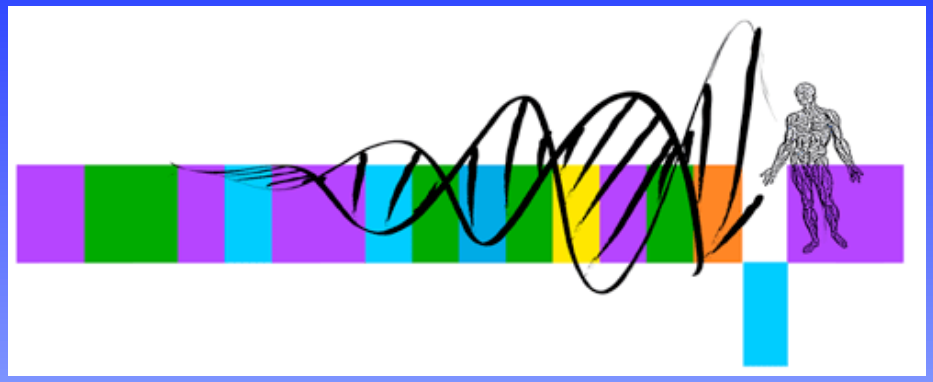


Bioinformatics for Biologists

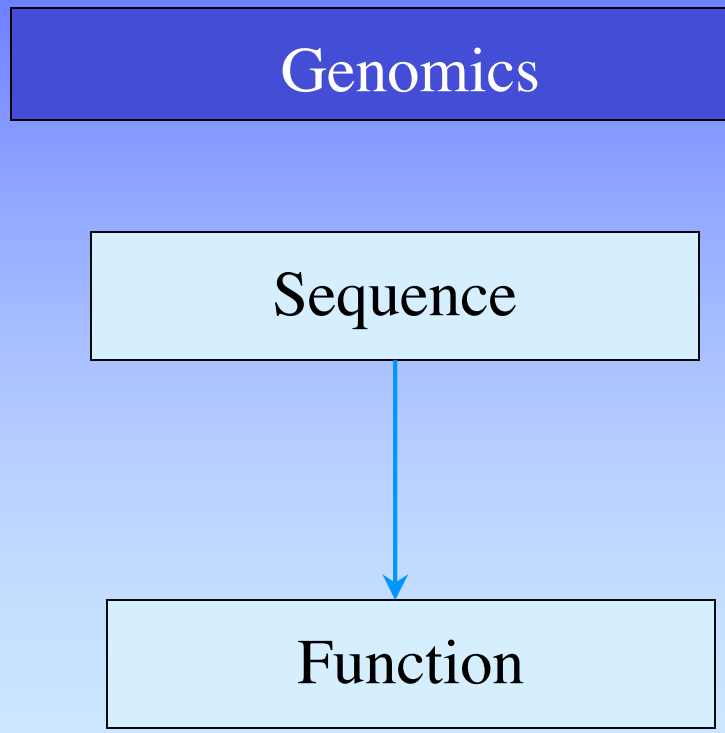
Microarray Data Analysis. Lecture 1.

Fran Lewitter, Ph.D.
Director
Bioinformatics and Research Computing
Whitehead Institute

Outline

- Introduction
- Working with microarray data
  - Normalization
  - Analysis
    - Distance metrics
    - Clustering methods

Research Trends



- How are genes regulated?
- How do genes interact?
- What are the functional roles of different genes?
- How does expression level of a gene differ in different tissues?

Transcriptional Profiling

(Adapted from Quackenbush 2001)

- Study of patterns of gene expression across many experiments that survey a wide array of cellular responses, phenotypes and conditions
- Simple analysis - what's up/down regulated?
- More interesting - identify patterns of expression for insight into function, etc.

