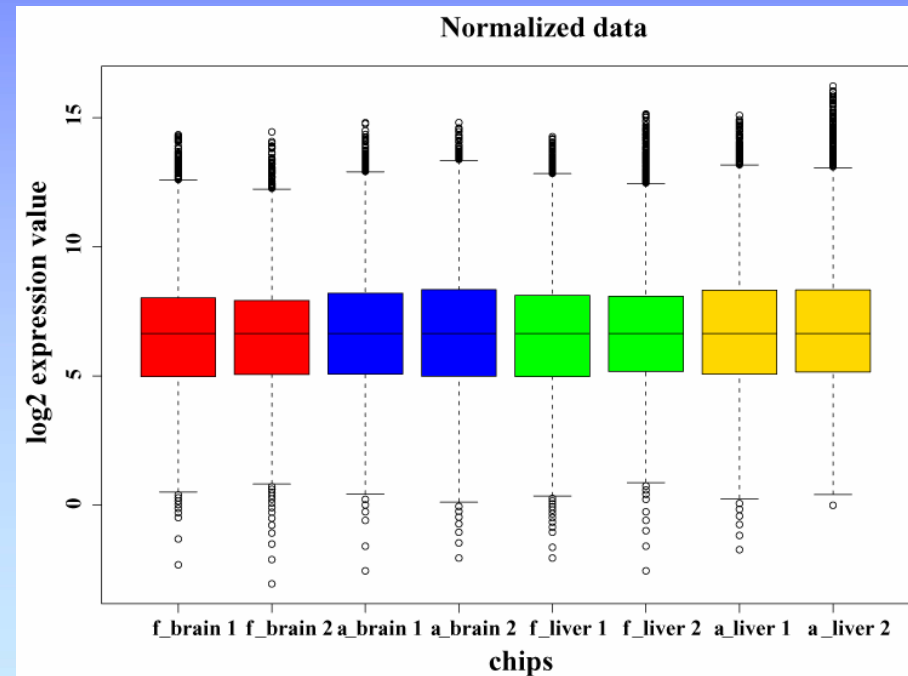
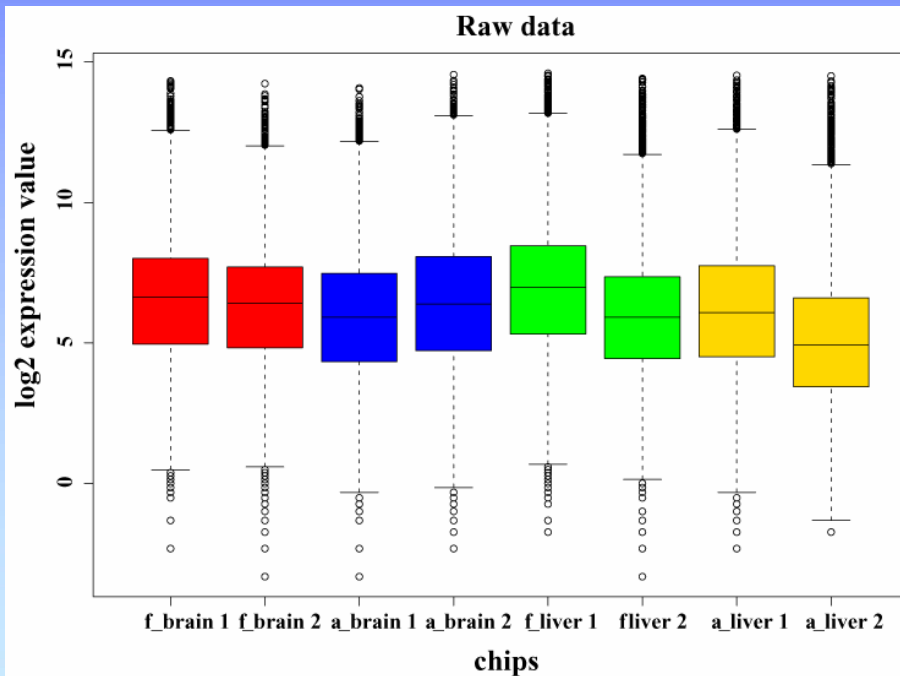
- Procedure: Multiply/divide all expression values for one color (or chip if one-color) by a factor calculated to produce a constant mean (or total intensity) for every color.
- Example with 2 one-color arrays:

Chip	Mean expr (raw)	Total expr (raw)	Norm. factor	Mean expr (norm)	Total expr (norm)
A	2.000	100,000	0.5	1.000	50,000
B	2.200	110,000	0.4545	1.000	50,000

