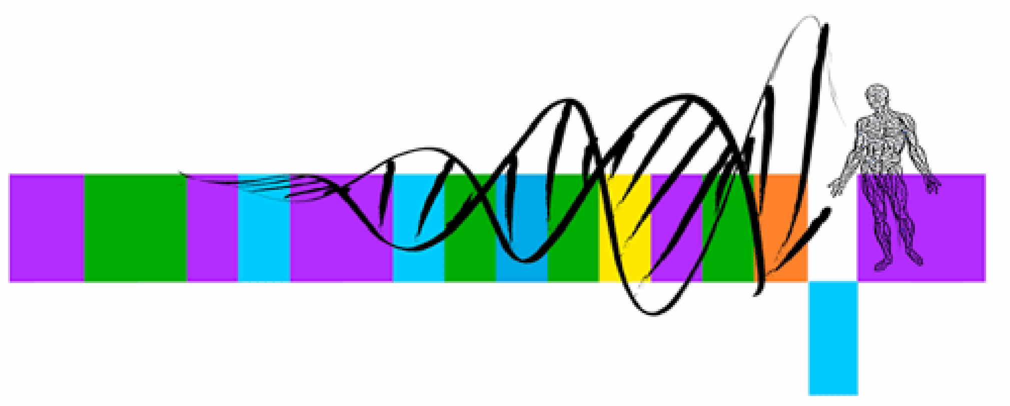


Analysis of Microarray Data

Lecture 1:

Experimental Design and Data Normalization

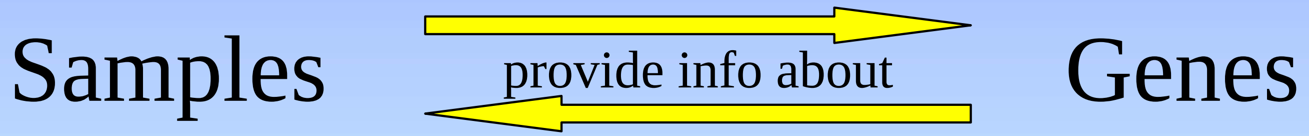
George Bell, Ph.D.
Senior Bioinformatics Scientist
Bioinformatics and Research Computing
Whitehead Institute

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.

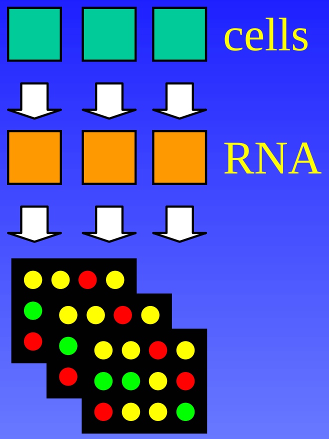
Comparisons

- 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?

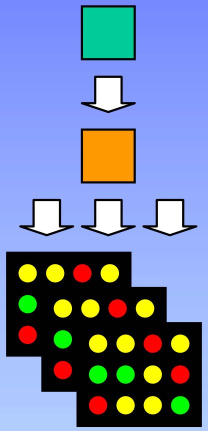
  Spike-in controls?

Replication

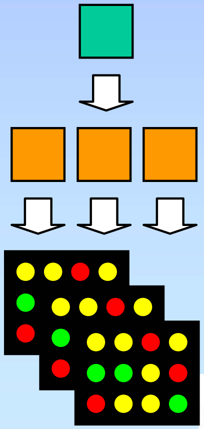
- ***Biological** replicates: use different cell cultures prepared in parallel



- **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?

