





THE UNIVERSITY OF ALABAMA AT BIRMINGHAM



Section ON Statistical Genetics



David B. Allison, Ph.D.
Professor

Department of Biostatistics
Dept of Nutrition Sciences
Clinical Nutrition Research Unit
School of Public Health

Recruitment actively underway for

- graduate students
- post-doctoral fellows
- research track faculty
- tenure track faculty

Applying High-Dimensional Approaches to Microarray Research

Contributors

- Grier Page
- Jode Edwards
- Prinal Patel
- Tapan Mehta
- Murat Tanik
- Moonseong Heo
- Jose Fernandez
- Gary Gadbury
- Tom Prolla
- Cheol-Koo Lee
- Rick Weindruch
- Jaap Brand
- Christopher Coffey
- Mark Beasley
- Stephen Barnes

Collaborators

- Anthony Ferrante
- Antonio Tataranni
- Angelo Del Parigi
- Kei Cheung
- Ken Williams
- Bob Eckel
- Upender Manne
- W. Timothy Garvey
- John Mountz
- Bill Grizzle
- Stephen Barnes
- Brad Yoder
- Lisa Guay-Woodford
- William E. Grizzle
- Dale Benos
- Robert Kesterson

Supported by:

NIDDK – CNRU (Allison); Biotechnology Center (Yale; Williams); Short Course on Stat Genet (Allison)

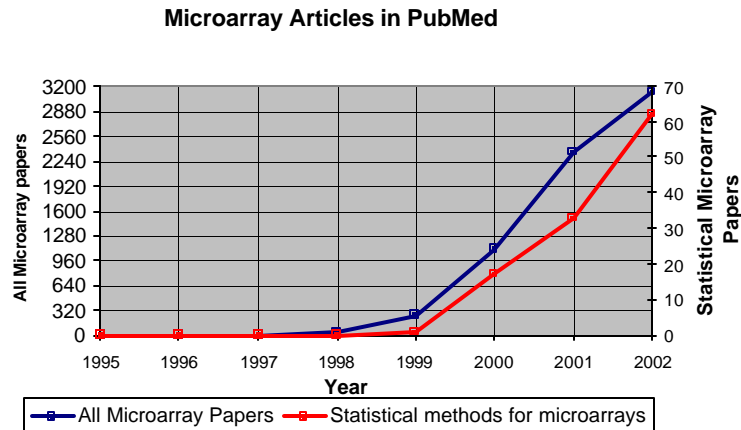
NSF – Plant Genomics grant (Allison); RCN on Microarray Analysis (Allison)

NIA – P01 & R01 (Weindruch)

NCI – U54 CNGI Grant (Barnes)

NIEHS – R01 on Stat Genet Methods (Amos)

Keeping Up with the Microarray Literature: How Many Can You Read Per Day?



From Mehta, Tanik, & Allison (in preparation).

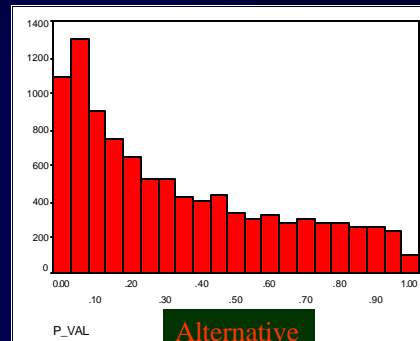
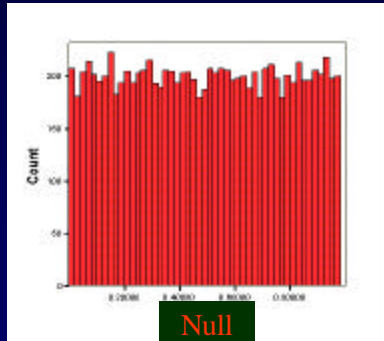
Some of The Challenges We Face

- ◆ Massive Multiple Testing
 - Inference
 - Designing studies with many hypotheses
 - Estimation
 - Model Selection
- ◆ Lack of a null hypothesis
- ◆ Keeping Our Own Field Solid.

Mixture Model Approach

Allison et al. (2002). *Comp Stat & Data Analysis*.

Under the null hypothesis, the distribution of p-values is uniform on the interval [0,1] regardless of the sample size and statistical test used (as long as that test is valid).



Under the alternative hypothesis, the distribution of p-values will tend to cluster closer to zero than to one.

Mixtures of Betas

Any distribution on the interval [0,1] can be modeled as a mixture of $v+1$ separate component distributions where the j^{th} component is a beta distribution with parameters r_j and s_j . The beta's PDF is:

$$\mathbf{b}(r,s)(x) = I_{(0,1)}(x) \frac{x^{r-1} (1-x)^{s-1}}{\mathbf{B}(r,s)}$$

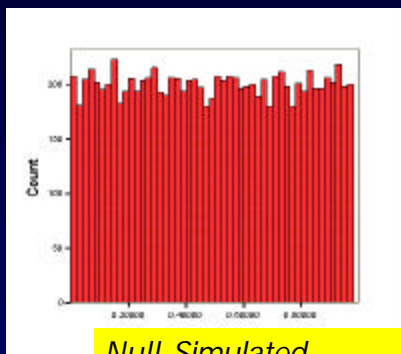
When $r=s=1$, the beta distribution is a uniform distribution.

