



Analysis of microarray data

Sandra Rodriguez-Zas
Department of Animal Sciences
Department of Statistics
Institute for Genomic Biology
Neuroscience Program

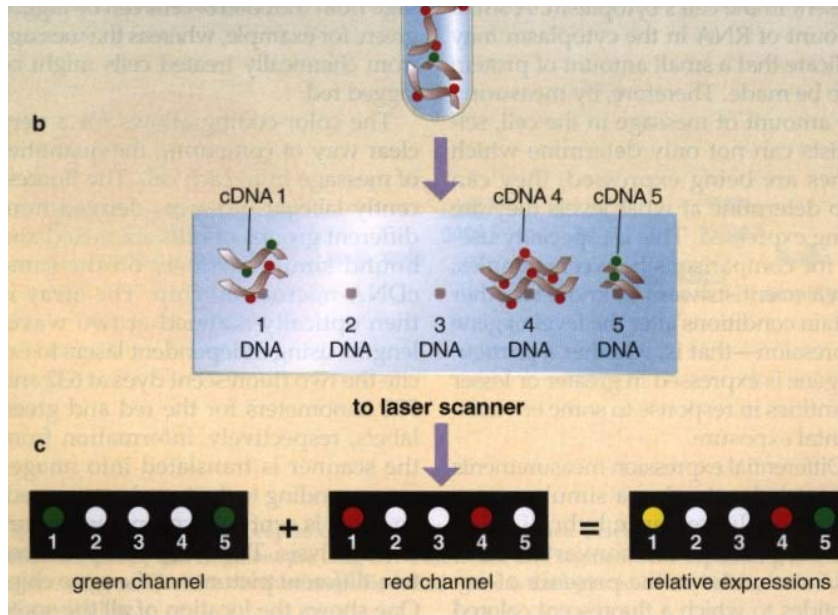
Objectives

- Summarize data
- Make inferences, test hypothesis
- Find common patterns and classify samples

Steps

- Background subtraction, data filtering
- Normalization
- Statistical analysis
- Experimental design
- Clustering and pattern finding

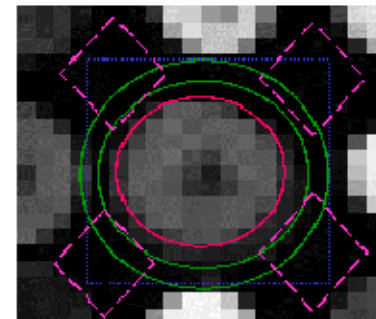
Two-dye spotted arrays



(Hamadeh and Afshari, 2000)

Background subtraction

--- GenePix
 --- QuantArray
 --- ScanAnalyze



Filtering

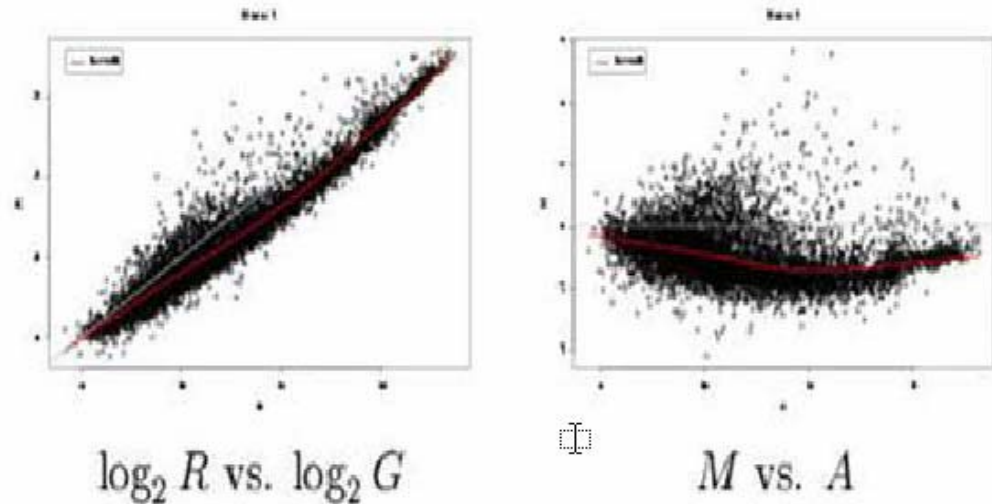
- Remove dubious intensity observations
 - spots with flags
 - spots that do not reach some criteria

Normalization

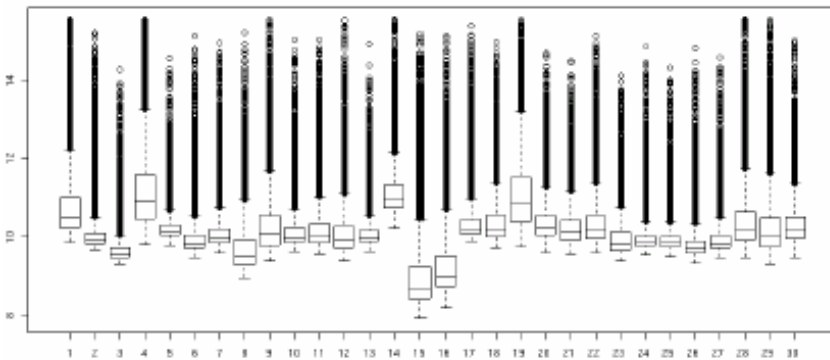
- Removal of variation other than factor of interest
- Enables combination of data from different dyes and arrays

Examples of technical variation:

■ A) Within array



■ B) Across arrays



- Multiple normalizations must be evaluated:
 - Global and local
 - Shift, LOESS, Quantile, etc.