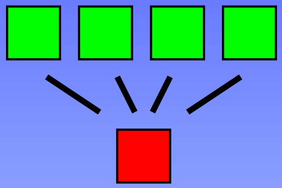
- Replication is needed to have confidence about your results.
- To determine the optimal number using statistics,
  - How large an effect do you want to identify?
  - How confident do you want to be of your conclusions?
  - How variable is gene expression in your system?
  - Perform a test of statistical 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,

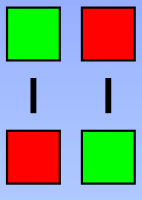
- **Reference** design

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



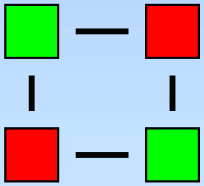
- **Balanced block** design

  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

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)

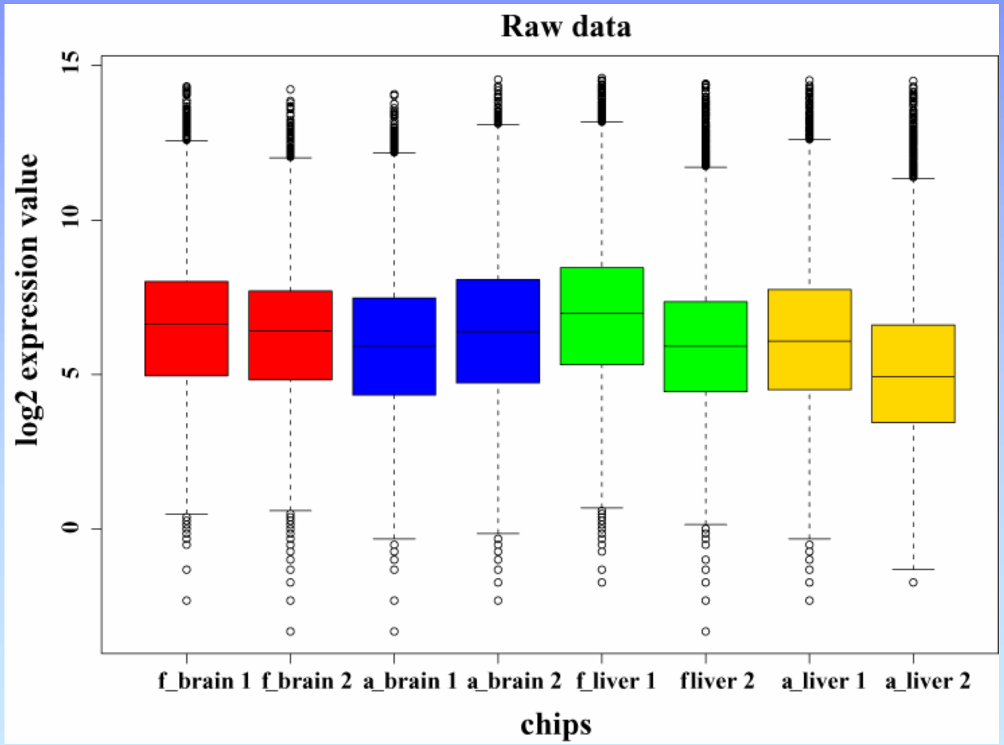
- Procedure: Multiply/divide all expression values for one color (or chip if one-color) by a factor calculated to produce a constant mean (or total intensity) for every color.
- Example with 2 one-color arrays with a total intensity target of 50,000:

