Tentacular analysis of microarray data

Dhammika Amaratunga

Senior Research Fellow, Nonclinical Biostatistics

Johnson & Johnson
PHARMACEUTICAL RESEARCH
& DEVELOPMENT, L.L.C.

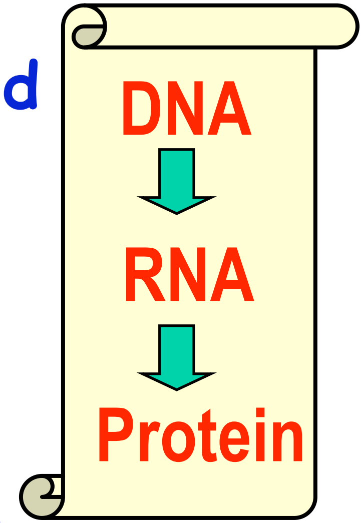
Joint work with Javier Cabrera, Hinrich Göhlmann,
Nandini Raghavan, Jyotsna Kasturi, Willem Talloen, Luc Bijnens,
James Colaianne and others

NCS2008, Leuven, Belgium, September 2008

A brief history of omics

About 60 years ago:

♦ Realization that genetic information is carried by DNA (Avery et al 1944), structure of DNA deduced (Watson and Crick, 1953), mode of DNA expression elucidated (Crick, 1958)



DNA

RNA

Protein

About 10 years ago:

♦ Sequencing of human genome near completion

♦ Work on understanding the functions of these genes under various conditions goes into overdrive with the development of microarrays, with which expression levels of several thousand genes can be simultaneously measured

♦ Expectation of better disease management via biotechnology and the various omics (accompanied by lots of hype such as the promise of "personalized medicine" within a few years)

Where are we now?

♦ Progress being made but evolution slow

♦ Technical difficulties encountered but e.g. microarrays reaching maturity as a core technology

♦ Biologists are gaining a deeper understanding of various diseases but progress related to disease management has been slow, in part because (a) genetic factors contribute only partially to common complex diseases (b) new findings have little supporting body of knowledge

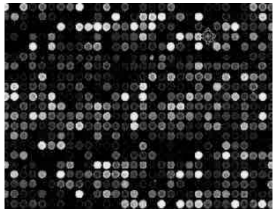
♦ Interpretation of omics data reaching maturity as a practice but very slow recognition of the emergence of data management and data analysis as bottlenecks

A typical microarray experiment

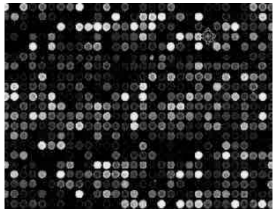
- ♦ Premise: Physiological changes ↔ Gene expression changes ↔ mRNA abundance level changes
- ♦ Objective: Use gene expression levels measured via DNA microarrays to identify a set of genes that are differentially expressed across two sets of samples (e.g., in diseased cells compared to normal cells)

Normal cells:

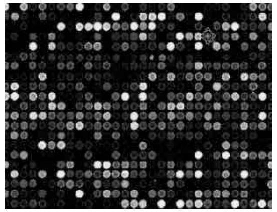
N1



N2

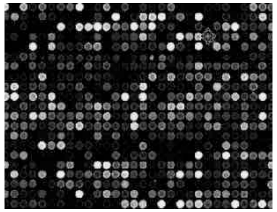


N3

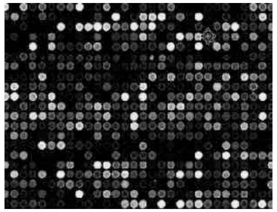


Diseased cells:

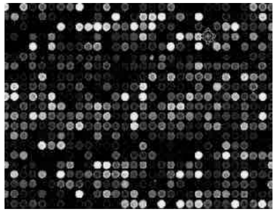
D1



D2



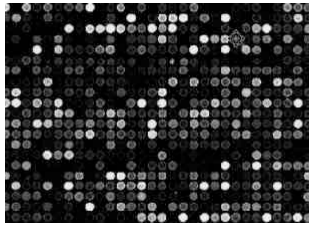
D3



Data

Expression levels for *G* genes in *N* samples

	C1	C2	C3	T1	T2	T3	…
G1	83	94	82	111	130	122	
G2	16	14	7	2	11	33	
G3	490	879	193	604	1031	962	
G4	46458	49268	74059	44849	42235	44611	
G5	32	70	185	20	25	19	
G6	1067	891	546	906	1038	1098	
G7	118	111	95	896	536	695	
G8	10	30	25	24	31	28	
G9	166	132	162	27	109	213	
G10	136	139	44	62	23	135	
	. . .	. . .	. .	. .	. .		
	(22283 genes)						



