THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

David B. Allison

Dept of Biostatistics, Section on Statistical Genetics

&

Dept of Nutrition Sciences, Clinical Nutrition Research Center

Supported by:

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

NSF – RCN on Microarray Analysis (Allison)

NIA – P01 & R01 (Weindruch)

NCI – P20 Planning Grant (Barnes)

NIEHS – R01 on Stat Genet Methods (Amos)

Recruitment actively underway for post-doctoral fellows, research track faculty, and tenure track faculty!

STATISTICAL ISSUES IN MICROARRAY GENE EXPRESSION STUDIES

Contributors

- Moonseong Heo
- Jose Fernandez
- Gary Gadbury
- Tom Prolla
- CK Lee
- Rick Weindruch

Collaborators

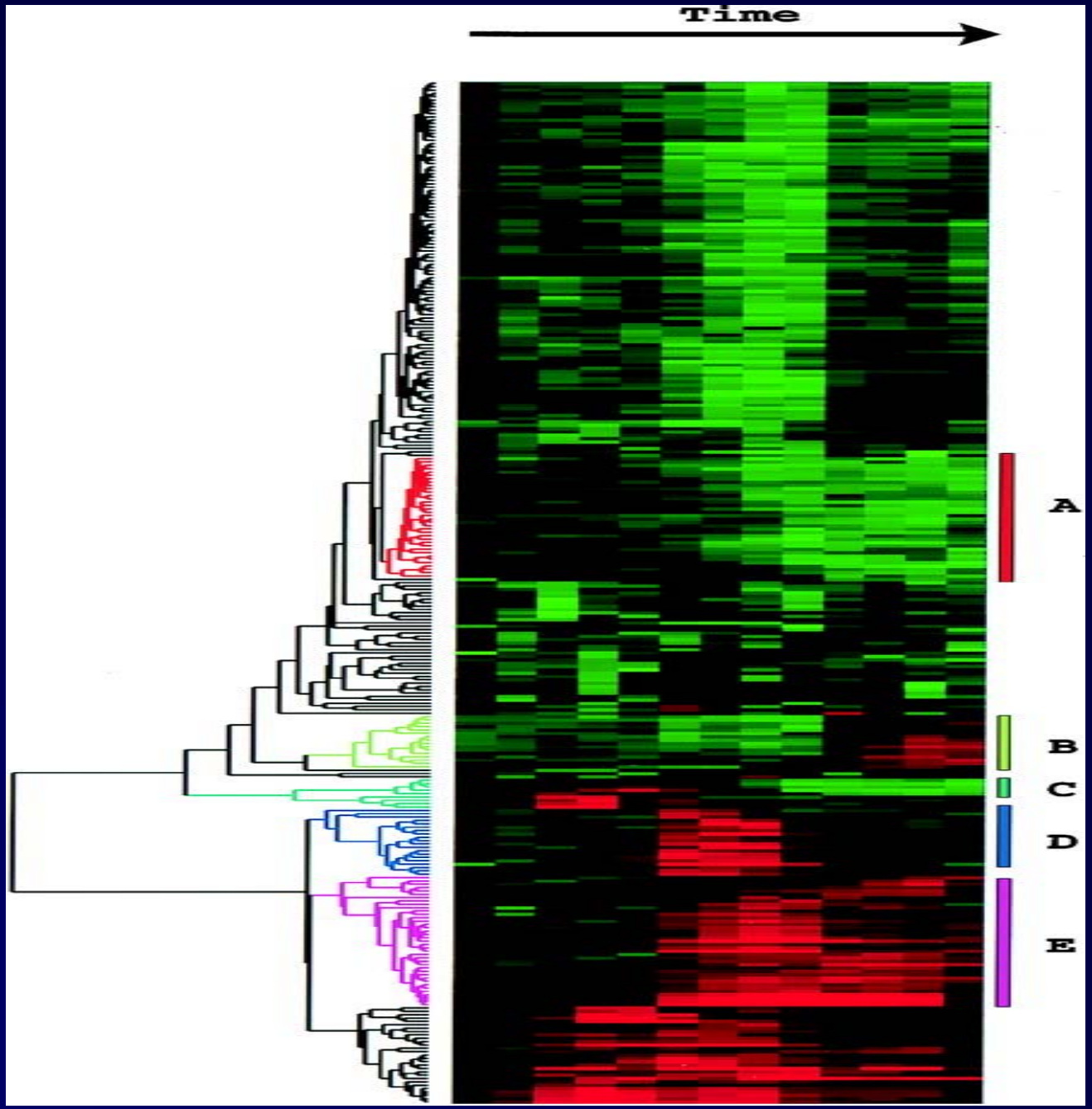
- Bill Grundy
- Anthony Ferrante
- Antonio Tataranni
- Kei Cheung
- Ken Williams
- Bob Eckel
- Stephen Barnes
- Brad Yoder
- Susan Sell

Gene Expression Profile of Aging and Its Retardation by Caloric Restriction

Cheol-Koo Lee,[1,3] Roger G. Klopp,[2]
Richard Weindruch,[4*] Tomas A. Prolla[3*]

The gene expression profile of the aging process was analyzed in skeletal muscle of mice. Use of high-density oligonucleotide arrays representing 6347 genes revealed that aging resulted in a differential gene expression pattern indicative of a marked stress response and lower expression of metabolic and biosynthetic genes. Most alterations were either completely or partially prevented by caloric restriction, the only intervention known to retard aging in mammals. Transcriptional patterns of calorie-restricted animals suggest that caloric restriction retards the aging process by causing a metabolic shift toward increased protein turnover and decreased macromolecular damage.