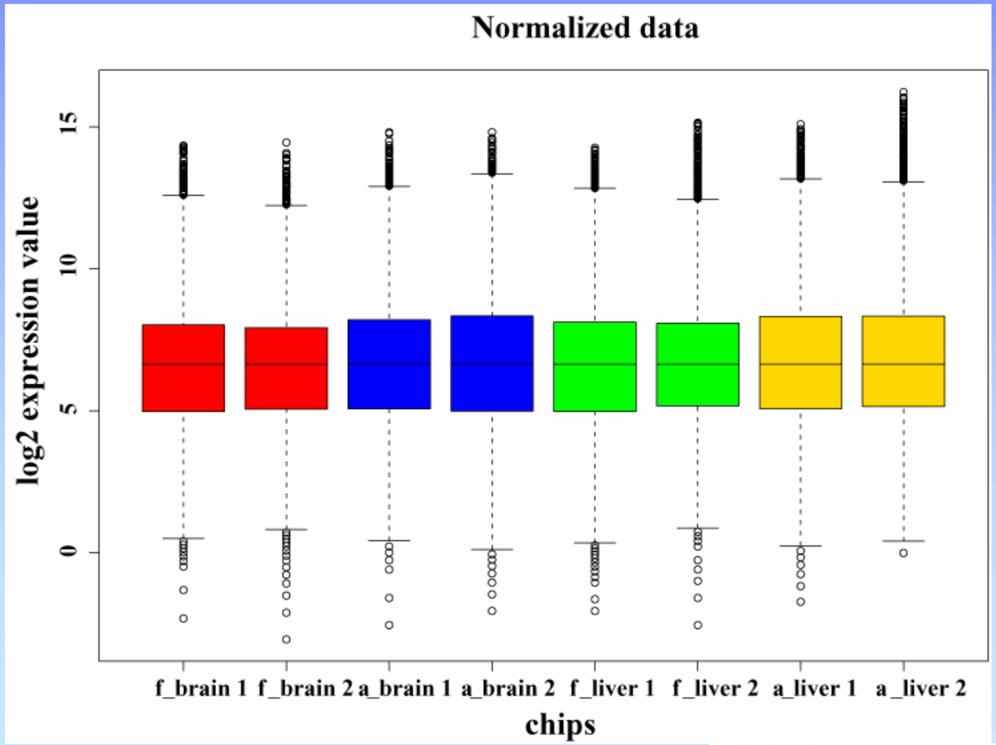
Chip	Sample gene expr (raw)	Total expr on chip (raw)	**Norm. factor (tot_{des} / tot_{obs})**	Sample gene expr (norm)
A	2.0	100,000	**50,000 / 100,000 =** **0.5**	**2.0 x 0.5 =** 1.000
B	2.2	125,000	**50,000 / 125,000 =** **0.4**	**2.2 x 0.4 =** 0.88