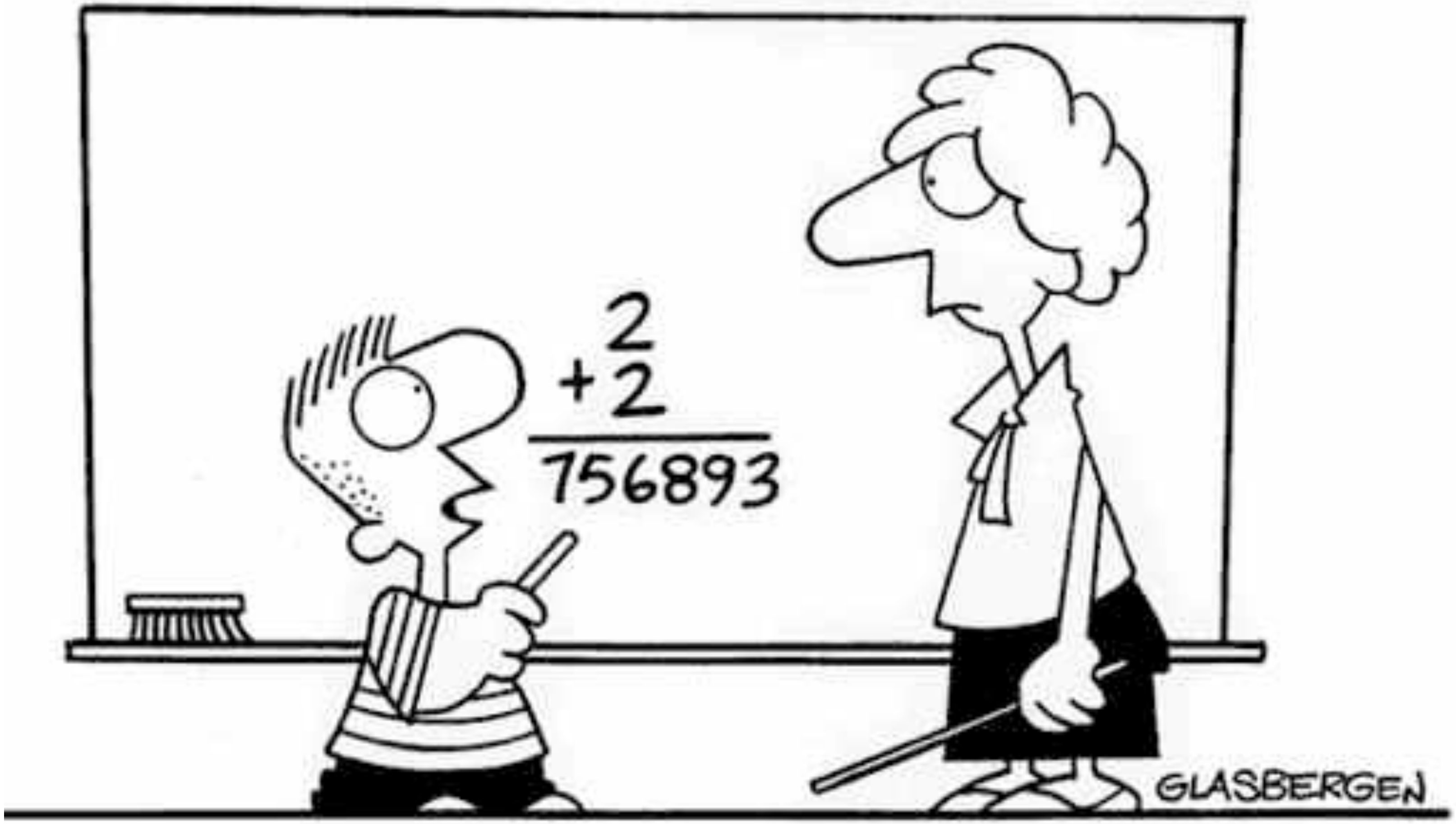
Stage 1: Assess quality & preprocess

Stage 2: Analyze

Note: *N* is small, *G* is very large.

Preprocessed data

	C1	C2	C3	T1	T2	T3	
G8521	6.89	7.18	6.60	7.40	7.15	7.40	
G8522	6.78	6.55	6.37	6.89	6.78	6.92	
G8523	6.52	6.61	6.72	6.51	6.59	6.46	
G8524	5.67	5.69	5.88	7.43	7.16	7.31	
G8525	5.64	5.91	5.61	7.41	7.49	7.41	*
G8526	4.63	4.85	5.72	5.71	5.47	5.79	
G8527	8.28	7.88	7.84	8.12	7.99	7.97	
G8528	7.81	7.58	7.24	7.79	7.38	8.60	
G8529	4.26	4.20	4.82	3.11	4.94	3.08	
G8530	7.36	7.45	7.31	7.46	7.53	7.35	
G8531	5.30	5.36	5.70	5.41	5.73	5.77	
G8532	5.84	5.48	5.93	5.84	5.73	5.73	
G8533	9.45	9.56	9.92	10.15	9.81	9.36	
G8534	7.57	7.55	7.30	7.48	7.82	7.46	



"In an increasingly complex world,
sometimes old questions require new answers."