The log of the likelihood for the collection of k p-values from a model with $v+1$ components can then be expressed as:

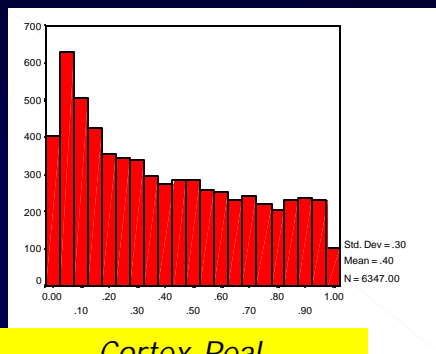
$$L_{v+1} = \sum_{i=1}^k \ln \left[I_0 \mathbf{b}(1,1)(x_i) + \sum_{j=1}^v I_j \mathbf{b}(r_j, s_j)(x_i) \right]$$

Allison et al. (2002). *Comp Stat & Data Analysis*.

Weindruch et al. Mouse Cortex Data (Old Control vs. Old Calorically Restricted)



Null-Simulated



Cortex-Real

Estimated parameters: $l_1 = .29$; $r_1 = .78$; $s_1 = 3.87$.

Allison et al. (2002). *Comp Stat & Data Analysis*.

Posterior (Bayesian) Probabilities

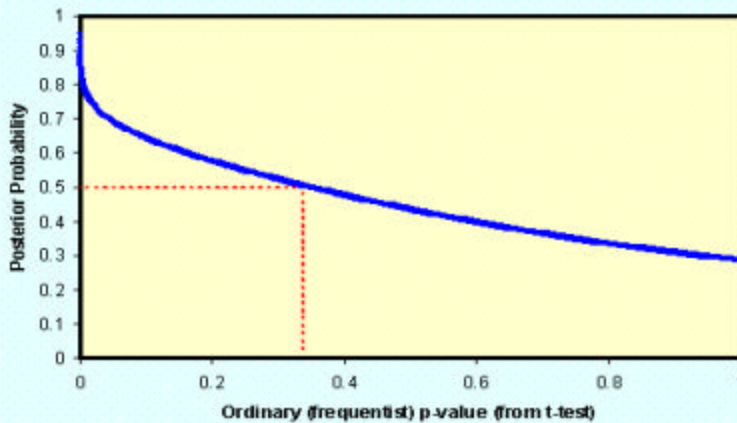
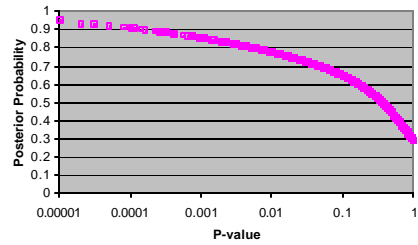
Gene	Group A			Group B			P-Values (t-test)
	A1	A2	A3	B1	B2	B3	
1	74	947	157	-566	-747	-663	0.16982
2	12	124	14	438	420	273	0.00142
3	18	68	21	15	99	36	0.48675
...
634	59	66	95	163	193	230	0.00027

Gene (i)	P-Values (t-test)	Posterior Probability
1	0.16982	0.89544
2	0.00142	0.84211
3	0.48675	0.44121
...
6347	0.00027	0.88561

Posterior Probability
Probability that the gene i is a gene from the population of genes for which the null hypothesis is not true

Allison et al. (2002). *Comp Stat & Data Analysis*.

Use of posterior probability



Allison et al. (2002). *Comp Stat & Data Analysis.*

Suppose we conduct a t-test of the difference between two means and obtain a p-value $< .05$. Does this mean:

- a) There is less than a 5% chance that the results are due to chance.
- b) If there really is no difference between the population means, there is less than a 5% chance of obtaining a difference this large or larger.
- c) There is a 95% chance that if the study is repeated, the result will be replicated.
- d) There is a 95% chance that there is a real difference between the two population means.

Adapted from: Wulff et al. (1987): What do doctors know about statistics?
Statistics in Medicine 6:3-10

FDR Defined

		Truth		
		Null	Alt	
Conclusion	Null	a	b	K-R
	Alt	c	d	R
		K-M	M	K

$$FDR = E\left(\frac{c}{c+d}\right)$$

Estimating the FDR

- ◆ The FDR is not observable in a data set but can be estimated.
- ◆ Call the alternative estimators of FDR Q_x where x indicates which particular estimator we are referring to.
- ◆ Desirable characteristics of FDR estimators include:
 1. $E(Q) \geq FDR$
 2. $E(Q-FDR)$ is minimized
 3. $E(Q-FDR)^2$ is minimized

Three Broad Approaches to FDR

- ◆ Multiple Comparison Approaches
 - Benjamini & Hochberg (1995; 2000; 2001)
- ◆ Global Null Expectation Subtraction
 - Tusher et al. (2001) [SAM]
 - Xu et al (2002)
 - Van der Wiel (submitted)
 - Edwards, Allison, et al. (unpublished)
- ◆ Model-based Approaches
 - Allison et al (2002)
 - Others?? - Manducci et al; Lee et al; Baldi & Long; Pan
- ◆ Other
 - Storey methods

False Discovery Rates (FDR) (Benjamini & Hochberg, 1995)

Let there be k p-values.