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

Global median normalization

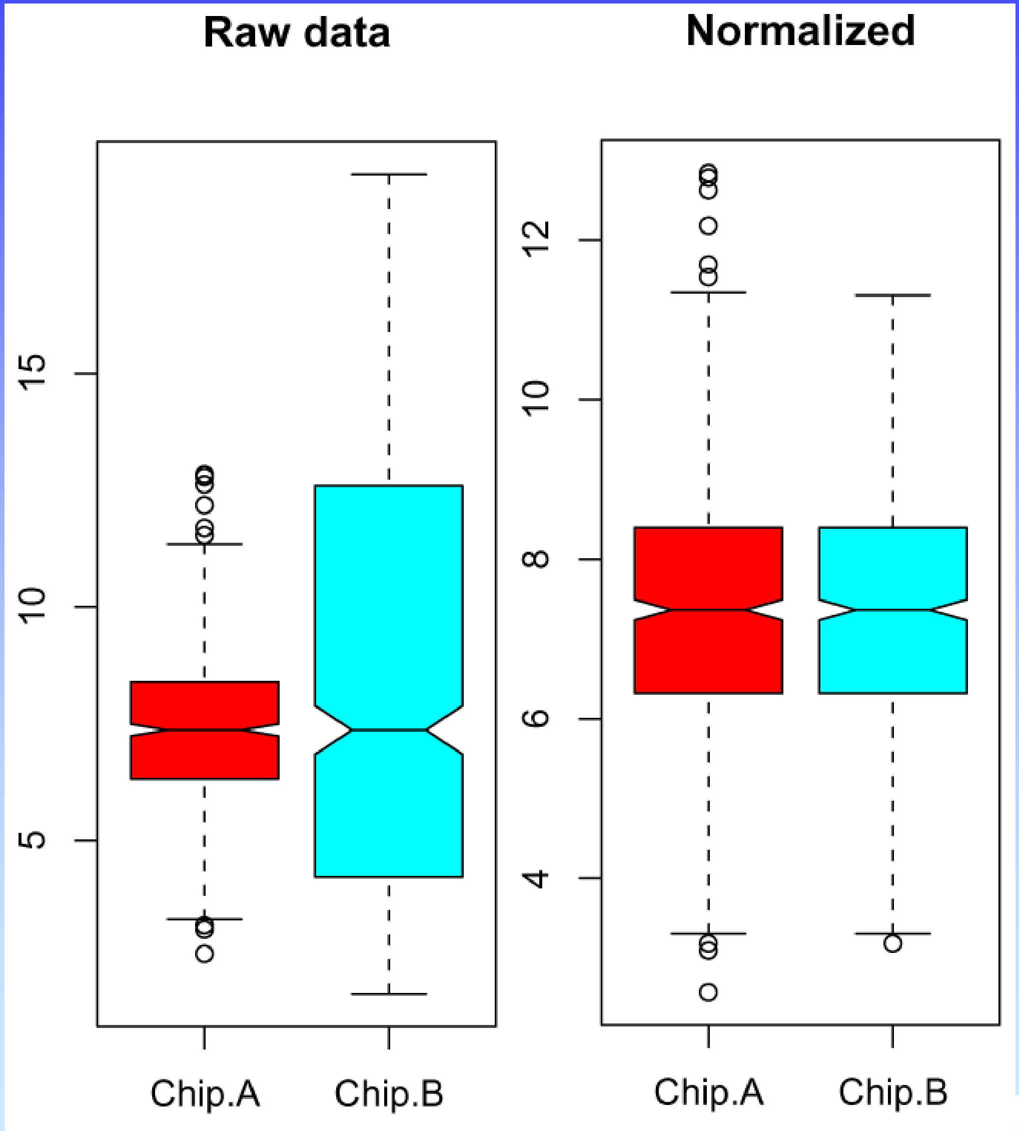
- Procedure: Transform all expression values to produce a constant median
- More robust than using the mean