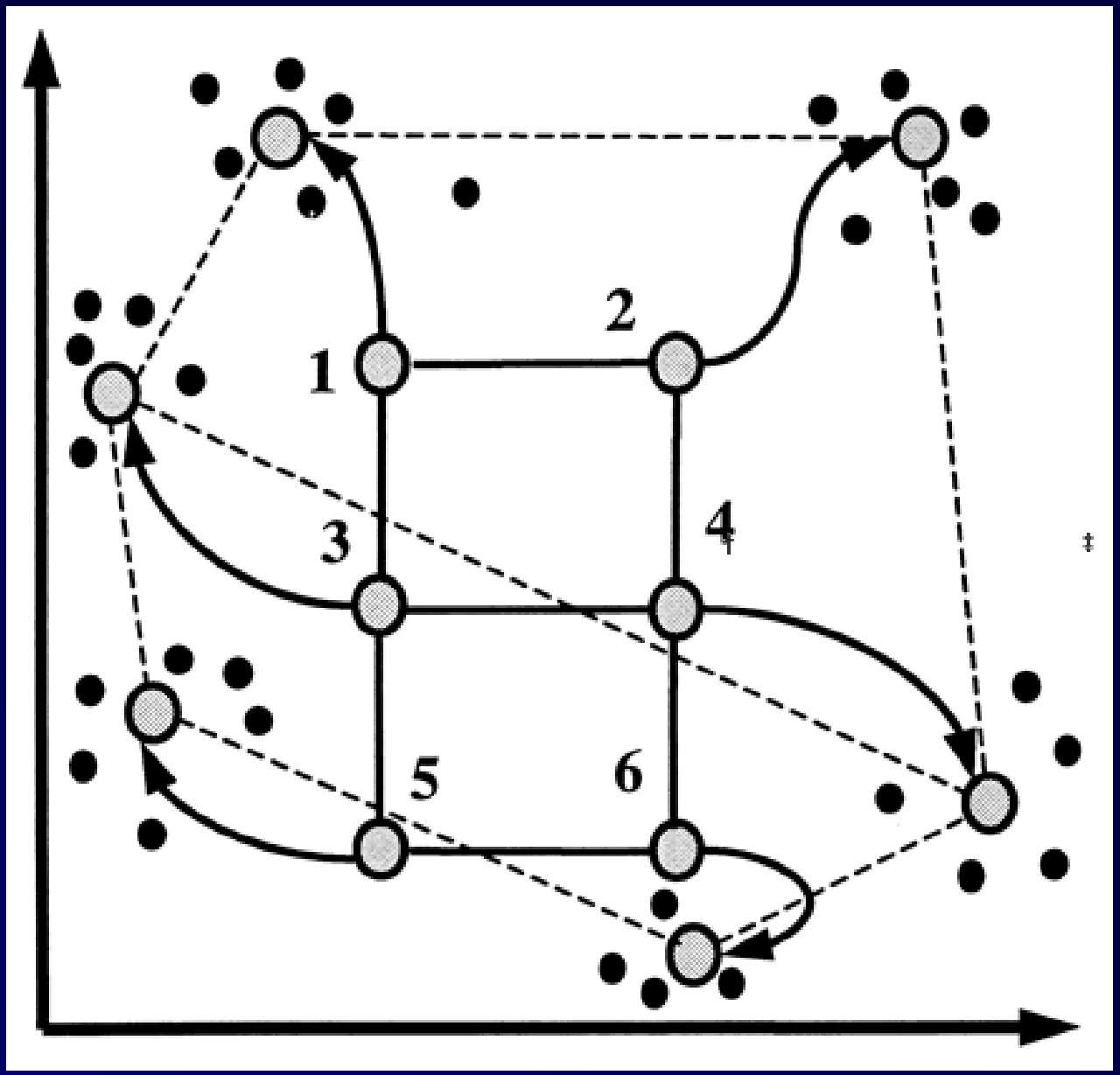
27 AUGUST 1999 VOL 285 SCIENCE www.sciencemag.org