Microarray Data

Collect data on *n* DNA samples (e.g. rows, genes, promotors, exons, etc.) for *p* mRNA samples of tissues or experimental conditions (eg. columns, time course, pathogen exposure, mating type, etc)

Matrix (*n x p*) =

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Multivariate Analysis

Concerned with datasets with more than one response variable for each observational or experimental unit (e.g. matrix X with *n* rows (genes) and *p* columns (tissue types))

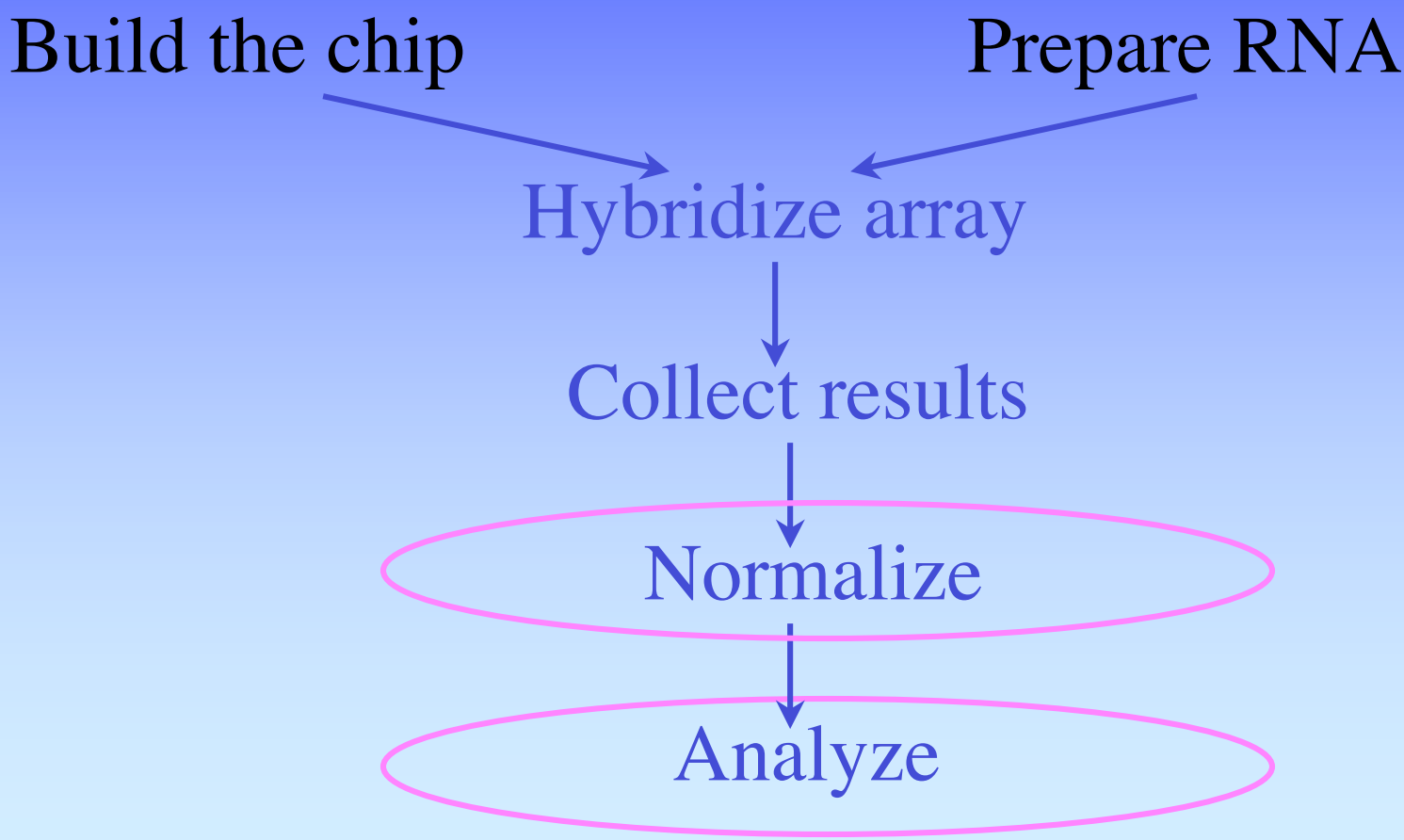
- Hierarchical (phylogenetic trees) vs non-hierarchical (k-means)
- Divisive vs agglomerative
- Supervised vs unsupervised
  - Divide cases into groups vs discover structure of data