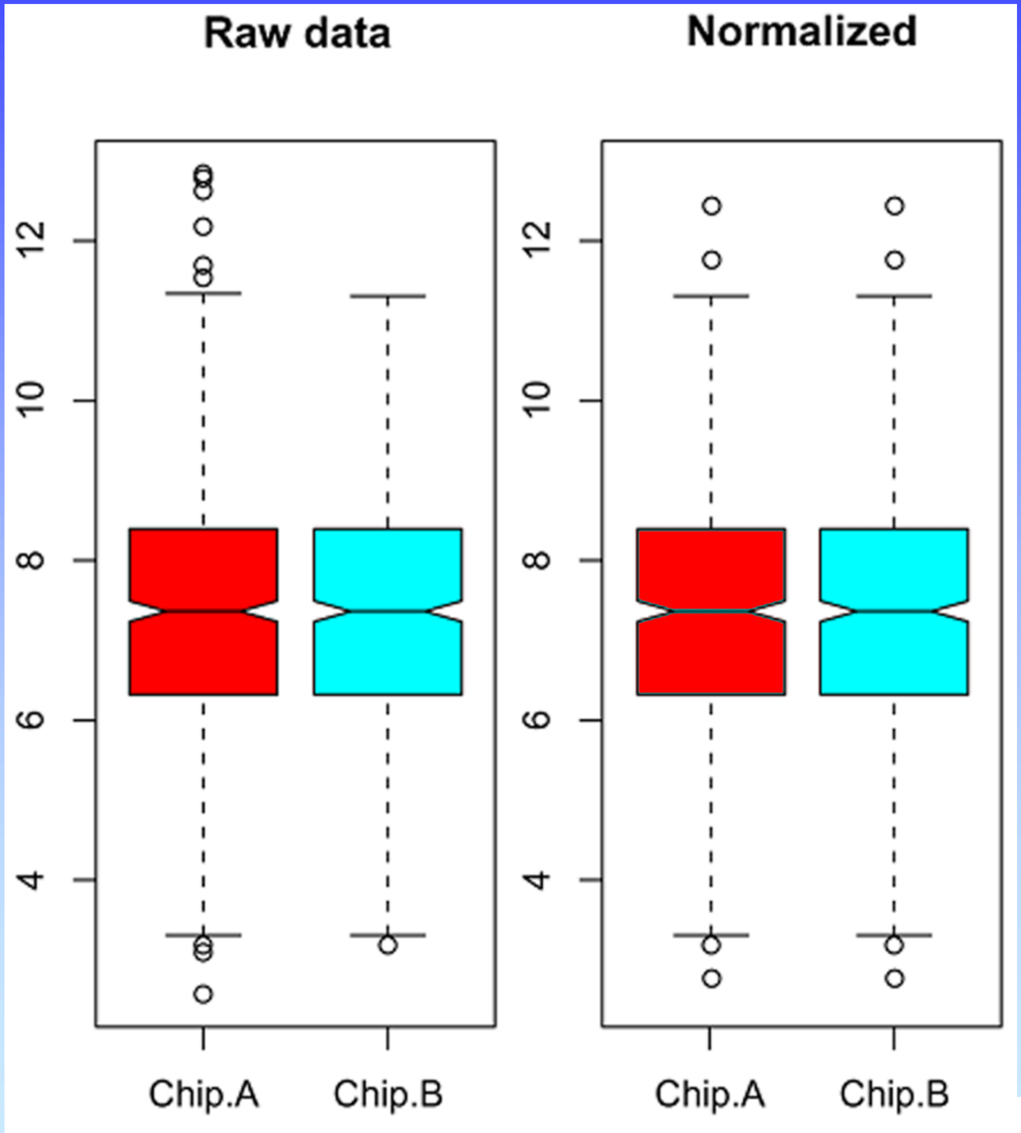
Variance normalization

- 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



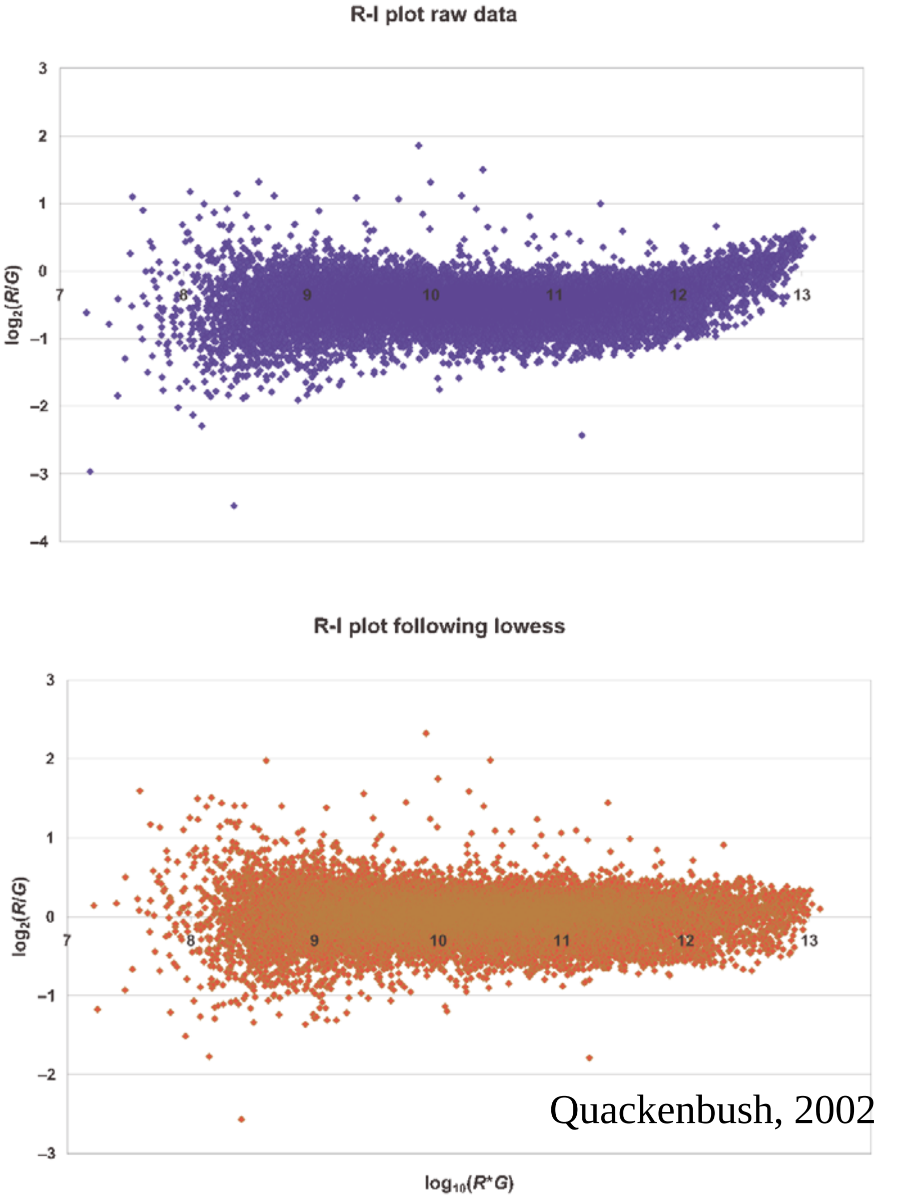
Quantile normalization

- 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



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

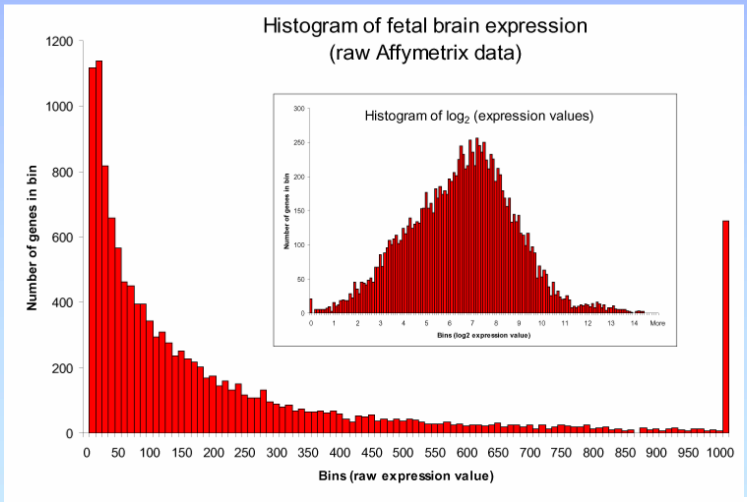
- What is the background intensity of the chip?
- What expression values are just noise?
- Filtering / flagging low values
- Settings floors and ceilings
- Effects on fold changes and determination of differential expression

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

- Dov Stekel. Microarray Bioinformatics. Cambridge, 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]
- Wu Z and Irizarry RA. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. Proceedings of RECOMB ’04. [GCRMA]