Characteristics of microarray data

- Lots of data but usually many features (*G*=10000-50000) measured on few samples (*N*=5-100)

⇒ Information content per feature is low

⇒ Potential for overfitting of data and misinterpretation of findings is very high

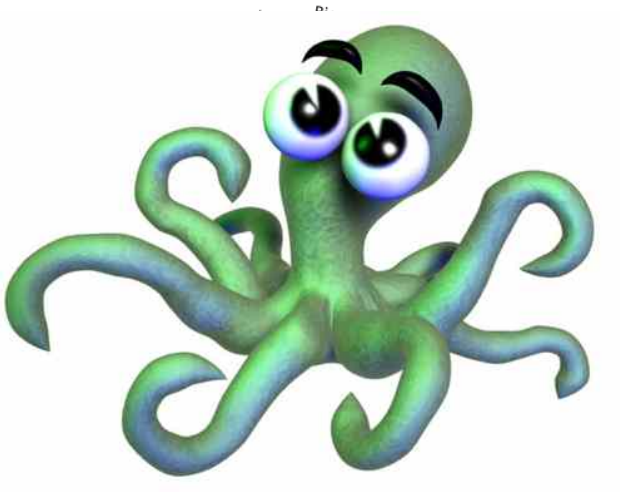
- Data is complex (not just a matrix)

⇒ Ancillary biological information

⇒ Database management

⇒ Specialized statistical tools

⇒ Multi-armed (tentacular) approach needed for interpretation



What are we really looking for?

♦ A “gene expression signature”:

Flexible definition depending on potential use:
- To understand the underlying biology.
- A classifier of sorts or a composite biomarker.

1. Set of genes differentially expressed in D vs N.
2. Not necessarily an exhaustive list.
3. Not necessarily a classifier or discriminant in the strict statistical sense; redundancy low but not necessarily zero.
4. Not necessarily unique.
5. Reasonably specific to D vs N.

(a) Excludes highly non-specific genes such as stress genes.

(b) Excludes potentially non-specific genes such as genes that may differentiate D' vs N where D' is similar but not identical to D.

Individual gene analysis

- ♦ Fold change: Seek genes that exhibit at least a certain specified fold increase or decrease in mean expression level.

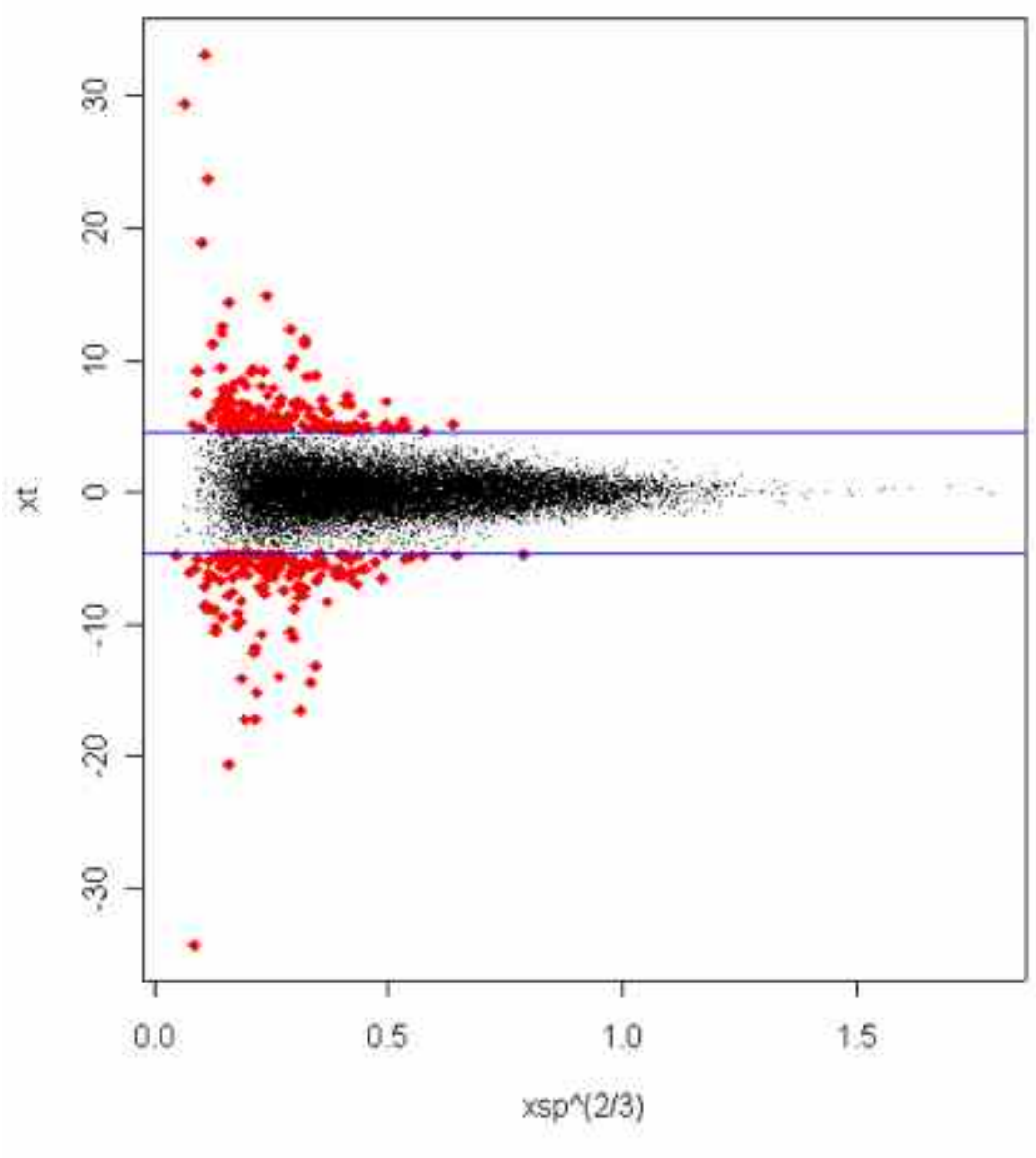
- ♦ Statistical analysis of individual genes: Seek genes that exhibit a statistically significant difference across the groups (via e.g., t, permutation test, Ct, SAM, limma, Bayes/EmpiricalBayes procedures).