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

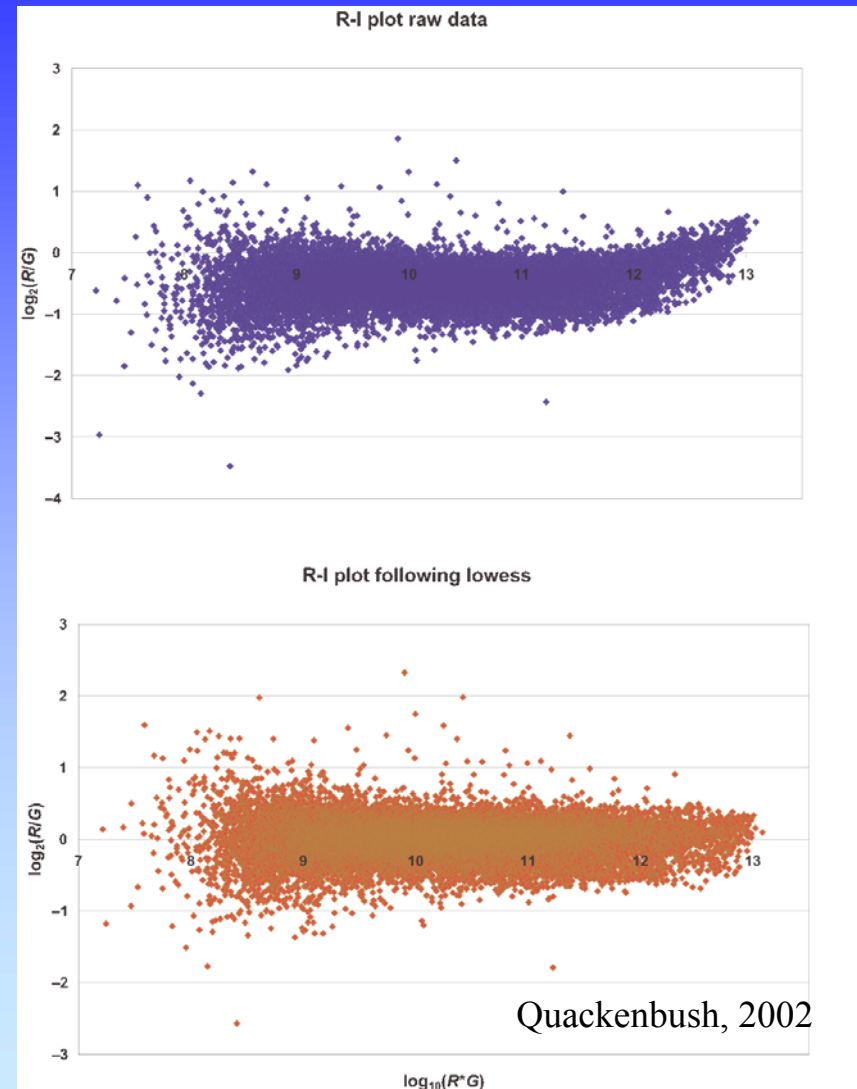
Global median normalization

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



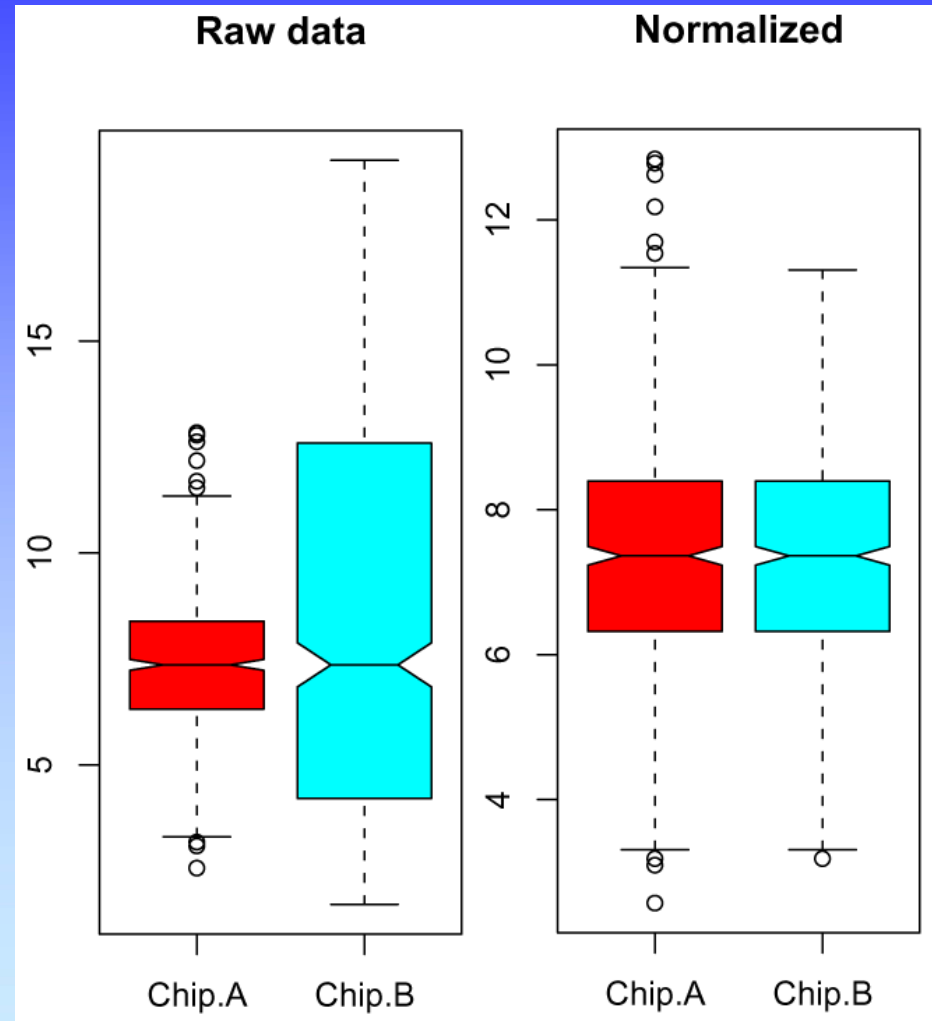
Lowess normalization

- Some 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



Quantile normalization

- Different chips may have the same median or mean but still very different distributions
- Assuming the chips have a common distribution of intensities, they may be transformed to produce similar distributions



Local normalization

- Sometimes normalization is required before one can compare expression values even within a chip
- 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

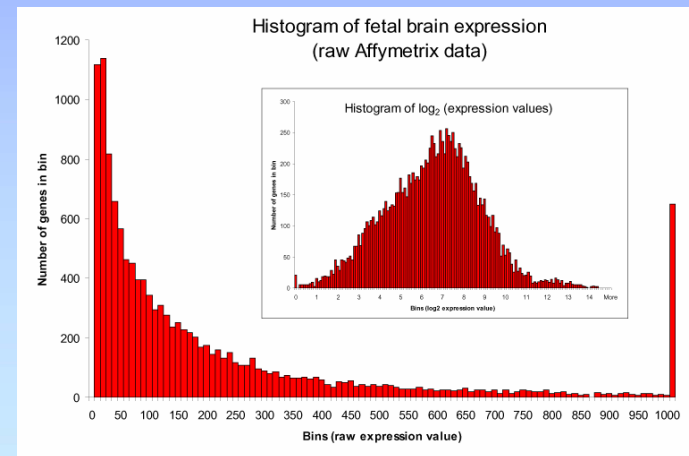
- What is the background intensity of the chip?
- What expression values are just noise?
- Filtering / flagging low values
- Settings floors and ceilings
- Effects on fold changes and determination of differential expression

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, GC-RMA, 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
- Since most genes exhibit low expression, logarithms produce a more normal distribution of expression levels



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

- Causton HC et al. Microarray Gene Expression Data Analysis: A Beginner's Guide. Blackwell, 2003.