PNAS, Vol. 95, Issue 25, 14863-14868, December 8, 1998

Cluster analysis and display of genome-wide expression patterns

Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein

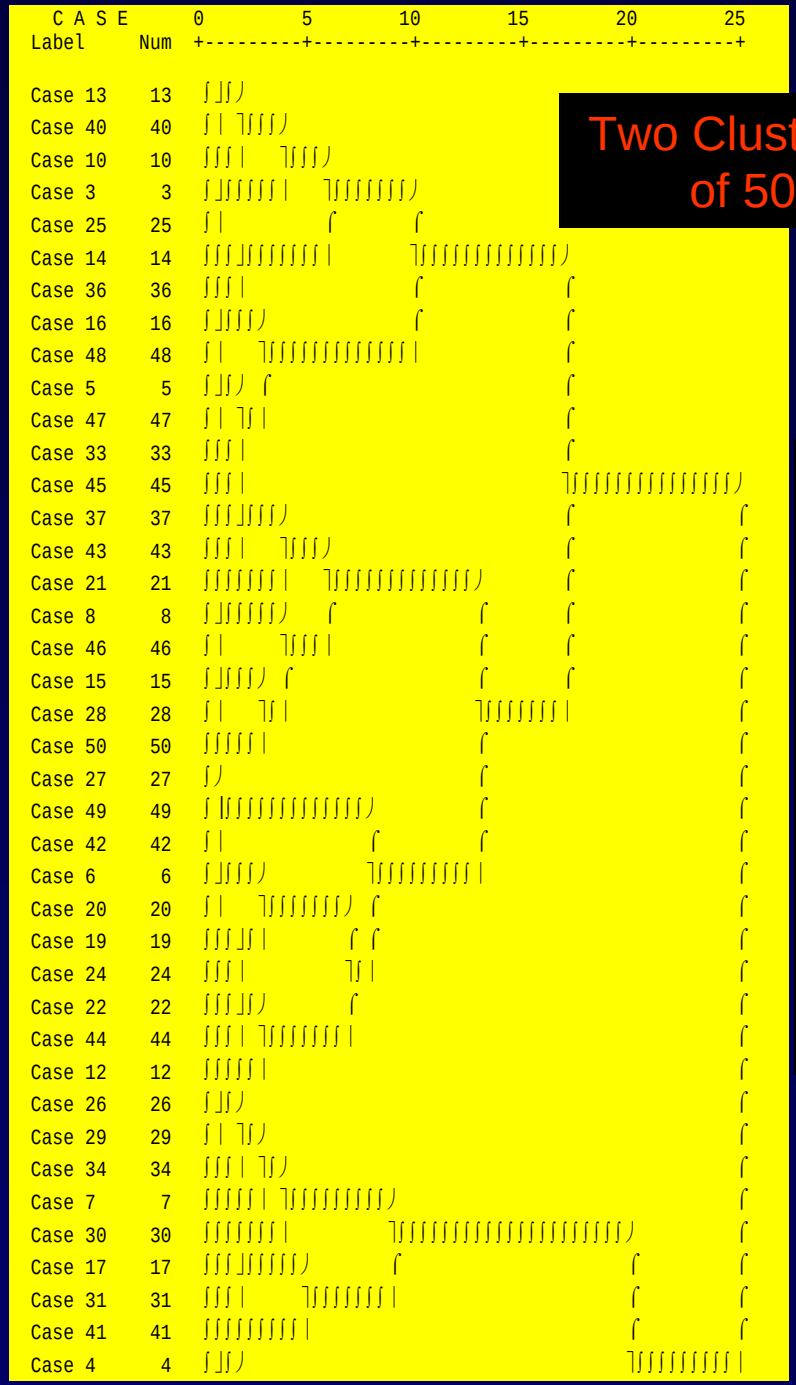


PNAS, Vol. 96, Issue 6, 2907-2912, March 16, 1999

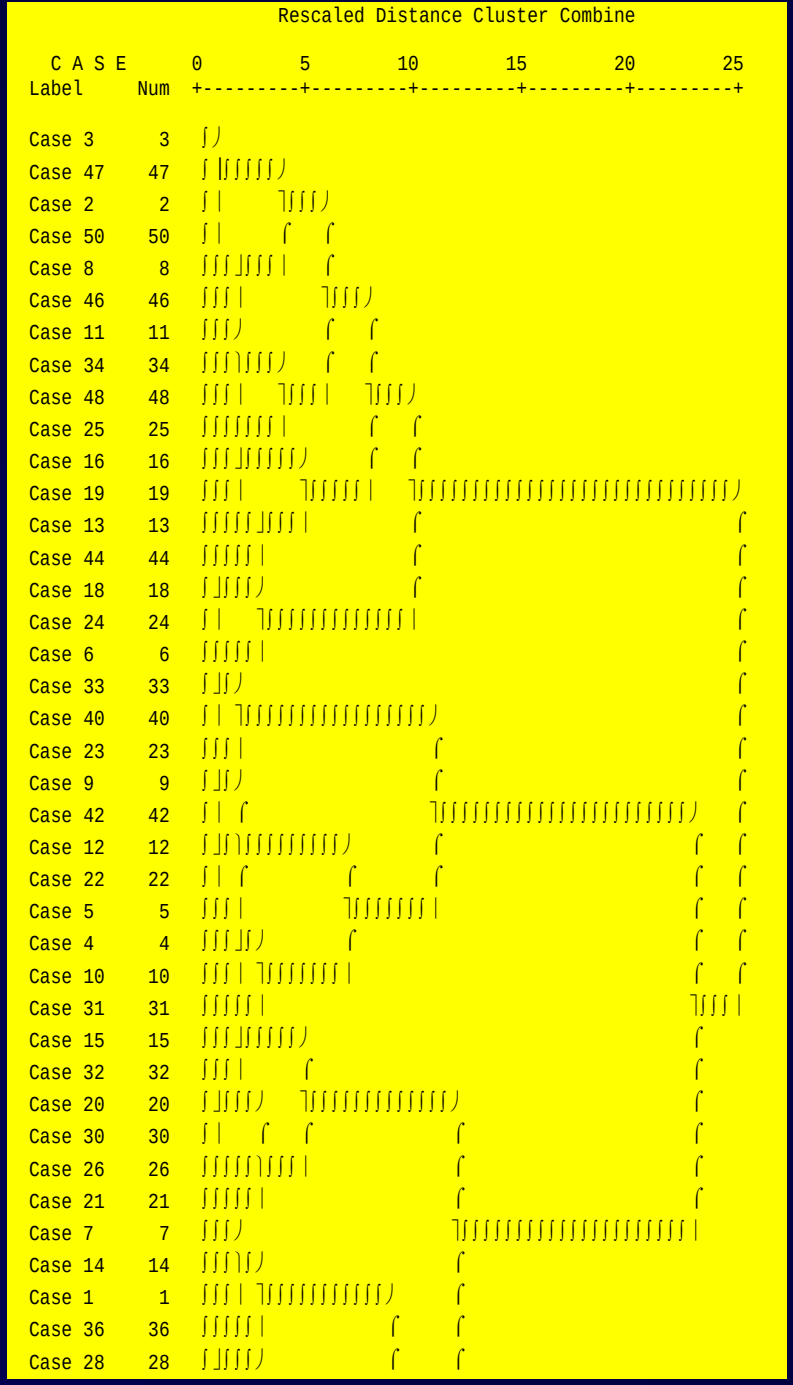
Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation

Pablo Tamayo, Donna Slonim, Jill Mesirov, Qing Zhu, Sutisak Kitareewan, Ethan Dmitrovsky, Eric S. Lander, and Todd R. Golub

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.

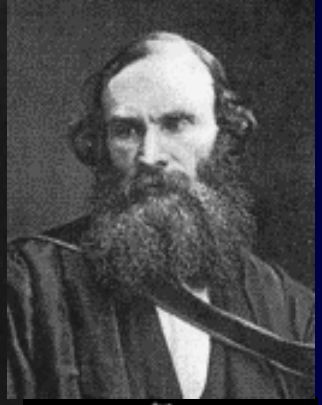
"...availability of computer calculation methods has led many psychiatrists to using the new data techniques uncritically. Spurious findings may result. The available clustering techniques should be validated, as by applying them to sets of data of known structure. Most present methods may be defective."

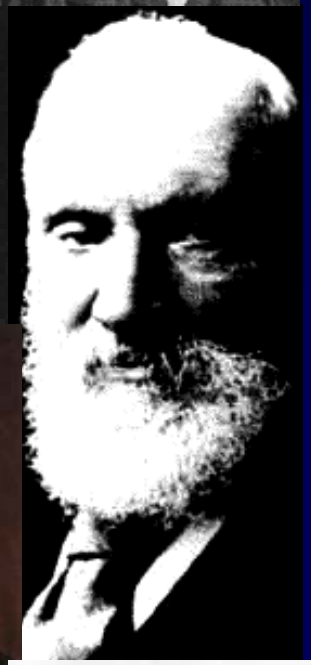
3. It can be very computationally demanding.