Multivariate Methods

- Cluster analysis - discover groupings among cases of X
  - Hierarchical produces dendograms
  - K-means - choose a prespecified number of clusters
  - Self Organizing Maps

- Principal component analysis (PCA)
  - Linear method, unsupervised, seeks linear combinations of the columns of X with maximal (or minimal) variance (graphical)

DNA Microarrays



Data Normalization

- Correct for systematic bias in data
  - Avoid it, recognize it, correct it, discard outliers
- First step for comparing data from one array to another

Sources of variation

wanted vs unwanted

Across experimental conditions

Chip, slide

Hybridization conditions

Imaging

Normalization Approaches

Compensate for experimental variability

- Housekeeping genes
- Spiked in controls
- Global median normalization
- Total intensity normalization
- LOWESS correction

Expression Ratios

- Let R = a query sample
- Let G = a reference sample
- Then the ratio, $T_i = R_i/G_i$
- Need to transform these to $\log_2$
- Examples: T = 2/1 = 2; T=1/2 = .5
- Examples: $\log_2(2) = 1$; $\log_2(.5) = -1$

Total Intensity Normalization

(Adapted from Quackenbush 2002)

Assumptions: (1) start with equal amounts of RNA for the two samples; (2) arrayed elements represent random sample of genes in the organism

a. $$N_{total} = \frac{\sum_{i=1}^{Narray} R_i}{\sum_{i=1}^{Narray} G_i}$$

c. $$T_i' = \frac{R_i'}{G_i'} = \frac{1}{N_{total}} \frac{R_i}{G_i}$$

b. Rescale intensities:

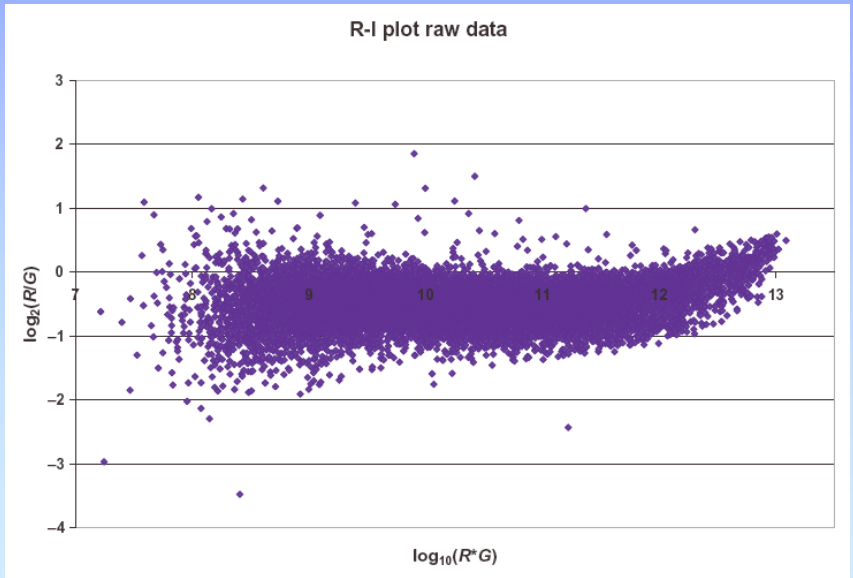
$$G_i' = N_{total} G_i \text{ and } R_i' = R_i$$