Rank the p-values in ascending order from 1 to k .

Let i denote the rank.

Find the largest p-value satisfies the inequality:

$$p_i < (i/k)Q$$

Reject all null hypotheses with $p < p_i$.

Then, the expected proportion of rejected null hypotheses that are falsely rejected (i.e., the FDR) is $\leq Q$.

Example: Weindruch, Prolla et al. CR vs. non-CR Mice (N = 3 per group) 6347 genes in heart.

No gene passes Bonferroni or Sidak even with family-wise α of .10.

14 genes pass FDR at .10.

Adaptive False Discovery Rates (FDR) (Benjamini & Hochberg, 2000)

Rank the p-values in ascending order from 1 to k.

Let i denote the rank.

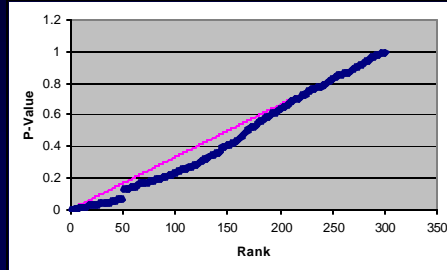
Calculate $S_i = (1-p_i)/(k+1-i)$

Starting with $i = 1$, proceed towards larger i as long as $S_i \geq S_{i-1}$.

Stop when the first time as $S_j < S_{j-1}$.

Estimate $k_0 = \min[(1/S_j + 1), k]$

Then use k_0 in place of k in B&H (1995) FDR procedure.



Example: Dr. X's obese vs lean humans.

29 genes pass Bonferroni with family-wise α of .05.

1,198 genes pass FDR at .05.

1,223 genes pass adaptive FDR at .05.

'Global Null' FDR Estimation (e.g., Tusher et al., 2001; Xu, 2002)

Figure out the proportion of true null hypotheses you expect to incorrectly reject by chance (i.e., α), then let:

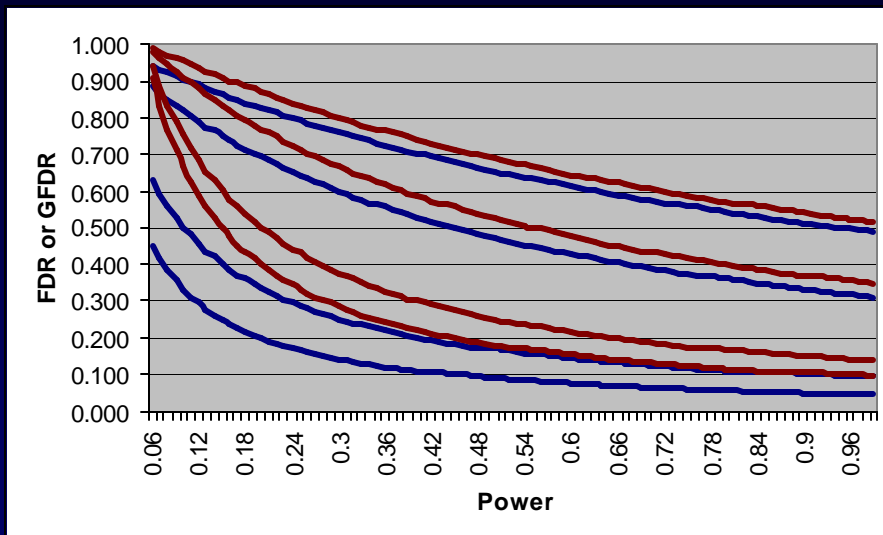
$$Q_G \equiv I(Ka < R > 0) \frac{Ka}{R} + I(Ka > R > 0)$$

Problem with 'Global Null' FDR Estimation

$$Q_G \approx FDR \frac{1}{1-m},$$

where m is the proportion of null hypotheses which are false (i.e., the proportion of genes that *are* differentially expressed).

FDR vs GFDR at alpha = .05 as a function of m (proportion of false null hypotheses) and power.



$m = .05; .10; .33; .50$

Idea Behind Q_{EA}

Let $K = 10,000$.

Let $\alpha = .05$.

Let $R = 1,000$.

Then we expect to get 500 significant results just by chance. Thus, $Q_G = 500/1,000 = .5$. This means we think 500 significant results were mistakes and 500 were 'real.'

But if 500 were real, then there were only 9,500 plausible null hypotheses and then we would expect only 475 significant results by chance.....

...and so on.

Idea Behind Q_{EA} - II

Let K_j be the plausible number of true null hypotheses at the j^{th} iteration.

For $j = 0$, $K_j = K$.

