# Design and Analysis of Microarray Gene Expression Experiments

Guilherme J. M. Rosa

MICHIGAN STATE UNIVERSITY

CSIRO

Brisbane, Australia

20-24 October, 2003

## INTRODUCTION

- ➡ Genomics
- ➡ The Central Dogma of Molecular Biology
- ➡ Microarray Experiments
- ➡ Statistical Issues

## GENOMICS

Phenotype = Genotype + Environment

➨ Traditional Animal/Plant Breeding:

- Use P to predict G.
- Attempt to make permanent change in P.

➨ Genomics:

- Whole-genome genetics.
- Use G to predict P.
- Selection based on G to make permanent changes in P.

## BACKGROUND

➨ What is G?

- Complete sequence of DNA in an individual.

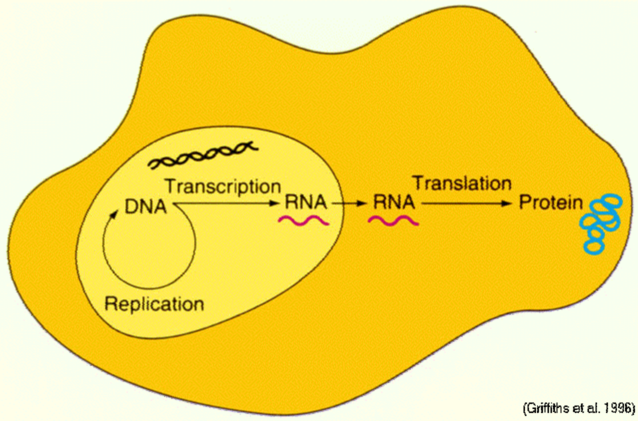
➨ What is a gene?

- Unit of inheritance.
- Specific sequence of DNA with defined beginning and end that codes for a protein.
- To have function gene must be expressed:

  ☞ Functional Genomics.

DNA → RNA → Protein

# THE CENTRAL DOGMA OF MOLECULAR BIOLOGY

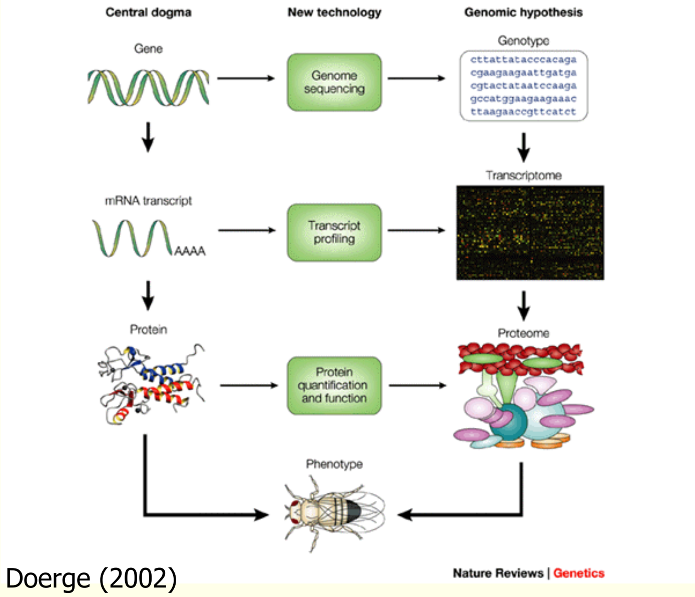


(Griffiths et al. 1996)

**Idea:** measure the amount of mRNA to see which genes are being expressed in (used by) the cell.

Measuring protein might be better, but is currently harder.