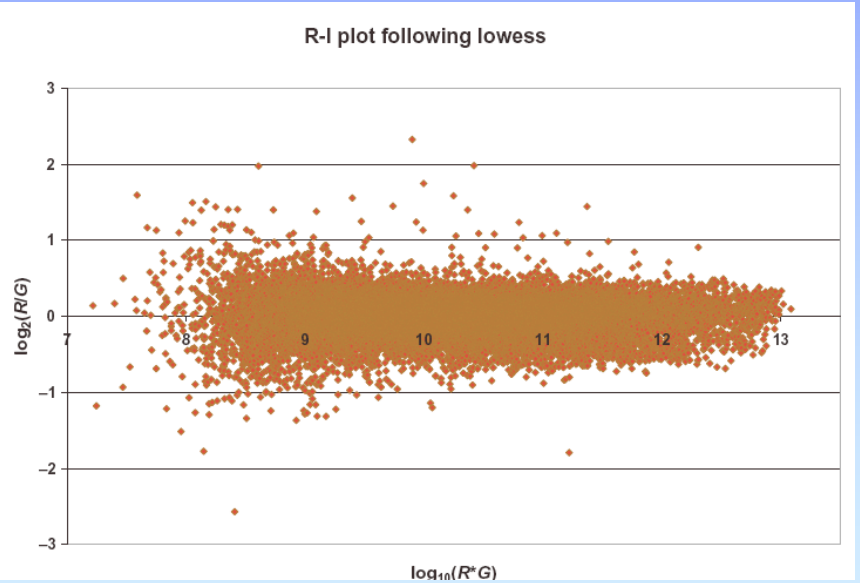
d. $$\log_2(T_i') = \log_2(T_i) - \log_2(N_{total})$$

LOWESS - The R-I Plot

(Adapted from Quackenbush 2002)

- Data exhibit an intensity-dependent structure
- Uncertainty in intensity and ratio measurements is greater at lower intensities





LOWESS - The R-I Plot

(Adapted from Quackenbush 2002)

- Plot $\log_2(R/G)$ ratio as a function of $\log_{10}(R*G)$ product intensity
- Shows intensity specific artifacts in the measurements of ratios
- Correct using a local weighted linear regression

LOWESS Normalization

(From Quackenbush 2002)

If we set $x_i = \log_{10}(R_i{}^{*}G_i)$ and $y_i = \log_2(R_i/G_i)$, lowess first estimates $y(x_k)$, the dependence of the $\log_2$(ratio) on the $\log_{10}$(intensity), and then uses this function, point by point, to correct the measured $\log_2$(ratio) values so that

$$\log_2(T_i')=\log_2(T_i)-y(x_i)=\log_2(T_i)-\log_2(2^{y(x_i)}),$$

or equivalently,

$$\log_2(T_i')=\log_2\left(T_i * \frac{1}{2^{y(xi)}}\right)=\log_2\left(\frac{R_i}{G_i} * \frac{1}{2^{y(xi)}}\right).$$

As with the other normalization methods, we can make this equation equivalent to a transformation on the intensities, where

$$G_i' = G_i * 2^{y(x_i)} \text{ and } R_i' = R_i.$$

After normalization

(Adapted from Quackenbush 2001)

- Data reported as an "expression ratio" or as a logarithm of the expression ratio
- Expression ratio is the normalized value of the expression level for a particular gene in the query sample divided by its normalized value for the control
- Use log of expression ratio for easier comparisons

Citations

- Brazma A and Vilo J. Minireview: Gene expression data analysis. *FEBS Letters* 480:17-24, 2000.
- Quackenbush J. Computational Analysis of Microarray Data. *Nature Review | Genetics* 2:418-427, 2001.
- Quackenbush J. Microarray data normalization and transformation. *Nature Genetics Supp*. 32:496-501, 2002.
- Dudoit S and Gentleman R. Classification in microarray experiments. Statistics and Genomics Short Course - Lecture 5, January 2002 (http://www.bioconductor.org/workshop.html)

Lists of Tools

- Local WI Page
  - http://jura.wi.mit.edu/bio/microarrays/biopage5tools.html
  - WADE
- R Statistics Package Microarray Tools
  - http://www.stat.uni-muenchen.de/~strimmer/rexpress.html
- Bioconductor Project
  - http://www.bioconductor.org/
- EBI
  - http://ep.ebi.ac.uk/Links.html
  - http://ep.ebi.ac.uk/EP/

Exercise 1

Excel Conventions

• 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

Functions

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

Exercise 1

To Do

Affy - fetal & human adult liver & brain tissue

- Normalize data - 8 chips (replicates)
  - Global median normalization
  - (expression signal/chip median value)*100
- Filter low intensity signals
  - Based on A/P
  - Eliminate signal similar to background
- Calculate ratios
  - Reduce data (replicates)
  - Use AVERAGE function
  - Ratio of fetal tissue/adult tissue
  - Log_2