For $j > 0$, $K_j = K - (R - \alpha K_{j-1})$

This will converge at a value K^* when:

$$K_{j-1} = K - R + \alpha K_{j-1}$$

Therefore,

$$K^* = (K-R)/(1-\alpha)$$

$$Q_{EA} \equiv \frac{\frac{K-R}{1-\alpha} \alpha}{R}$$

Q_{EA} vs Q_G

$$Q_{EA} \equiv \frac{\frac{K-R}{1-a}a}{R}$$

$$Q_G \approx FDR \frac{1}{1-m},$$

$$Q_{EA} = FDR \frac{1-c-d}{a}$$

Therefore, if

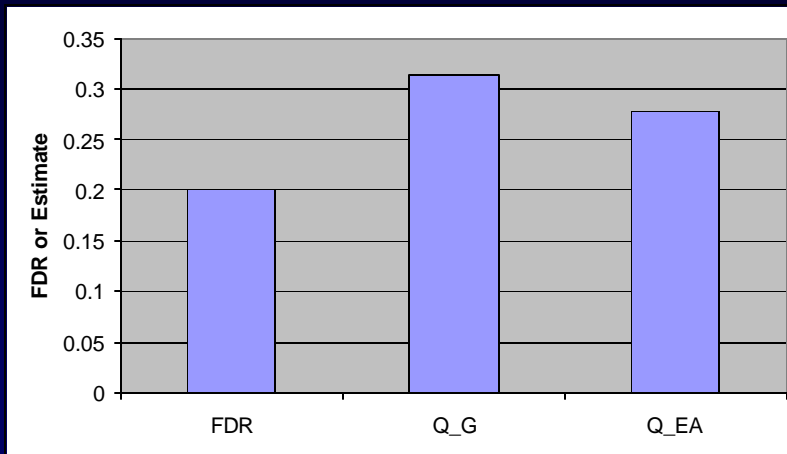
$$1 \leq \frac{1-c-d}{a} < \frac{1}{1-m}$$

Q_{EA}

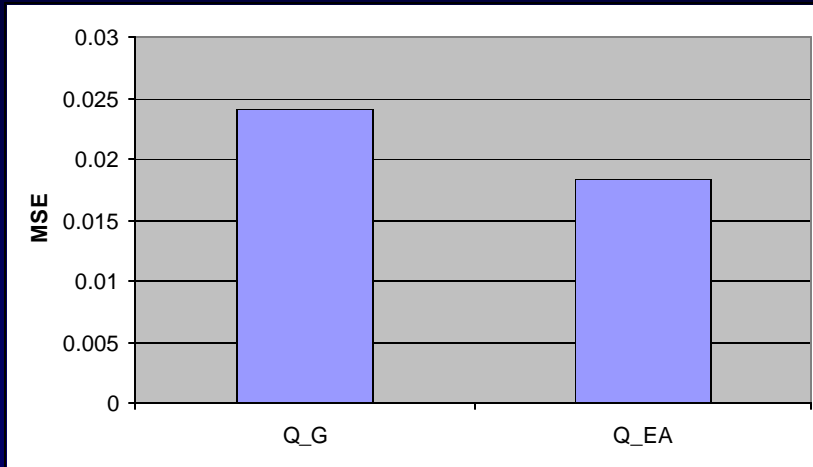
May have advantages over

Q_G

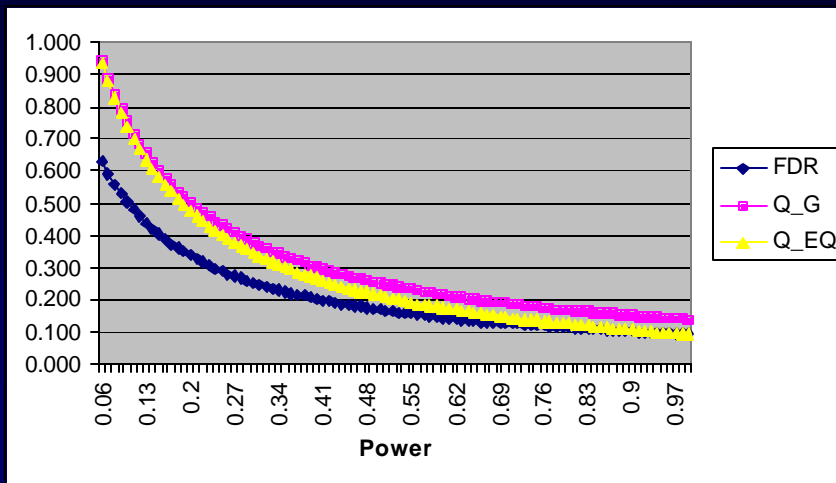
Simulation (10,000 datasets of 40,000 genes each; $a=.05$) - I



Simulation (10,000 datasets of 40,000 genes each) ; $\alpha = .05$ - II



Q_{EA} vs Q_G $\alpha = .05$ as a function of power. $\underline{m} = .33$.



Some of The Challenges We Face

- ◆ Massive Multiple Testing
 - Inference
 - Designing studies with many hypotheses
 - Estimation
 - Model Selection
- ◆ Lack of a null hypothesis
- ◆ Keeping Our Own Field Solid.