# MICROARRAY EXPERIMENTS



Source: Doerge (2002)

## MICROARRAY EXPERIMENTS

✱ mRNA levels compared in many different contexts

⇨ Different tissues, same organism (brain vs. liver)

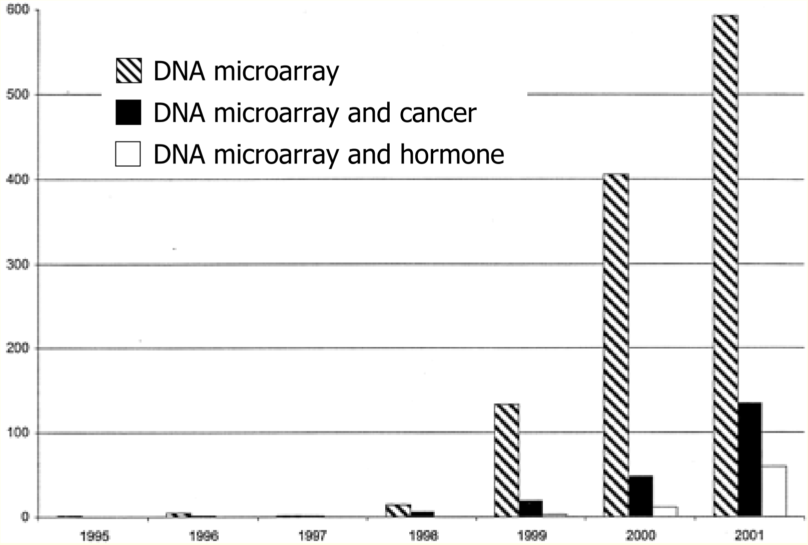
⇨ Same tissue, same organism (trt vs. ctl, tumor vs. non-tumor)

⇨ Same tissue, different organisms (wt vs. mutant)

⇨ Time course experiments (developmental differences)

⇨ Etc.

## TOTAL MICROARRAY ARTICLES - MEDLINE (Afshari, 2002)



# STATISTICAL TREATMENT OF MICROARRAY DATA

## Supervised:

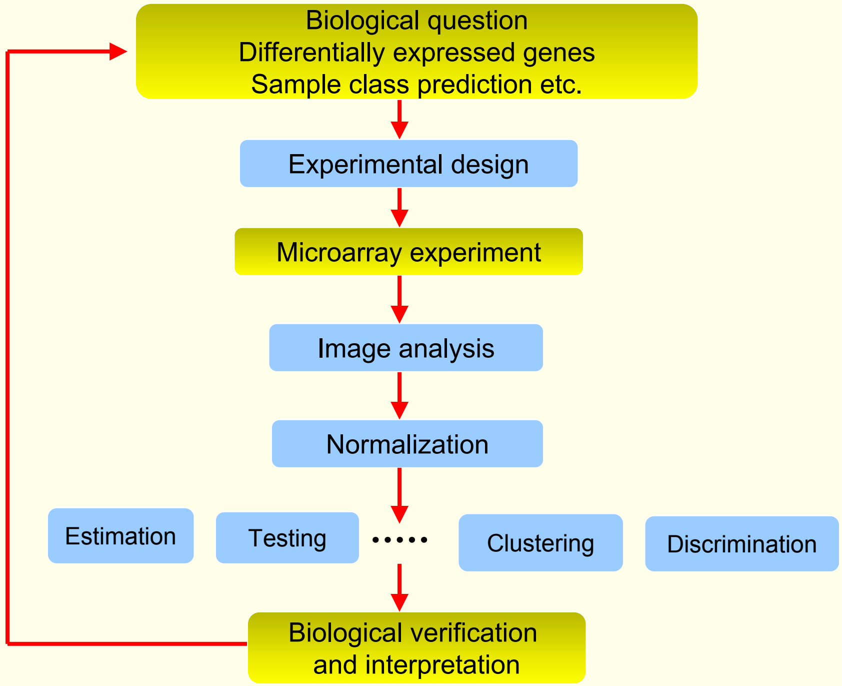
- Hypothesis driven.
- Example: Significance testing, Classification
- Some issues: Models assumptions, multiplicity

## Unsupervised:

- Not hypothesis driven; allows the discovery of “hidden” features
- Example: Cluster analysis

## Additional approaches:

- Integration with other sources of data
- Example: Molecular marker information, phenotypes (trait) scores
- Tool from the two previous items may be used together



Biological question
Differentially expressed genes
Sample class prediction etc.