Statistical analysis of gene expression

- Identification of genes that across conditions:
 - are differentially expressed
 - have similar patterns

- General per-gene model:

$$Y_{ijkm} = \mu + A_i + D_j + (AD)_{ij} + C_k + e_{ijkm}$$

- Y_{ijkm} = signal of spot m , array i , dye j , condition k
- μ = average signal across all factors
- A_i : global effect of array i , D_j : global effect of dye j
- $(AD)_{ij}$: interaction between array and dye
- $(C)_k$: effect of condition level k (factors and/or covariates and/or interaction and/or blocks)

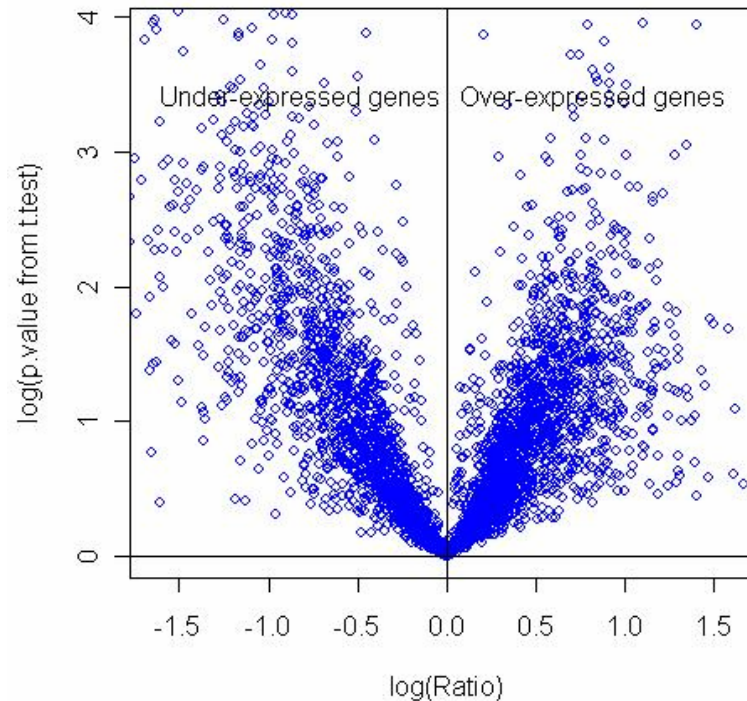
Potential terms in the BeeSpace models:

- Subspecies (e.g. *mellifera*, *ligustica*, *dorsata*)
- Genotype (e.g. SDI, NMQ)
- Colony type (e.g. SCC, TCC)
- Age (e.g. 5d, 15d)
- Maturation rate (e.g. precocious, normal)
- Treatment (e.g. cGMP, Mn, starvation)
- Roles (e.g. guards, soldiers)
- Food location (e.g. tunnel, field)
- Location (e.g. Illinois, Mexico, India)
- Colony (multiple colonies per condition)
- Bees (multiple bees nested within colony per condition)
- Adjustments for environment (e.g. temperature), etc.

Assessment of statistical significance

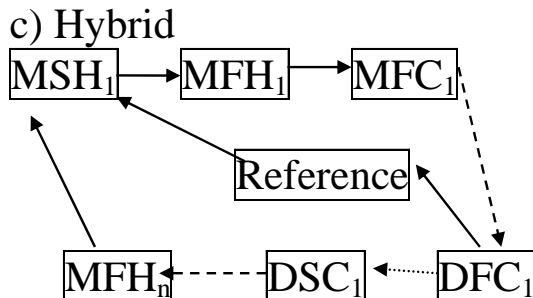
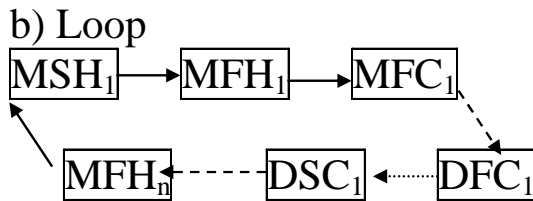
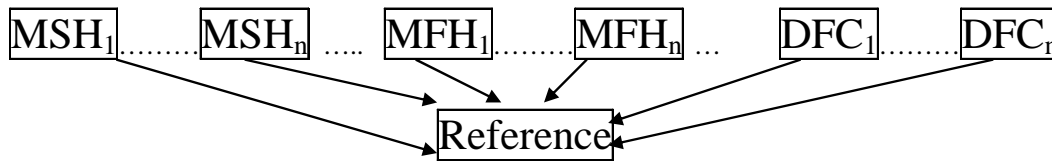
- a) adjust gene significance value for multiple testing
 - Bonferroni, FDR, re-sampling methods
- b) consider statistical and biological significance