EDR Defined

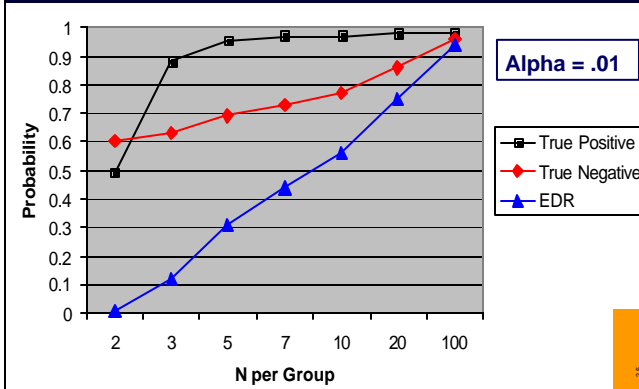
		Truth		
		Null	Alt	
Conclusion	Null	a	b	K-R
	Alt	c	d	R
		K-M	M	K

$$EDR = E \left(\frac{d}{b+d} \right)$$

Gadbury et al.
(submitted)

A parametric bootstrap approach to sample size estimation for microarray research

Gadbury et al (submitted)

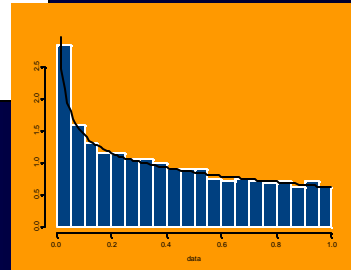


Alpha = .01

To determine targets of TNF signaling of NF- κ B in RASF, replicate samples of RASF transfected with Ad κ B-DN (n=3) or control AdTet (n=3).

Steps:

1. Fit a model to a set of observed data.
2. Simulate data from a model with the estimated parameters and observe what happens.
3. Repeat step 2 m times.
4. Summarize results.



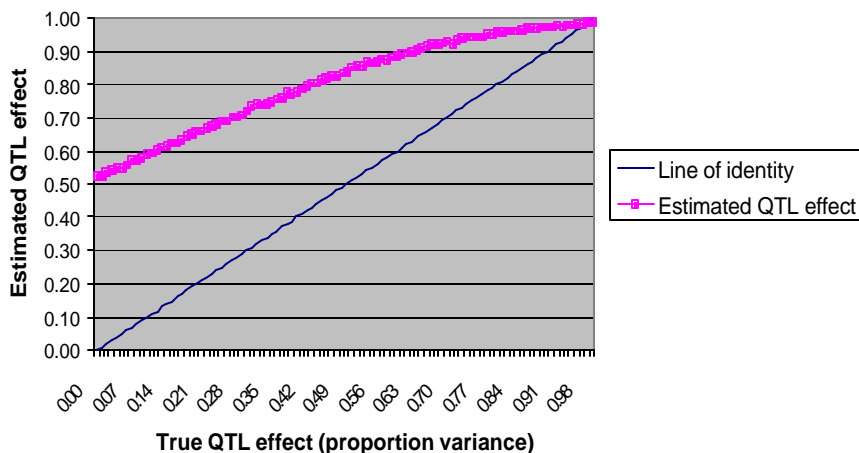
The problem of error has preoccupied philosophers since the earliest antiquity. According to the subtle remark made by a famous Greek philosopher, the man who makes a mistake is twice ignorant, for he does not know the correct answer, and he does not know that he does not know it.

Borel, Emile
Probability and Certainty

Some of The Challenges We Face

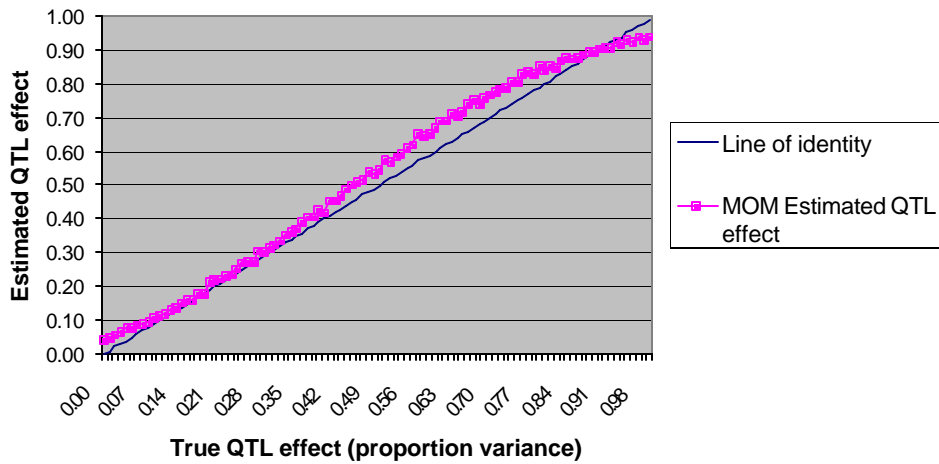
- ◆ Massive Multiple Testing
 - Inference
 - Designing studies with many hypotheses
 - Estimation
 - Model Selection
- ◆ Lack of a null hypothesis
- ◆ Keeping Our Own Field Solid.

Figure 2. Estimated vs. true QTL effect when selecting only significant results



Allison et al. (2002) *American Journal of Human Genetics*, 70, 575-585

Figure 4. MOM Estimated vs. true QTL effect when selecting only significant results from simulation 3.