Experimental design

Microarray experiment

Image analysis

Normalization

Estimation

Testing

•••••

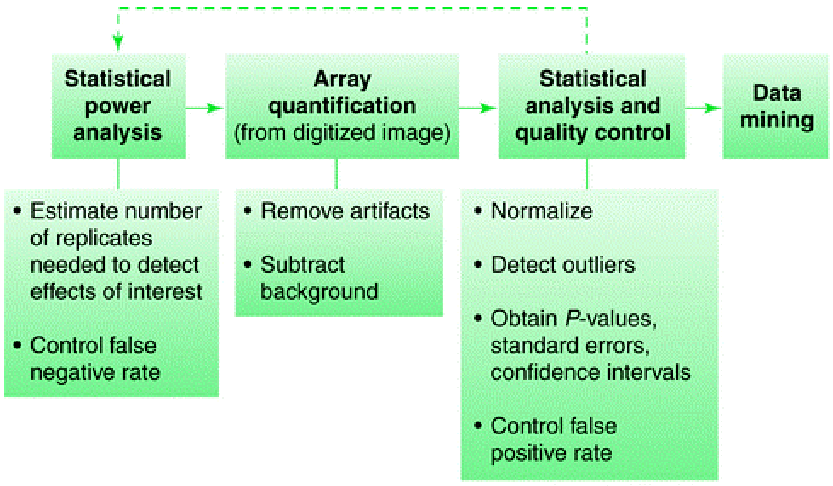
Clustering

Discrimination

Biological verification
and interpretation

# DATA ANALYSIS WORKFLOW

(Nadon and Shoemaker, 2002)



Statistical power analysis
Array quantification (from digitized image)
Statistical analysis and quality control
Data mining
• Estimate number of replicates needed to detect effects of interest
• Control false negative rate
• Remove artifacts
• Subtract background
• Normalize
• Detect outliers
• Obtain P-values, standard errors, confidence intervals
• Control false positive rate

TRENDS in Genetics

# HYPOTHESIS TESTING

- Data set:
  - $N \times p$ observations

$$N = \sum_{i=1}^{k} n_i$$

  - Usually $p >> N$
  - Goal: Compare gene expression patterns across groups

Experimental or observational groups

| | | Expression profiling | | | |
|---|---|---|---|---|---|
| Group | Subj. | $\text{gene}_1$ | $\text{gene}_2$ | … | $\text{gene}_p$ |
| 1 | 1 | $y_{111}$ | $y_{112}$ | | $y_{112}$ |
| 1 | 2 | $y_{121}$ | $y_{122}$ | | $y_{122}$ |
| … | … | | | | |
| 1 | $n_1$ | $y_{1n1}$ | $y_{1n2}$ | | $y_{1n2}$ |
| 2 | 1 | $y_{211}$ | $y_{212}$ | | $y_{212}$ |
| … | … | | | | |
| 2 | $n_2$ | $y_{2n1}$ | $y_{2n2}$ | | $y_{2n2}$ |
| … | … | | | | |
| k | 1 | $y_{k11}$ | $y_{k12}$ | | $y_{k12}$ |
| … | … | | | | |
| k | $n_k$ | $y_{kn1}$ | $y_{kn2}$ | | $y_{kn2}$ |

# CLUSTERING

➨ Data set:

- $N \times p$ observations
- Usually $p >> N$
- Goal: Groups subjects and/or genes with similar expression profiling