- ♦ Adjust for multiplicity: Try to control the False Discovery Rate: FDR = E(#FalsePositives /#Positives).

Compare C1-C3 vs T1-T3 usina t tests

<u>**Test**</u>**:** ***t*** **tests with $\alpha = 0.05$ (after preprocessing)**

<u>**Result**</u>**: If $X \sim N(0,\sigma^2)$, $T_g|s_g \sim N(0,\sigma^2/s_g^2)$**



Can this be improved upon?

Often the sample size per group is small.

➡ Unreliable variances (inferences).

However the number of genes is large.

➡ Borrow strength across genes.

A model for borrowing strength

- Let X_{gij} denote the preprocessed intensity measurement for gene g in array i of group j.

- Model: $X_{gij} = \mu_{gj} + \sigma_g\, \varepsilon_{gij}$

- Effect of interest: $\Delta_g = \mu_{g2} - \mu_{g1}$

- Error model: $\varepsilon_{gij} \sim F(\text{location}=0, \text{scale}=1)$

- Gene mean-variance model: $(\mu_{g1}, \sigma_g) \sim F_{\mu,\sigma}$

Possible approaches (1)

Parametric: Assume functional forms for $\boldsymbol{F}$ and $\boldsymbol{F}_{\mu,\sigma}$ and apply either a Bayes or Empirical Bayes procedure → regularized test statistics.

$$T_g = (\bar{X}_{g1} - \bar{X}_{g2}) / s_g$$

$$T_g(c) = (\bar{X}_{g1} - \bar{X}_{g2}) / (s_g + c)$$ SAM

or

$$T_g(d) = (\bar{X}_{g1} - \bar{X}_{g2}) / \sqrt{(d_g s_g^2 + d_0 s_0^2)}$$ LIMMA

Refs: Tusher, Tibshirani, and Chu (*Proc Natl Acad Sci USA,* 2001)
Smyth (*Stat Appl Genet Mol Biol.* 2004)

Possible approaches (2)

Nonparametric:

Estimate F: $=\{ \qquad \}$

Estimate $F\sigma$: $=\{s_g\}$

Resample: $r_{ij}^{*} \sim$ and $s^{*} \sim$ → $X_{ij}^{*} = s^{*} r_{ij}^{*}$

→ (t^{**}, s^{**}) → Repeat many many times

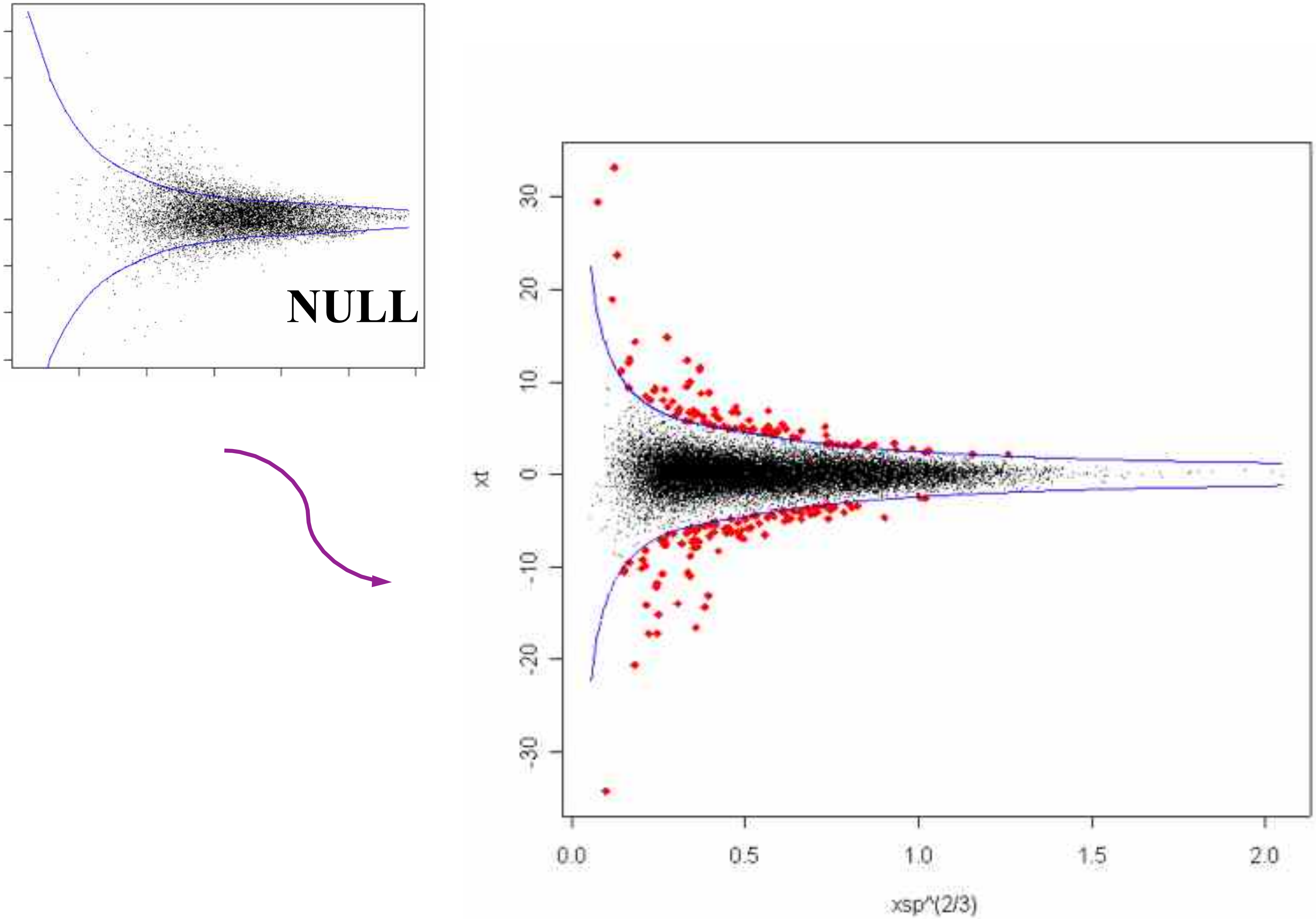
→ Form *critical envelope*, $t\alpha(s_g)$, defined by