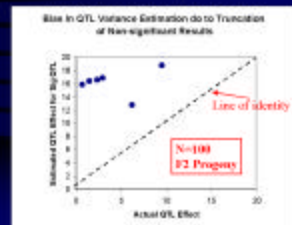
Allison et al. (2002). *American Journal of Human Genetics*, 70, 575-585.

Empirical Bayes (EB) Estimation of Gene-Specific Effects

- ◆ Why
- ◆ Who
- ◆ How

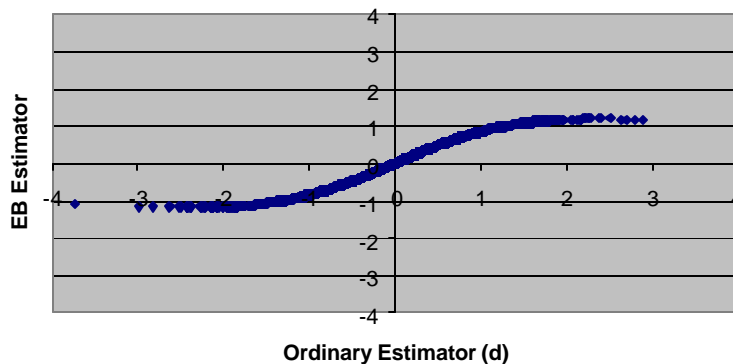


From Beavis, W.D. (1996).
In *Molecular Dissection of Complex Traits*. CRC Press.



Empirical Bayes (EB) Estimation of Gene-Specific Effects - II

Gene Expression Differences 5 Insulin Sensitive vs. 5
Insulin Resistant Humans
(Data Courtesy Paska Permana - NIDDK).



Some of The Challenges We Face

- ◆ Massive Multiple Testing
 - Inference
 - Designing studies with many hypotheses
 - Estimation
 - Model Selection
- ◆ Lack of a null hypothesis
- ◆ Keeping Our Own Field Solid.

Multifactor Dimensionality Reduction (MDR)

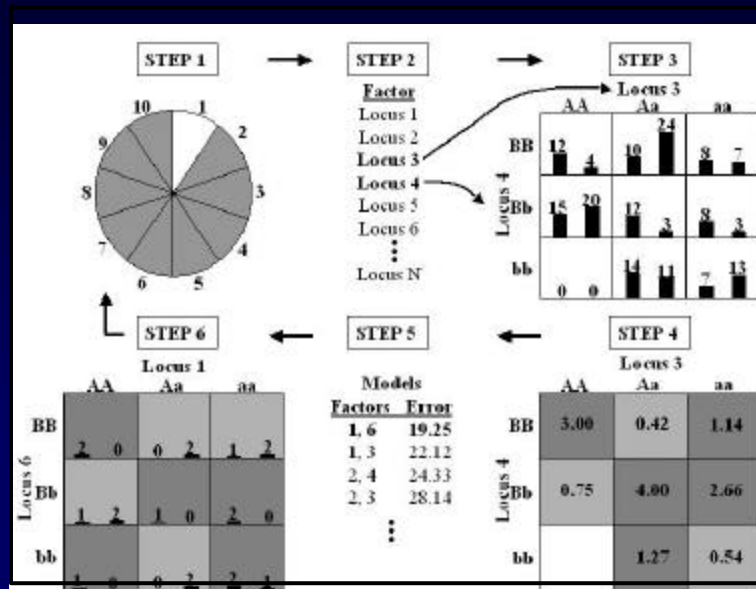
Ritchie et al., *American Journal of Human Genetics* 69:138-147 (2001)
Hahn et al., *Bioinformatics* 19:376-382 (2003)
Ritchie et al., *Genetic Epidemiology* 24:150-157 (2003)

- ◆ Criteria for significance
 - Evaluate all 2-locus, 3-locus, ... , n -locus models
 - Select a best model that
 - Maximizes the cross-validation consistency
 - Minimizes the prediction error
 - Is most parsimonious (if necessary)
 - Use permutation testing to estimate empirical p -value

© Jason H. Moore

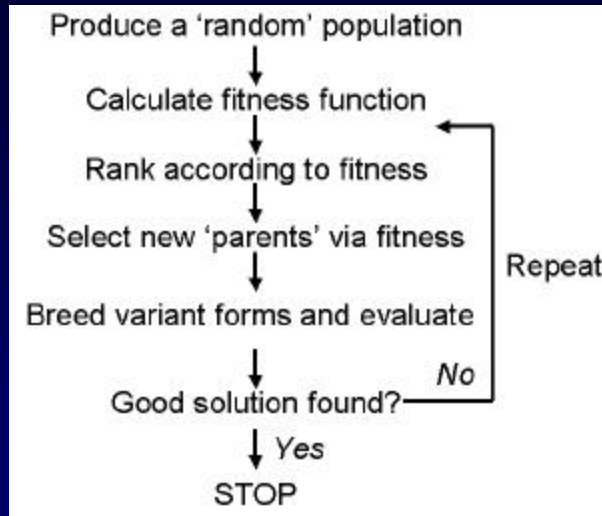
Multifactor Dimensionality Reduction (MDR)

Ritchie et al., *American Journal of Human Genetics* 69:138-147 (2001)