Microarray tools

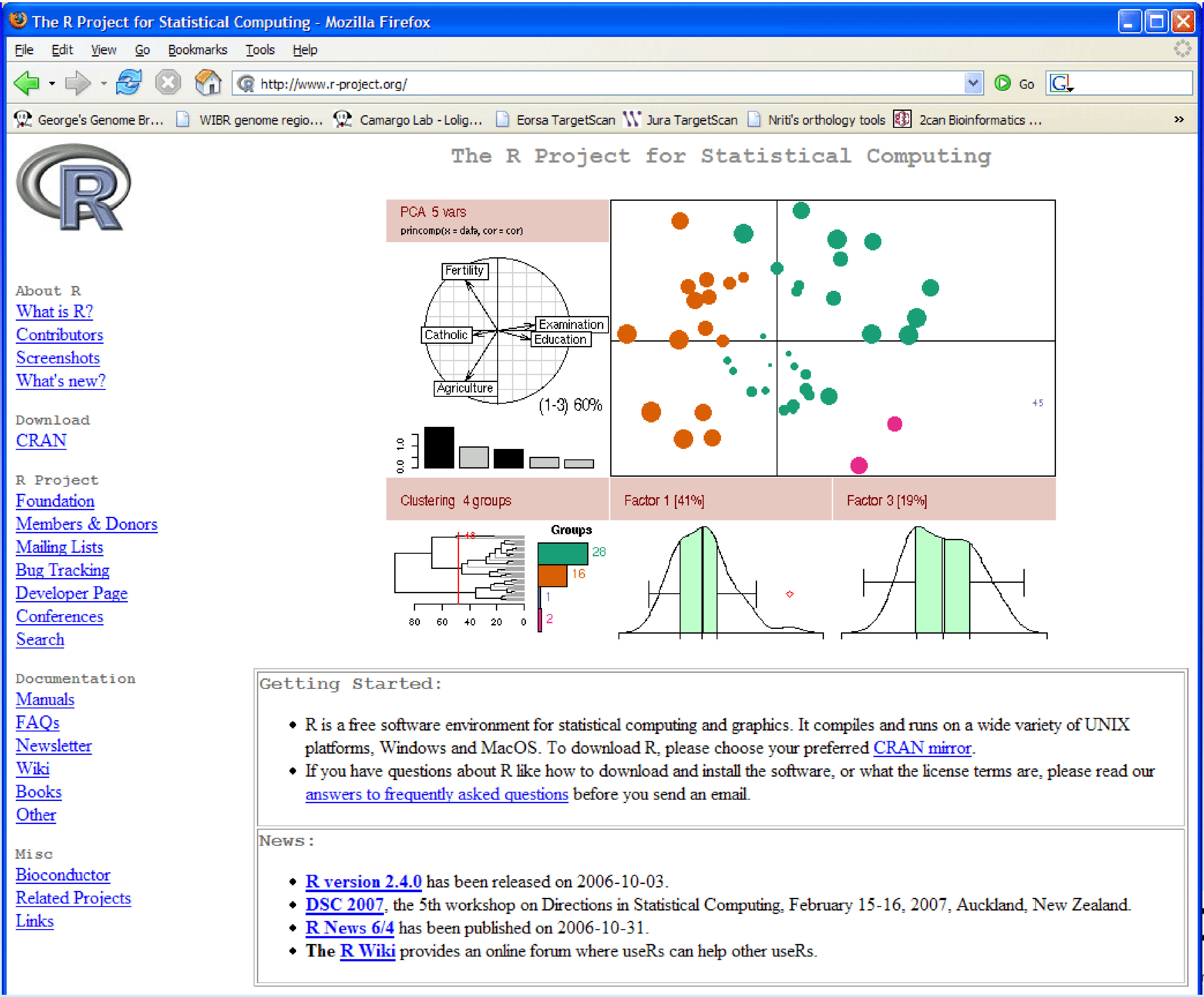
- Bioconductor (R statistics package)
  - http://www.bioconductor.org/
- BaRC analysis tools:
  - http://iona.wi.mit.edu/bio/tools/bioc_tools.html
- Excel
- TIGR MultiExperiment Viewer (MeV)
- Many commercial and open source packages

Exercise 1 - Excel syntax

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

Exercise 1: Excel functions

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


The R Project for Statistical Computing - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.r-project.org/
Go
George's Genome Br... WIBR genome regio... Camargo Lab - Lolig... Eorsa TargetScan Jura TargetScan Nriti's orthology tools 2can Bioinformatics ...
PCA 5 vars
princomp(x = data, cor = cor)
Fertility
Examination
Education
Catholic
Agriculture
(1-3) 60%
0.0 1.0
Clustering 4 groups
Factor 1 [41%]
Factor 3 [19%]
Groups
28
16
1
2
80 60 40 20 0

The R Project for Statistical Computing

About R
- What is R?
- Contributors
- Screenshots
- What's new?

Download
- CRAN

R Project
- Foundation
- Members & Donors
- Mailing Lists
- Bug Tracking
- Developer Page
- Conferences
- Search

Documentation
- Manuals
- FAQs
- Newsletter
- Wiki
- Books
- Other

Misc
- Bioconductor
- Related Projects
- Links

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.
- **DSC 2007**, the 5th workshop on Directions in Statistical Computing, February 15-16, 2007, Auckland, New Zealand.
- **R News 6/4** has been published on 2006-10-31.
- **The R Wiki** provides an online forum where useRs can help other useRs.

RC
nputing

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 = (expression signal / chip trimmed mean) * 100
- Calculate ratios
  - Reduce data (replicates)
  - Use AVERAGE function
  - Ratio of fetal tissue/adult tissue
- Calculate $\log_2$ of expression values and ratios