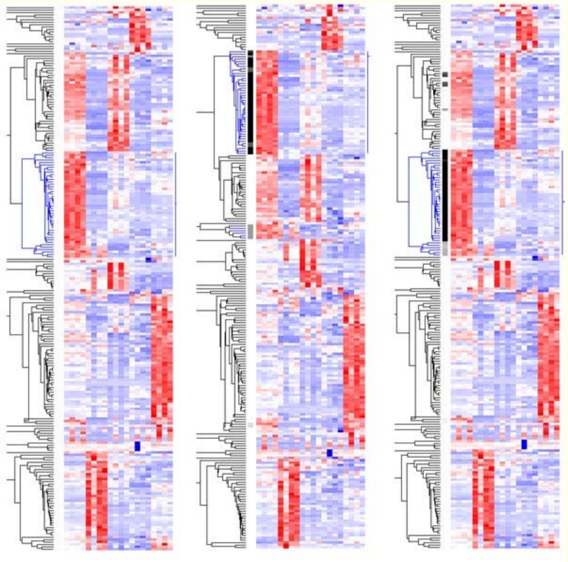
|  | Expression profiling |  |  |  |
|---|---|---|---|---|
| Subj. | $gene_1$ | $gene_2$ | … | $gene_p$ |
| 1 | $y_{11}$ | $y_{12}$ | … | $y_{12}$ |
| 2 | $y_{21}$ | $y_{22}$ | … | $y_{12}$ |
| … | … | … |  | … |
| N | $y_{N1}$ | $y_{N2}$ | … | $y_{N2}$ |

# A PRIME ON CLUSTERING ANALYSIS

- Clustering analysis of microarray data: two genes that are co-regulated (transcriptional level) present correlated expression values across samples.
- Clustering algorithm use these correlations (or the monotone transformation of correlations) to cluster genes.

➨ Example:

- Selected 225 genes: “Presence” proportion > 0.5 and CV > 70% across the 20 samples in array set 2.
- Standardize gene’s expression values ~ (0, 1)
- Blue and red represent lower and higher expression levels, respectively
- Suppose we are interested in the gene branch colored in blue
- Panel (b): clustering after a particular resampling
- Panel (c): after resampling 30 times (vertical gray-scale bar denotes the reliability of each gene belonging to the original cluster)



(a) 225 filtered genes are clustered based on their expression profiles across the 20 samples. (b) The clustering tree (blue) after a particular resampling. (c) After resampling 30 times, the reliability of the genes belonging to the original cluster is indicated by the vertical gray-scale bar on the right of the clustering tree.

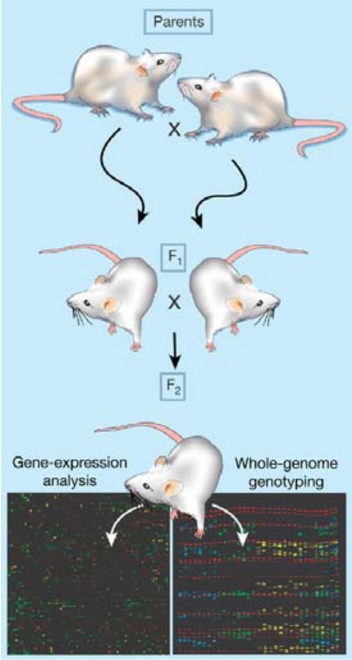
## GENETICAL GENOMICS

Integrating molecular marker information and gene expression profiling.

Alternatives:

i) 2 independent experiments; overlap of results

ii) gene expression treated as phenotype in QTL analysis (eQTL)



# REFERENCES

Afshari, C. A. (2002) Perspective: Microarray technology, seeing more than spots . Endocrinology 143(6): 1983-1989.

Doerge, R. W. (2002) Mapping and analysis of quantitative trait loci in experimental populations. Nature Reviews Genetics 3 (1): 43-52.

Parmigiani, G; E. S. Garrett; R. A. Irizarry and S. L. Zeger (Eds)(2003). The Analysis of Gene Expression Data: Methods and Software. Springer, New York. 455p.

Speed, T. (Ed)(2003). Statistical Analysis of Gene Expression Microarray Data. Chapman & Hall/CRC, Boca Raton, Fl. 222p.