© Jason H. Moore

The Basic Strategy of Evolutionary Computing



Adapted from Davison: <http://statwww.epfl.ch/davison/teaching/Microarrays/>

"Pluralitas non est ponenda sine neccesitate"

~~ William of Occam

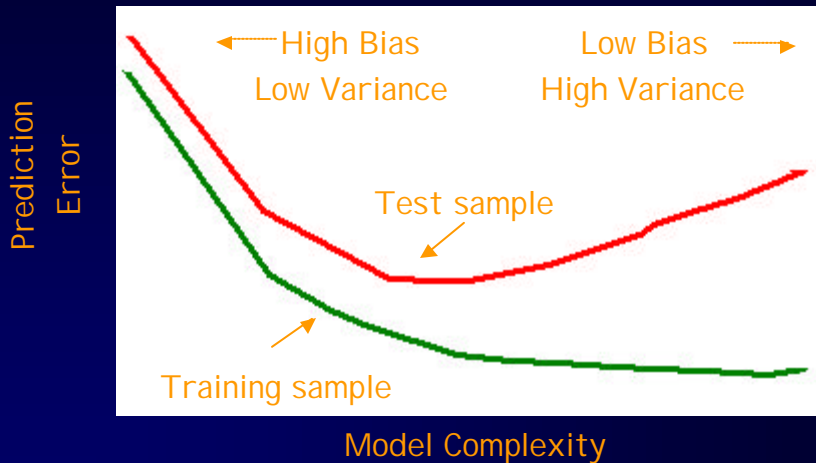
"Make things as simple as possible, but no simpler."

~~ Albert Einstein

"Science may be described as the art of systematic over- simplification."

~~ Karl Popper

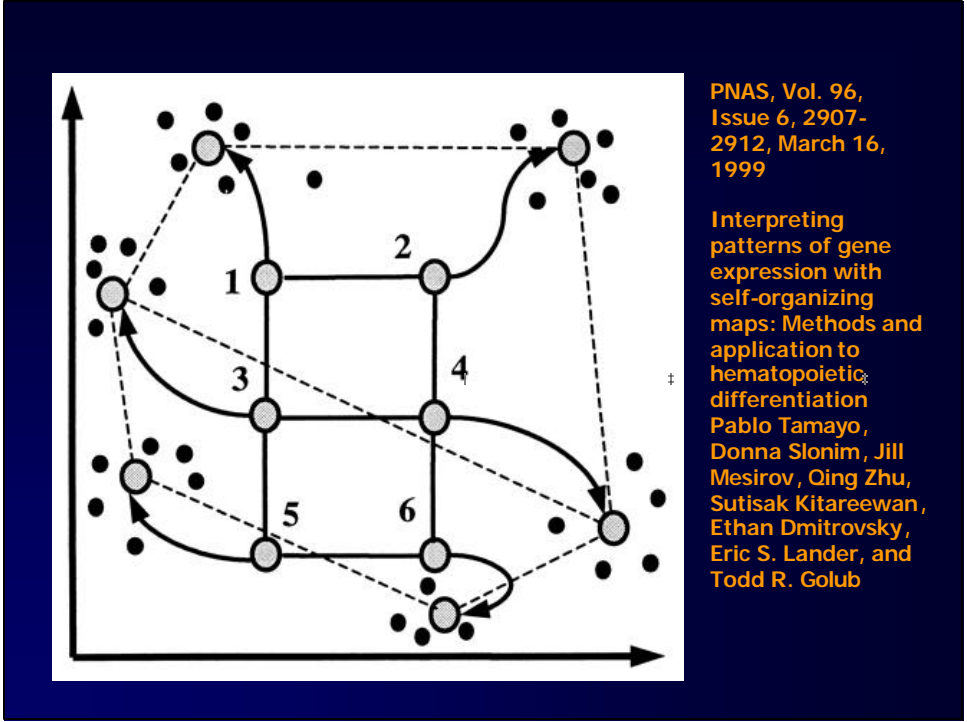
Prediction and Variable Selection: Error vs. Complexity or *What is the optimal amount of stupidity?*



Adapted from Davison: <http://statwww.epfl.ch/davison/teaching/Microarrays/>

Some of The Challenges We Face

- ◆ Massive Multiple Testing
 - Inference
 - Designing studies with many hypotheses
 - Estimation
 - Model Selection
- ◆ Lack of a null hypothesis
- ◆ Keeping Our Own Field Solid.



PNAS, Vol. 96,
Issue 6, 2907-
2912, March 16,
1999

Interpreting
patterns of gene
expression with
self-organizing
maps: Methods and
application to
hematopoietic
differentiation
Pablo Tamayo,
Donna Slonim, Jill
Mesirov, Qing Zhu,
Sutisak Kitareewan,
Ethan Dmitrovsky,
Eric S. Lander, and
Todd R. Golub