$P(|T| > t\alpha(s_g) \mid s_g \,;\, H_0) = \alpha$

$\hat{F} \qquad (X_{gij} \cdot$

$\hat{F}$

Ref: Amaratunga & Cabrera (*Statistics in Biopharmaceutical Research, 2008*)


NULL
30
20
10
0
-10
-20
-30
xt
0.0
0.5
1.0
1.5
2.0
xsp^(2/3)

Problems with individual gene analyses

♦ Individual gene analysis produces findings that are unstable and doesn't exploit the ability of a microarray to measure the expression levels of multiple genes simultaneously reflecting the inherent interactions among genes

However:

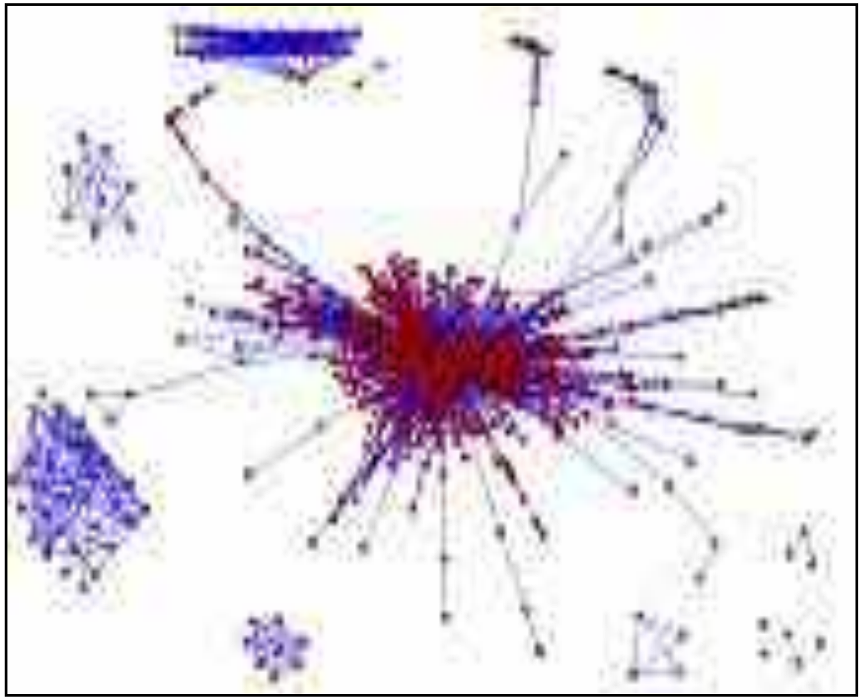
- correlations cannot be estimated well with small sample sizes
- correlations will occur both because of coexpression as well as sequence similarity
- some correlations may be understated because of biological or technical factors
- using only known associations will prevent novel genes from being detected

Multi-gene approach: co-expression network

- ♦ Co-expression networks

For example:
Calculate pairwise correlations and represent the correlation matrix as a network:

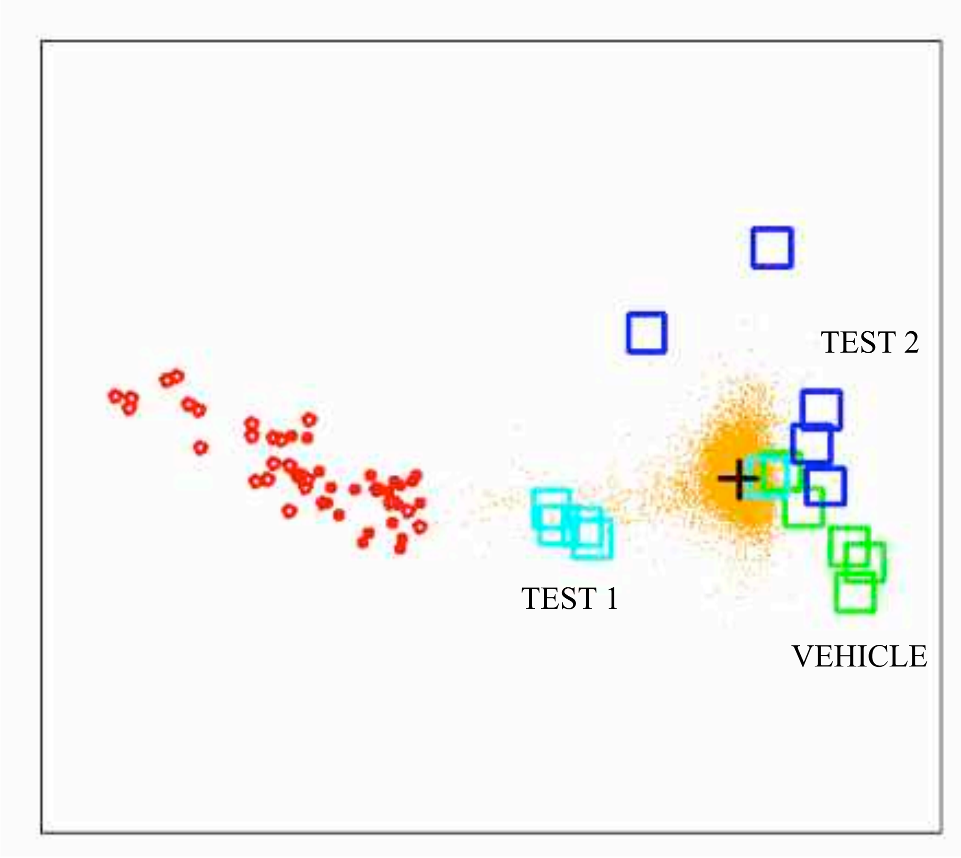
- Each gene corresponds to a node
- A gene pair is connected by an edge if and only if its correlation is high

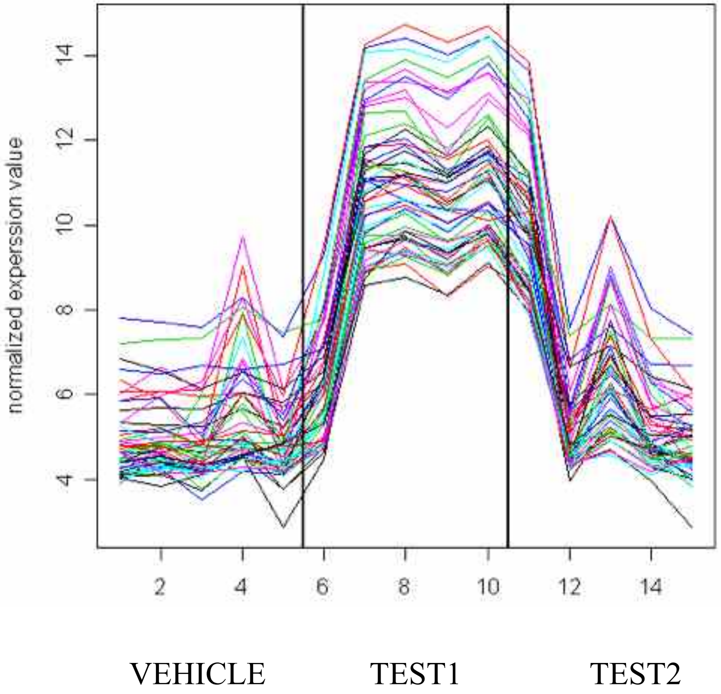


Ref: Zhang and Horvath (*Stat Applications in Genetics and Molecular Biology*, 2005)

Multi-gene approach: co-expressing differentiators

- ♦ Seek co-expressing genes that together separate the groups (via e.g., spectral maps).





Ref: Wouters et al (*Biometrics*, 2003)

Multi-gene approach: classification

♦ Seek combinations of genes that together separate the groups

- Discriminant analysis (e.g.,LDA?)
- Reduced LDA (e.g., FS+LDA, PCA+LDA)
- Penalized LDA (e.g., LASSO)
- Ensemble methods (e.g., Random Forest)

X (GxN) → X* (gxn) → LDA(X*) → Repeat many many times → Collate findings

♦ Use cross-validation or bootstrap to assess performance

Ref: Raghavan et al (*2008*)