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

The Major Activities of Statisticians/Data Analysts

- Measurement
- Design
- Inference
- Estimation
- Classification



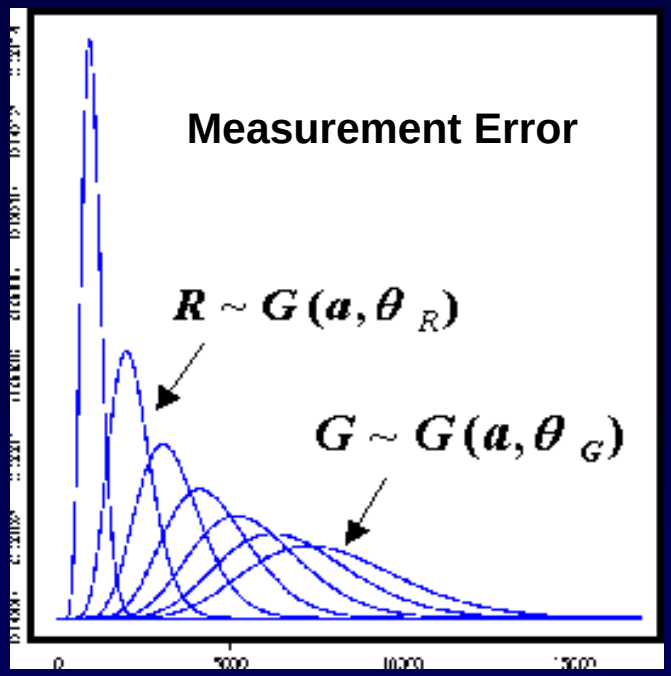


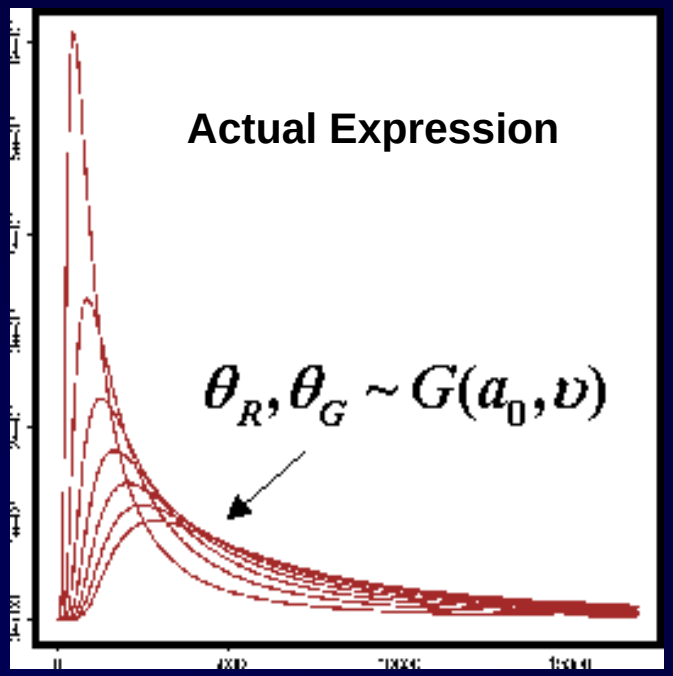
Measurement

"In physical science the first essential step in the direction of learning any subject is to find principles of numerical reckoning and practicable methods for measuring some quality connected with it. I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be."

William Thompson
(Lord Kelvin)

Estimation of differential expression by Bayesian Techniques





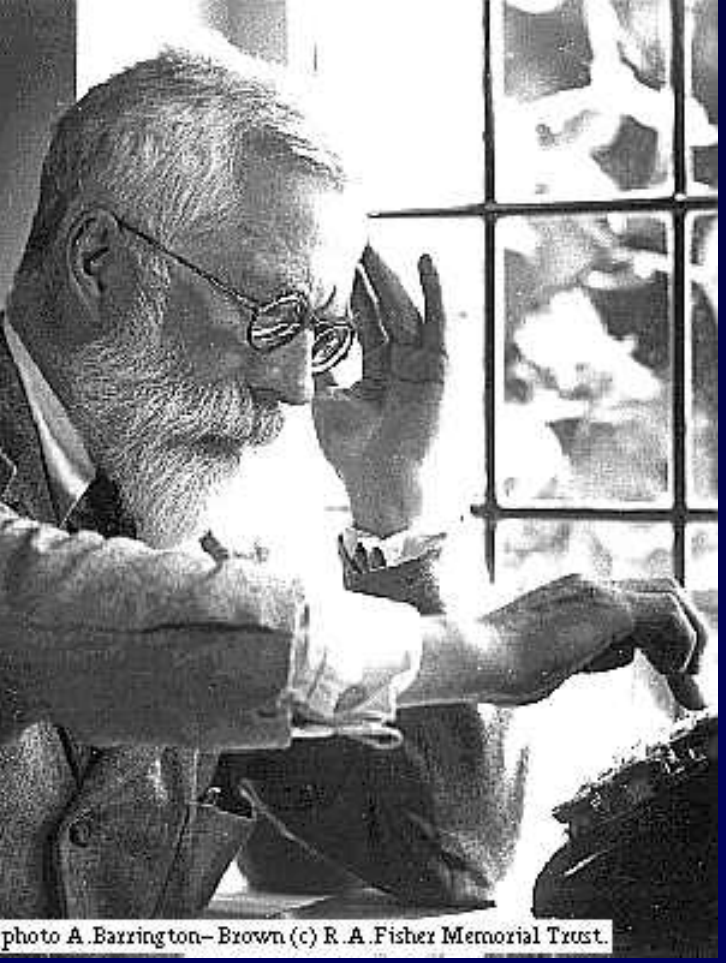
Within this Gamma-Gamma model set-up, the posterior distribution of true differential expression can be derived:

$$p(\rho \mid R, G, a, a_0, \upsilon) \propto \rho^{-(a+a_0+1)} \left\{ \frac{1}{\rho} + \frac{G+\upsilon}{R+\upsilon} \right\}^{-2(a+a_0)}$$

The Bayes estimate (BE), which lies between the mean and the mode of this distribution, is used as an estimate of the true differential expression:

$$BE = \frac{R+\upsilon}{G+\upsilon}$$

Newton et al. *J Computational Biology* (8: 37-52, 2001).



Design.

To consult the statistician after an experiment is finished is often merely to ask him to conduct a *post mortem* examination. He can perhaps say

Pooling Samples on Chips - I

Assumptions

Necessary

The value obtained from a pool of N subjects on a single chip equals the value that would be obtained from the mean of those same subjects measured on separate chips.

For Convenience

Data are normally distributed.

N per condition is equal and evenly divisible by 2.

Homogeneity of variance.