- Churchill, GA. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics Supp.* 32:490-495, 2002.
- Quackenbush J. Microarray data normalization and transformation. *Nature Genetics Supp.* 32:496-501, 2002.
- Smyth GK et al. Statistical issues in cDNA microarray data analysis. *Methods Mol Biol.* 224:111-36, 2003.
- Affymetrix. Statistical Algorithms Description Document.
http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf
- Irizarry RA et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4(2):249-64, 2003. [RMA]
- 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]

Microarray tools

- BaRC analysis tools:
 - http://iona.wi.mit.edu/bio/tools/bioc_tools.html
- Bioconductor (R statistics package)
 - <http://www.bioconductor.org/>
- Excel
- Many commercial and open source packages

Exercise 1 - Excel syntax

A2	Cell reference
A2:A100	Series of cells
=B5	Formula
=\$B\$5	Absolute link ('\$')
=data!B4	Reference other sheet
=[otherFile.xls]data!B4	Reference other file

Exercise 1: Excel functions

- MEDIAN
- SUM
- AVERAGE
- LOG
- IF
- TTEST
- VLOOKUP

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)
 - Global median normalization
 - $(\text{expression signal} / \text{chip median value}) * 100$
- Calculate ratios
 - Reduce data (replicates)
 - Use AVERAGE function
 - Ratio of fetal tissue/adult tissue
- Calculate \log_2 of expression values and ratios