Volcano Plot



Experimental design

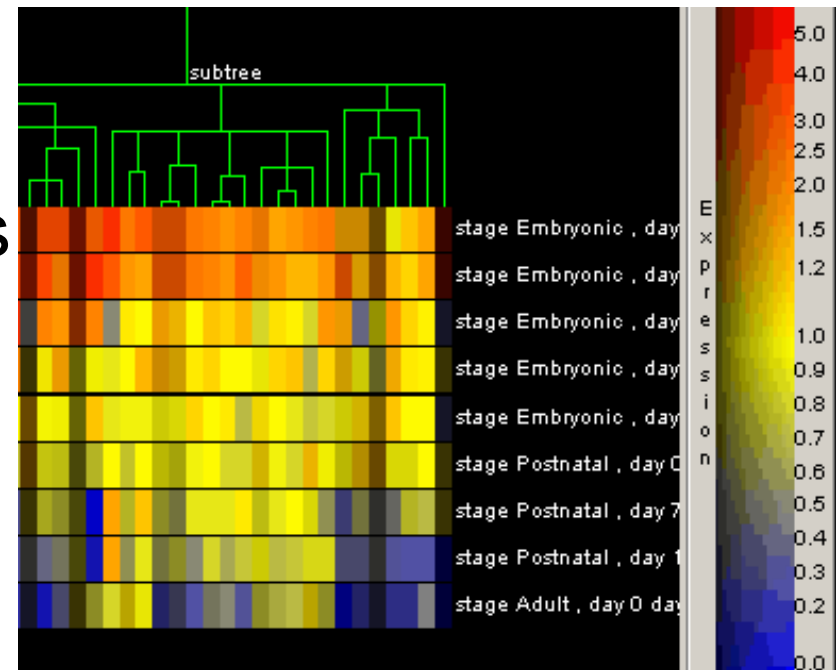
- Consider the conditions: genotype (*Mellifera*, *Ligustica*, *Dorsata*), season (**S**pring, **F**all), and treatment (**H**ormone, **C**ontrol)
- Each one-way arrow represents an array



- BeeSpace microarray experimental designs: reference, loop and hybrid

Clustering, data reduction, visualization

- Cluster analysis: techniques for classifying genes or conditions into groups
- Clustering results depend on distance measurement and clustering method:
 - Euclidean distance, correlation, etc.
 - Average, Complete, etc.
 - Clusters should be consistent
- Dendrograms depict groups



Multidimensional Data Reduction

- Reduce many genes or conditions into few factors

- Principal components (pc):

- p linear functions (indices) of p original variables

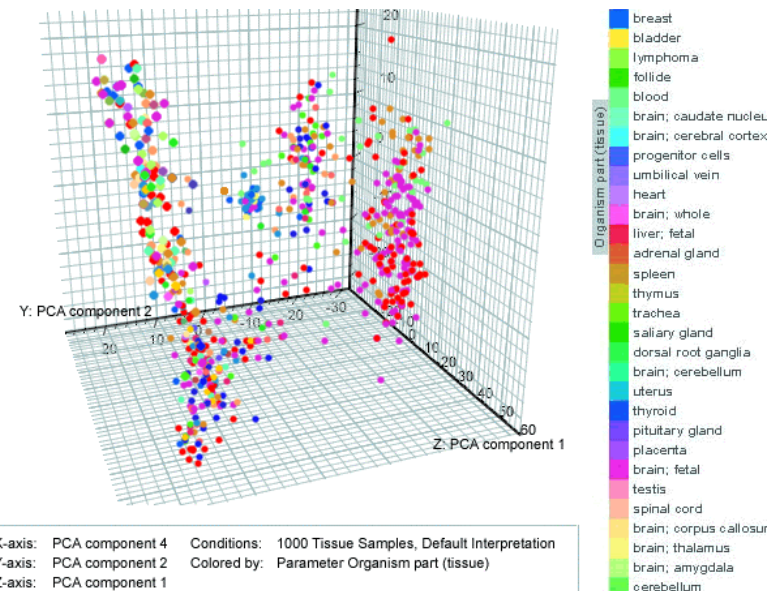
$$pc_1 = b_{11}(x_1) + b_{12}(x_2) + \dots + b_{1p}(x_p)$$

...

$$pc_p = b_{21}(x_1) + b_{22}(x_2) + \dots + b_{2p}(x_p)$$

x_p = signal of gene or condition p

b_{1p} = weight of p



Statistical Packages

- R (bioconductor or general functions)
- SAS

Depository of microarray data

- Array information will be deposited in MIAME compliant public databases (e.g. ArrayExpress)

- Thank you

- Questions?