Pooling Samples on Chips - II

Non-pooled

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{N} + \frac{S_2^2}{N}}} \sim t_{2N-2}$$

Pooled into 2 equal groups per condition

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{2} + \frac{S_2^2}{2}}} \sim t_2$$

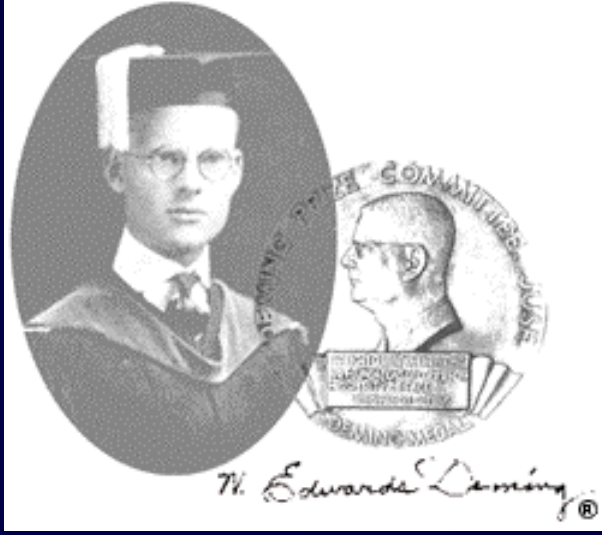
Non-Centrality Parameter

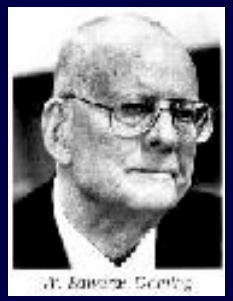
$$\frac{\mu_1 - \mu_2}{\sqrt{\frac{2\sigma^2}{N}}}$$

$$\frac{\mu_1 - \mu_2}{\sqrt{2\left[\frac{\sigma^2 / N/2}{2}\right]}} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{2\sigma^2}{N}}}$$

Inference Requires Knowledge of Variation

variation, including appreciation of a stable system, and some understanding of special causes and common causes of variation, are essential for management of a





-- W. Edwards Deming

The Fallacy of a Fixed 'Fold-Change' (or other variance-free metric) Cut-off for Significance

1. A cut-off that does not vary with sample size implies that the variability of the metric under the null hypothesis is the same regardless of the number of observations on which it is calculated. This can easily be proven to be in contradiction to several well established statistical theorems.

2. Even if the sample size is constant, this implies that the variability in expression measurements is constant across conditions, species, genes, etc. This seems implausible and is contradicted by data:

Human Brain Tissue	Mouse Adipose Tissue
3 Samples in Duplicate Aliquots	3 Samples in Duplicate Aliquots
Clontech System	Columbia 2-dye system
Gene to gene correlation	Gene to gene correlation
in absolute difference = **.63**	in absolute difference = **.57**

Data courtesy of Antonio Tataranni & Anthony Ferrante

Some Key Questions to Address

- **Is the difference in expression for the *such-and-such* gene statistically significant?**
- **Is there statistically significant evidence that <u>any</u> of the genes under study exhibit a difference in expression across groups.**
- **What is the best estimate of the number of genes for which there is a true difference in gene expression?**
- **What is the confidence interval around that estimate?**
- **If we set a threshold above which we declare genes '*interesting*' and worthy of follow-up study, what proportion of those genes are likely to be genes for which there is a real difference in expression and what proportion are likely to be false leads?**
- **What proportion of those genes not declared '*interesting*' are likely to be genes for which there is a real difference in expression (i.e., misses or false negatives)?**
- **For genes declared interesting, what is the best estimate of the difference in expression and the confidence interval around that estimate?**

Allison, et al. (in press). Statistical methods for the analysis of microarray gene expression data: a mixture model approach. *Computational Statistics & Data Analysis.*

Assumptions

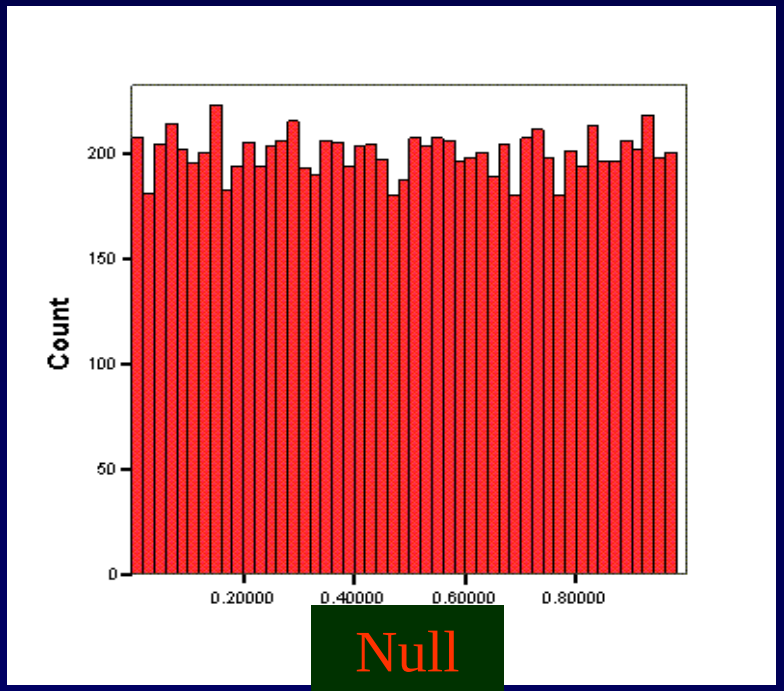
1) For each gene, the measurements of gene expression have a finite population mean and variance;
2) For each gene under study, there is a measure of expression available for each case and this measure has sufficient reliability and validity to be useful (we use the word ‘case’ generically to refer to any organism or tissue on which expression is measured).

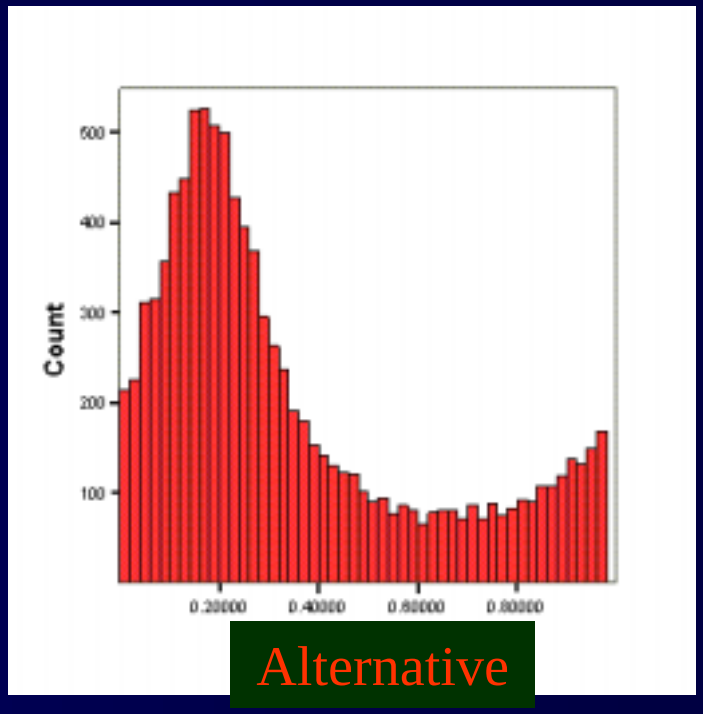
The Challenges When the Sample is not Extremely Large

- **Given low power, many false null hypotheses will be retained and important genes will not be followed-up.**
- **Without the near-perfect power and strict control of α_{ew} offered by extremely large samples, individual gene testing provides no direct information about the number of genes for which there is a true difference in expression, the false positive rate, or the false negative rate.**
- **Ordinary estimates of effect size calculated only for the statistically significant results will be biased upward even when those estimates are maximum likelihood estimates and asymptotically unbiased when considered across all results regardless of significance.**

Mixture Model Approach

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).





