

# Bioinformatics for Biologists

## Microarray Data Analysis. Lecture 1.

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

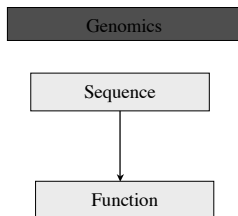
## Outline

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

WIBR Microarray Course, © Whitehead Institute, May 2004

2

## 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?

WIBR Microarray Course, © Whitehead Institute, May 2004

3

## 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.

WIBR Microarray Course, © Whitehead Institute, May 2004

4

## Microarray Data

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

Matrix ( $n \times p$ ) =

$x_{11}$	$x_{12}$	...	$x_{1p}$
$x_{21}$	$x_{22}$	...	$x_{2p}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_{n1}$	$x_{n2}$	...	$x_{np}$

WIBR Microarray Course, © Whitehead Institute, May 2004

5

## 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

WIBR Microarray Course, © Whitehead Institute, May 2004

6

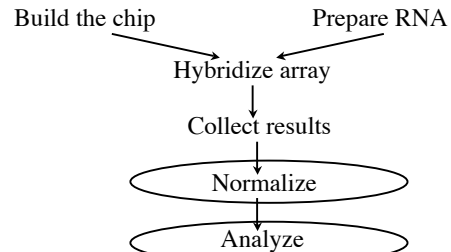
## Multivariate Methods

- 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)

WIBR Microarray Course, © Whitehead Institute, May 2004

7

## DNA Microarrays



WIBR Microarray Course, © Whitehead Institute, May 2004

8

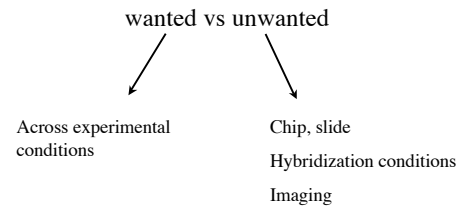
## Data Normalization

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

WIBR Microarray Course, © Whitehead Institute, May 2004

9

## Sources of variation



WIBR Microarray Course, © Whitehead Institute, May 2004

10

## Normalization Approaches

Compensate for experimental variability

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

WIBR Microarray Course, © Whitehead Institute, May 2004

11

## 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:  $\log_2(2) = 1$ ;  $\log_2(.5) = -1$

WIBR Microarray Course, © Whitehead Institute, May 2004

12

## 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. 
$$N_{total} = \frac{\sum_{i=1}^{N_{array}} R_i}{\sum_{i=1}^{N_{array}} G_i}$$

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. 
$$\log_2(T'_i) = \log_2(T_i) - \log_2(N_{total})$$

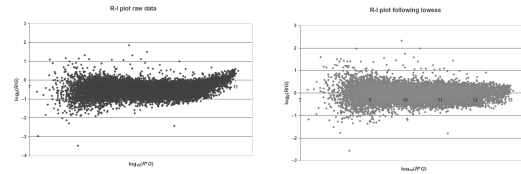
WIBR Microarray Course, © Whitehead Institute, May 2004

13

## 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



WIBR Microarray Course, © Whitehead Institute, May 2004

14

## 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

WIBR Microarray Course, © Whitehead Institute, May 2004

15

## 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  $\hat{y}(x_i)$ , the dependence of the  $\log_2(\text{ratio})$  on the  $\log_{10}(\text{intensity})$ , and then uses this function, point by point, to correct the measured  $\log_2(\text{ratio})$  values so that

$$\log_2(T'_i) = \log_2(T_i) - \hat{y}(x_i) = \log_2(T_i) - \log_2(2^{\hat{y}(x_i)}),$$

or equivalently,

$$\log_2(T'_i) = \log_2\left(T_i * \frac{1}{2^{\hat{y}(x_i)}}\right) = \log_2\left(\frac{R_i}{G_i} * \frac{1}{2^{\hat{y}(x_i)}}\right).$$

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

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

WIBR Microarray Course, © Whitehead Institute, May 2004

16

## 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

WIBR Microarray Course, © Whitehead Institute, May 2004

17

## 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>)

WIBR Microarray Course, © Whitehead Institute, May 2004

18

## 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/>

WIBR Microarray Course, © Whitehead Institute, May 2004

19

## Exercise 1

### Excel Conventions

---

- 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

WIBR Microarray Course, © Whitehead Institute, May 2004

20

## Exercise 1

### Functions

---

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

WIBR Microarray Course, © Whitehead Institute, May 2004

21

## Exercise 1

### To Do

---

Affy - fetal & human adult liver & brain tissue

- Normalize data - 8 chips (replicates)
  - Global median normalization
  - $(\text{expression signal} / \text{chip median value}) * 100$
- 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
  - $\text{Log}_2$

WIBR Microarray Course, © Whitehead Institute, May 2004

22