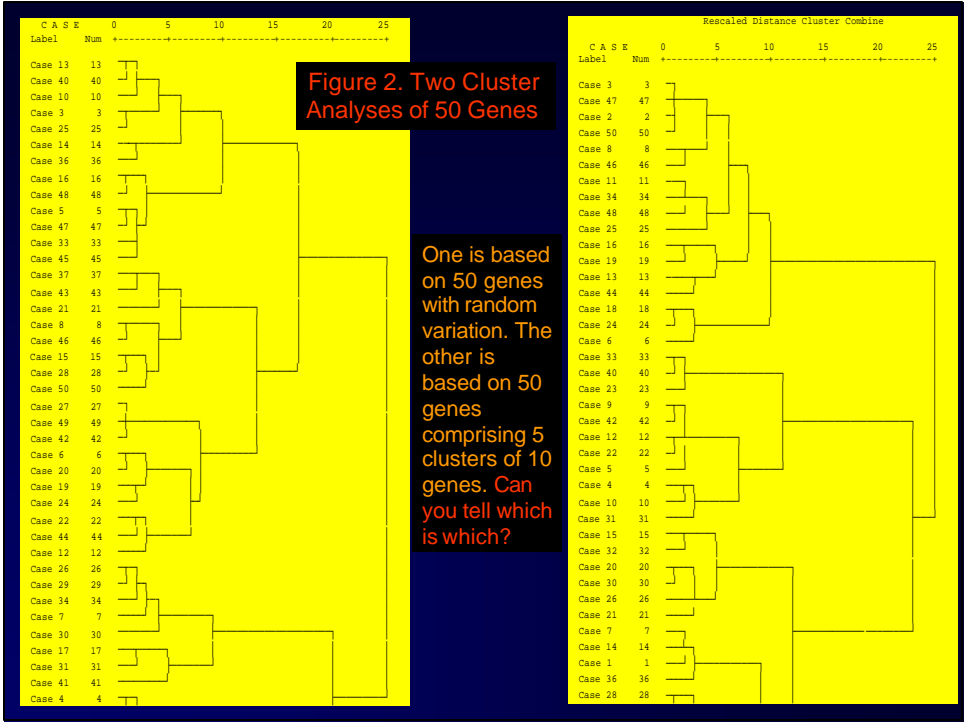


Figure 2. Two Cluster Analyses of 50 Genes

One is based on 50 genes with random variation. The other is based on 50 genes comprising 5 clusters of 10 genes. Can you tell which is which?

Cluster Analysis: Four Caveats

1. Its Not New.

Czekanowski, J.
Objectiv kriterien in der ethnologie.
Korrespondenzblatt der Deutschen Gesselschaft fur Anthropologie, Ethnologie, und Urgeschichte, 1911, 47, 1-5.

Harsh, C. M. Three applications of cluster analysis to an annoyance study. *Psychological Bulletin*. 33, 1936, 773.

2. It has well-recognized problems.

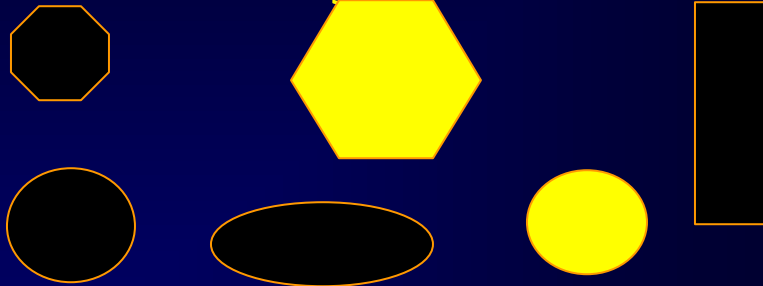
[Everitt, B. S.](#) **Cluster analysis:** A brief discussion of some of the problems. *Brit J Psychia*, 120, 1972, 143-45.

"...availability of computer calculation methods has led many psychiatrists to using the new data techniques uncritically. Spurious findings may result. The available clustering techniques should be validated, as by applying them to sets of data of known structure. Most present methods may be defective."

3. It can be very computationally demanding.

4. It may not answer a particular investigator's questions.

A Challenge for Classification: Lack of a Null Hypothesis.



If there is no right answer, how do we know if we got a good one?

Some of The Challenges We Face

- ◆ Massive Multiple Testing
 - Inference
 - Designing studies with many hypotheses
 - Estimation
 - Model Selection
- ◆ Lack of a null hypothesis
- ◆ Keeping Our Own Field Solid.

Epistemological Foundations

- The word *science* comes from the Latin *scientia*, meaning having knowledge.
- Epistemology: the study of how we come to have and what constitutes knowledge.
- The biological sciences generally use the empirical method and inductive reasoning.
- One way in which such inferences, predictions, and estimations are made from data is via use of statistical techniques.
- Thus, our confidence in the biological knowledge we obtain from examination of data stems, in part, from our confidence in the validity of the statistical methodology that we apply to those data.
- Given a set of statistical procedures judged to be valid, a sound epistemological foundation for biological science comes, in part, from the application of those procedures.
- But how do we derive knowledge about the validity of our statistical methods such that they are also enjoy a solid epistemological foundation?

From Mehta, Tanik, & Allison (in preparation).

Method Validation

A Circular & Epistemologically Invalid Framework

- ◆ Application to single or a few real data sets.

Two Epistemologically Valid Frameworks: Induction & Deduction

- ◆ Deduction: i.e., mathematical proof.
- ◆ Induction: simulations
- ◆ Combinations of the two

From Mehta, Tanik, & Allison (in preparation).