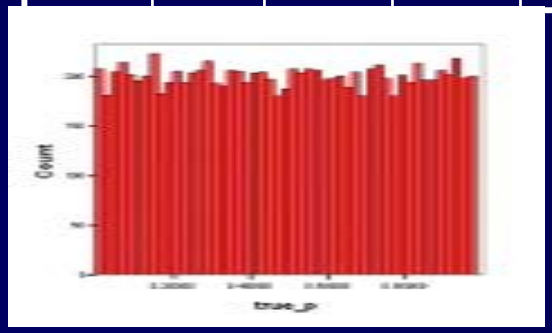
Under the alternative hypothesis, the distribution of p-values will tend to cluster closer to zero than to one.

Why a Mixture Model Approach (MMA)?

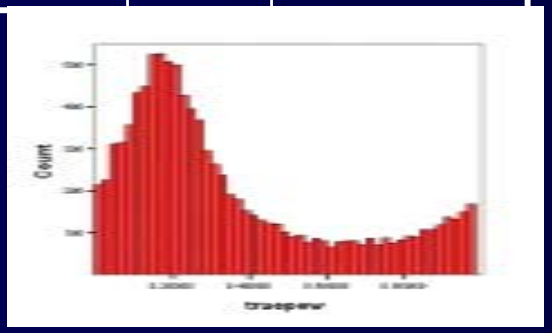
Gene (*i*)	Group A			Group B			P-Values (t-test)
	A1	A2	A3	B1	B2	B3	
1	-741	-947	-1578	-566	-747	-663	0.16982
2	122	124	14	438	420	373	0.00142
3	-18	-68	21	-15	99	-36	0.48675
...	...	...	...	...	...	...	...
6347	59	66	95	-163	-193	-230	0.00027

H_o

H_a



No difference in gene expression



Difference in gene expression

Mixtures of Betas

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

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

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:

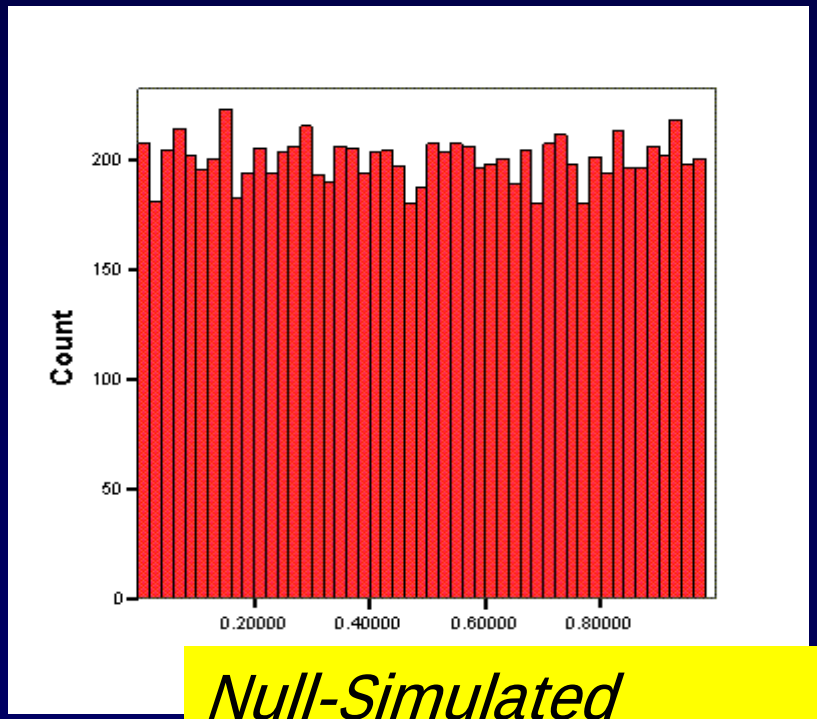
$$L_{V+1} = \sum_{i=1}^{k} \ln\left[\lambda_0 \beta(1,1)(x_i) + \sum_{j=1}^{v} \lambda_j \beta(r_j, s_j)(x_i)\right]$$

x_i is the p-value for the ith test,

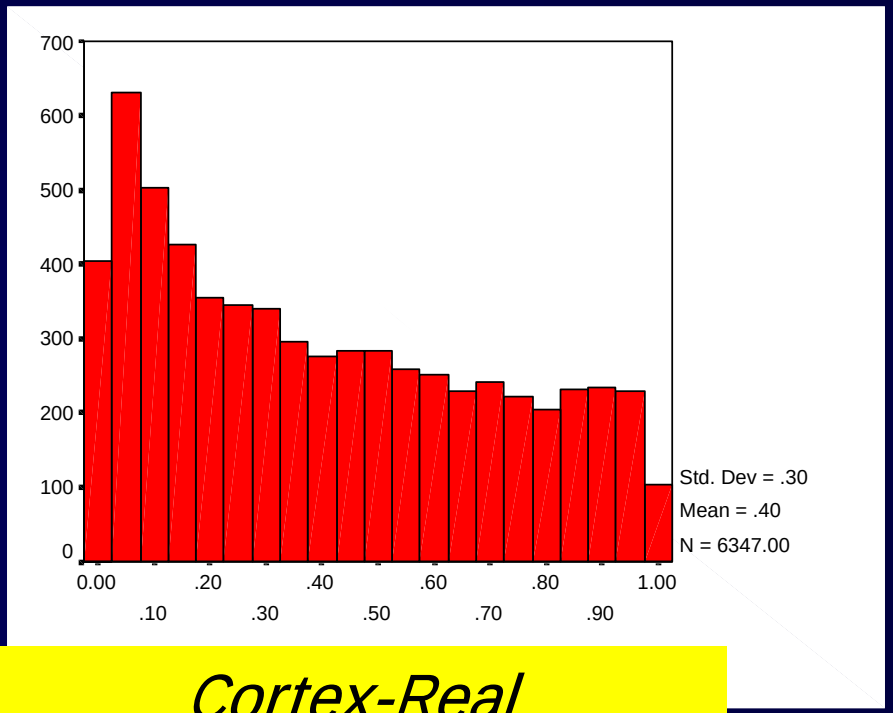
λ_o is the probability that a randomly chosen test from the collection of tests is for a gene for which there is no population difference in gene expression,

λ_j is the probability that a randomly chosen test is for a gene from the jth component distribution for which there is a true population difference in gene expression.