Multi-gene approach: gene-set analysis

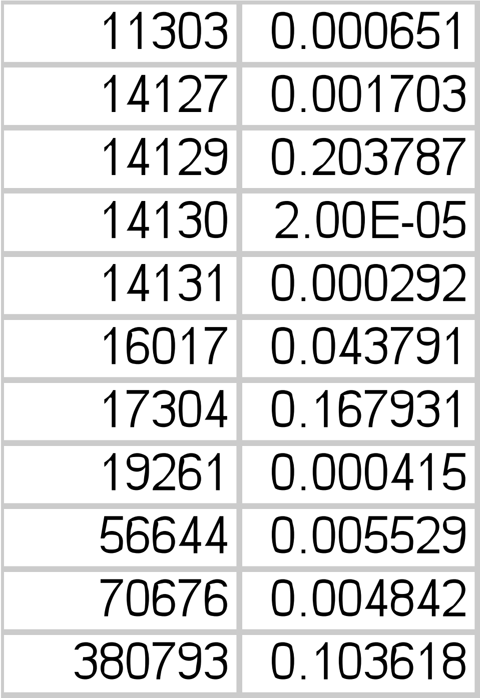
♦ Seek pre-defined gene sets that separate the groups.

Example:
Phagocytosis engulfment
in D vs N experiment

11 genes (*p*: 0.00002 - 0.2)

MLP = mean (-log *p*) = 2.34*

Significance assessed via a permutation test (permute the *p*-values across all the genes in the entire dataset).

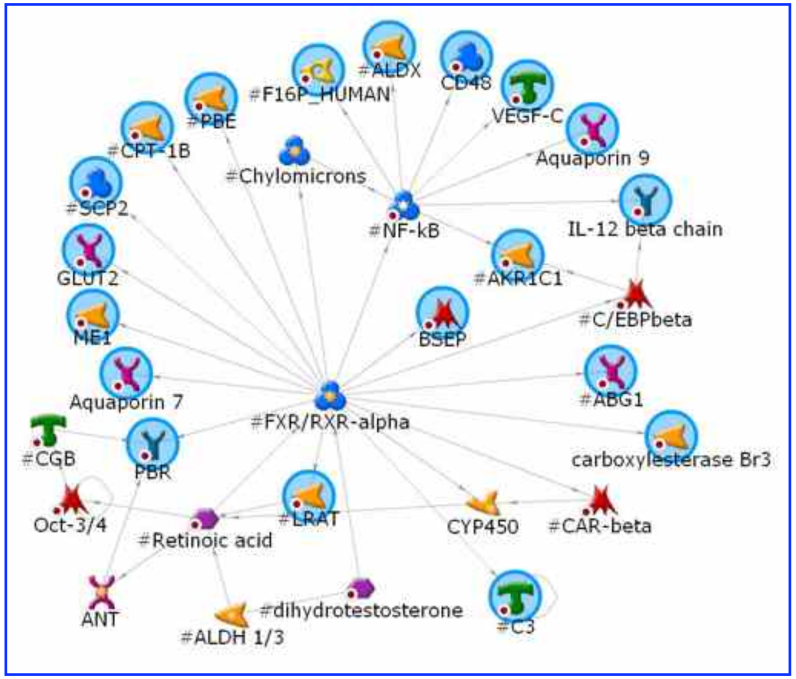


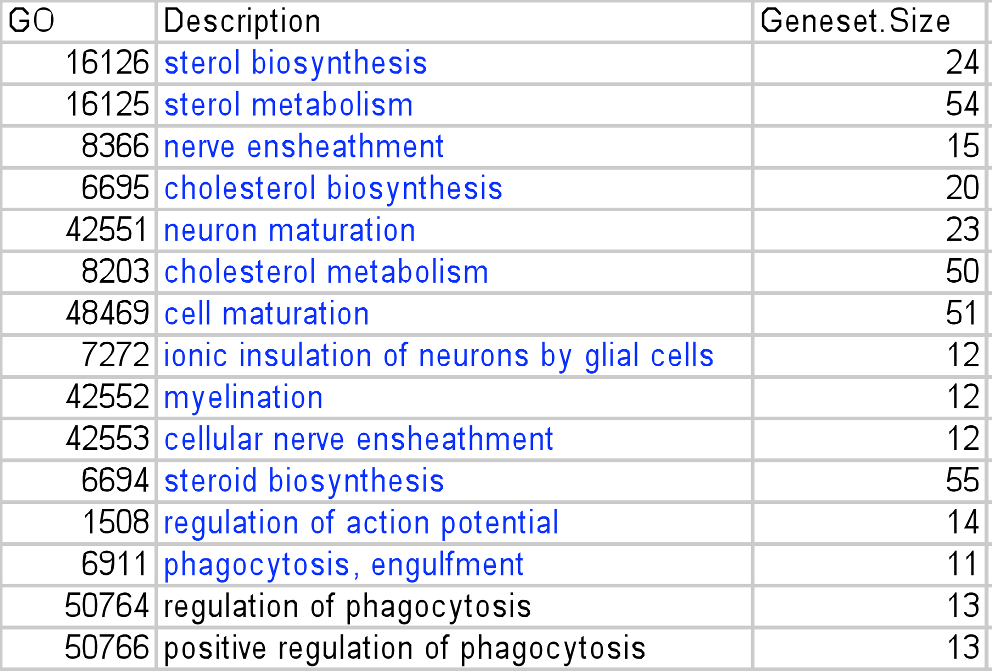
Gene	p-value
11303	0.000651
14127	0.001703
14129	0.203787
14130	2.00E-05
14131	0.000292
16017	0.043791
17304	0.167931
19261	0.000415
56644	0.005529
70676	0.004842
380793	0.103618