The Weindruch et al. Cortex Data (Old Ad Lib vs. Old CR)



Null-Simulated



Cortex-Real

Estimated parameters: λ_1 =.29; r_1=.78; s_1 =3.87.

Posterior (Bayesian) Probabilities

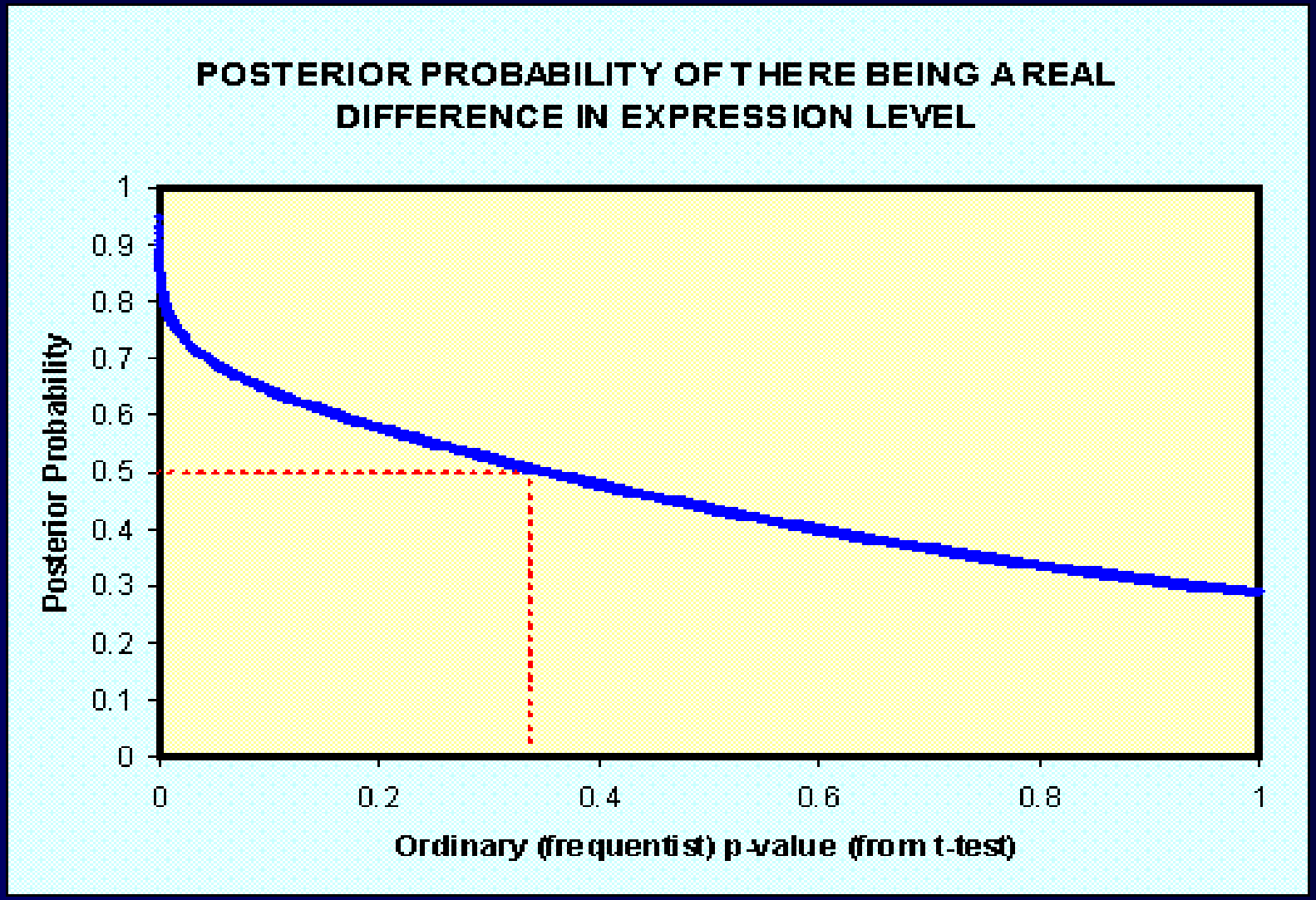
Gene	Group A			Group B			P-Values (t-test
	A1	A2	A3	B1	B2	B3	
1	-741	-947	-1578	-566	-747	-663	0.16982
2	122	124	14	438	420	373	0.00142
3	-18	-68	21	-15	99	-36	0.48675
...	...	...	...	...	...	...	...
6347	59	66	95	-163	-193	-230	0.00027

Gene (*i*)	P-Values (t-test	Posterior Probability
1	0.16982	0.59544
2	0.00142	0.84211
3	0.48675	0.44121
...	...	...
6347	0.00027	0.88561

Posterior Probability

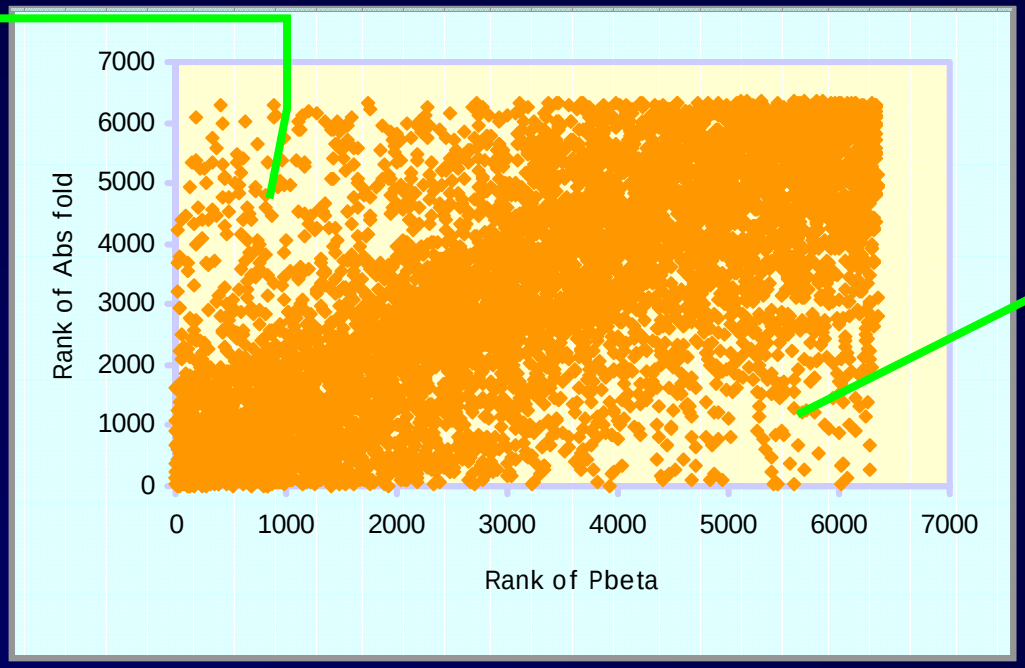
Probability that the gene *i* is a gene from the population of genes for which the null hypothesis is **not** true