Ref: Raghavan et al (*Journal of Computational Biology, 2006*)

Importance of gene set analysis

- Can detect groups of modestly changing genes
- Greater stability
- Better interpretability



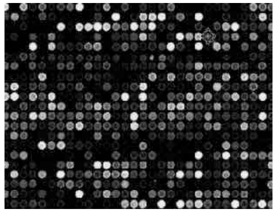


GO	Description	Geneset.Size
16126	sterol biosynthesis	24
16125	sterol metabolism	54
8366	nerve ensheathment	15
6695	cholesterol biosynthesis	20
42551	neuron maturation	23
8203	cholesterol metabolism	50
48469	cell maturation	51
7272	ionic insulation of neurons by glial cells	12
42552	myelination	12
42553	cellular nerve ensheathment	12
6694	steroid biosynthesis	55
1508	regulation of action potential	14
6911	phagocytosis, engulfment	11
50764	regulation of phagocytosis	13
50766	positive regulation of phagocytosis	13

Ref: Raghavan et al (*Journal of Computational Biology, 2006*)
and Raghavan et al (*Bioinformatics, 2007*)

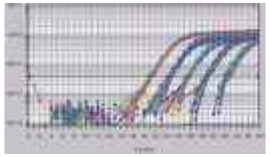
Towards a holistic approach

- Integrate data/findings with other -omics data /findings



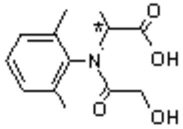
DNA microarray

genomics

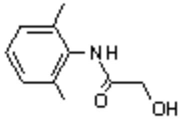


qPCR, siRNA, CNV, …

metabolomics

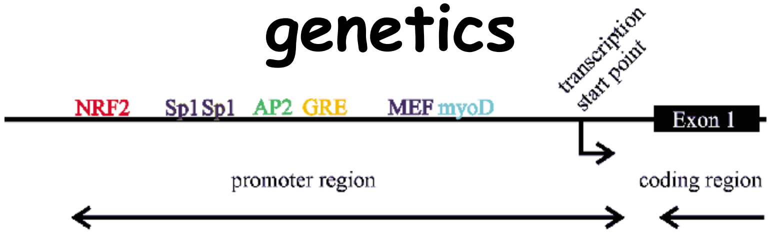


Metabolite 6

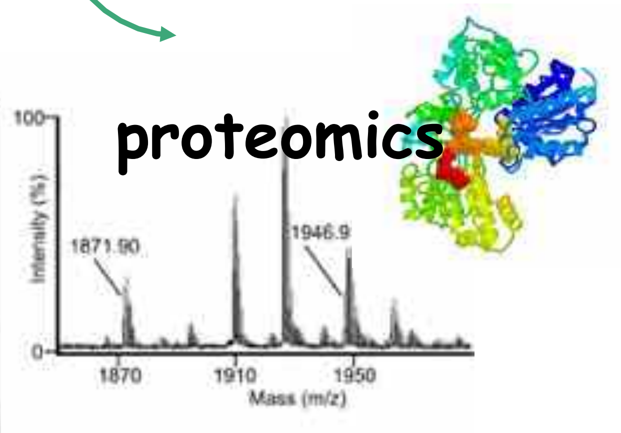


Metabolite 5

genetics



proteomics



Summary

- Microarrays are reaching maturity as a technology.
- Making sense of microarray data is an inter-disciplinary effort in which statistical considerations play an important role.
- From a statistician's perspective, it is important to keep in mind that microarray experiments are (over-parametrized under-sampled) screening experiments and a careful balance must be struck between finding a signal and overfitting.

Wrap up

- **References:**

D. Amaratunga and J. Cabrera (2004) *Exploration and Analysis of DNA Microarray and Protein Array Data*, New York: John Wiley.

D. Amaratunga, J. Cabrera and Y. S. Lee (2008) Enriched random forests, *Bioinformatics*.

D. Amaratunga and J. Cabrera (2008). A conditional *t* suite of tests for identifying differentially expressed genes in a DNA microarray experiment with little replication, *Statistics in Biopharmaceutical Research*.

D. Amaratunga, J. Cabrera and V. Kovtun (2008) Microarray learning with ABC, *Biostatistics*, 9:128-136.

D. Amaratunga and J. Cabrera (2001) Statistical analysis of viral microchip data, *Journal of the American Statistical Association*, 96: 1161-1170.

N. Raghavan, D. Amaratunga, J. Cabrera, A. Nie, Q. Jie and M.McMillian (2006) On methods for gene function scoring as a means of facilitating the interpretation of microarray results, *Journal of Computational Biology*, 13: 798-809

N. Raghavan, A. De Bondt, W. Talloen, D. Moechars, H. Göhlmann and D. Amaratunga (2007) The high-level similarity of some disparate gene expression measures, *Bioinformatics*, 23:3032-3038

N. Raghavan, A. Nie , M. McMillian and D. Amaratunga (2008), Development of a toxicogenomic signature for non-genotoxic carcinogenicity, *in review*.

- **Website:**

www.geocities.com/damaratung

Thank You!