Use of posterior probability



Relationship between Fold Change and Posterior Probability

Likely False Positives



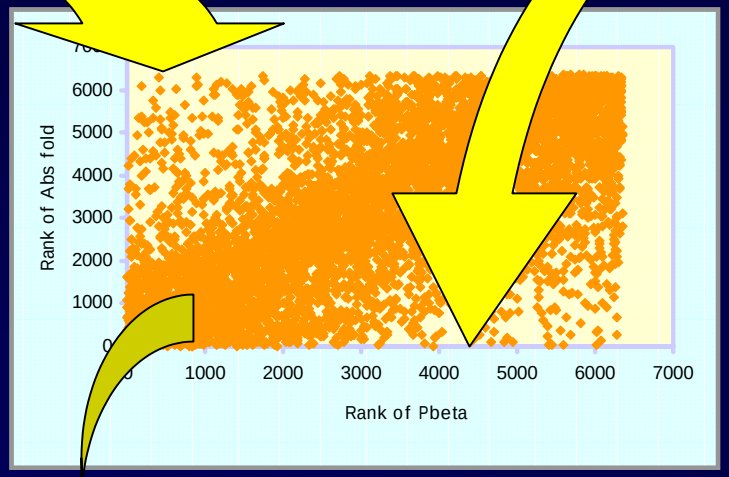
Likely False Negatives

Correlation = 0.671

Some applications of MMA

M74180

- Fold change of 2.7
- Posterior probability of 0.95
- P-value of 0.0001
- Encodes a protein homologous to mouse macrophage stimulating protein



L06451

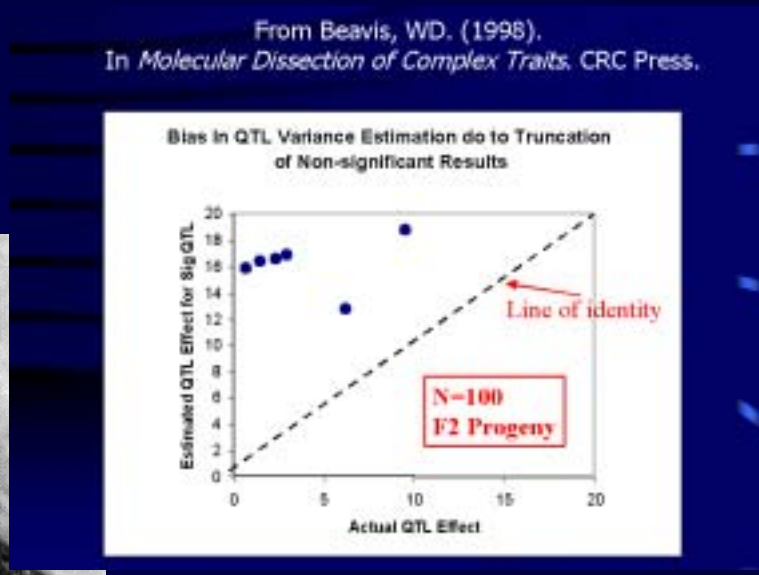
- Fold change of 2.2
- Posterior probability of 0.68
- P-value of 0.06
- Homologous to agouti signaling protein

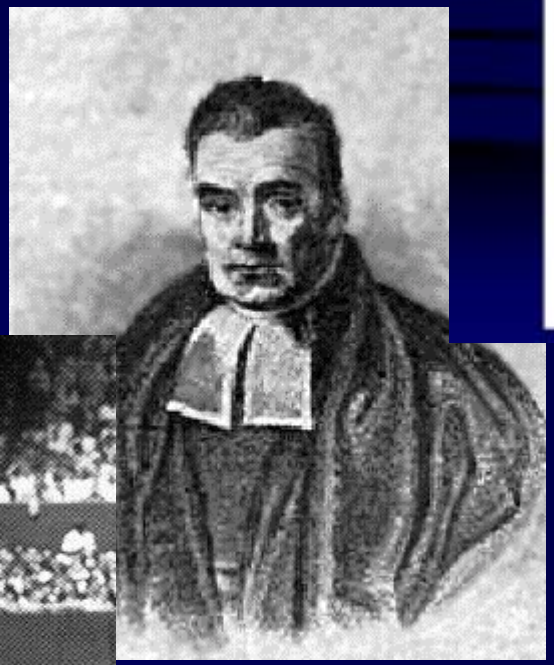
W75705

- Fold change of 3.3
- Posterior probability of 0.44
- P-value of .0.49
- Encodes a protein homologous to mouse cyclophilin

Empirical Bayes (EB) Estimation of Gene-Specific Effects

- Why
- Who
- How





Now it Gets Interesting!

Gene	Estimated Expression Difference	Estimated Linkage QTL	Protein Characteristic	Sequence
1				
2				
3				
4				
….				
k				

Empirical Bayes may be ideally suited.

Where will the computing and bioinformatics power come from?

Suggestions for Advancing

- Provide major infrastructure for analytic software hardening.
- Expand pipeline of statisticians.
- Demystify (we need experts not gurus).
- Convene blue-ribbon panel to issue guidelines (not rules).
- Understand that these are not trivial analytic undertakings (yet) and basic science grantees and reviewers need to be prepared to budget for real help.
- NIH should not need to shoulder this burden alone. This research impacts and is of interest to the EU, FDA, EPA, USDA, NSF and others.

TEXAS
JOSTENS
Will Analyze
Data for
Food