

UAB THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Section ON Statistical Genetics

Statistical Analysis of proteomic data

Grier P Page Ph.D. Associate Professor

Section on Statistical Genetics
Department of Biostatistics
School of Public Health

Gpage@uab.edu
4-4930
Ryals 317D

Section on Statistical Genetics - Microsoft Internet Explorer

Address: http://www.soph.uab.edu/ssg_content.asp?id=1174

UAB SCHOOL OF PUBLIC HEALTH Department of Biostatistics

Section ON Statistical Genetics

Activities Events on Video Linkage & Association Projects Microarray Projects Opportunities People Publications Software

RealPlayer

Statistical analysis of microarrays

UAB THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Section ON Statistical Genetics

Gpage@uab.edu
4-4930
Ryals 317D

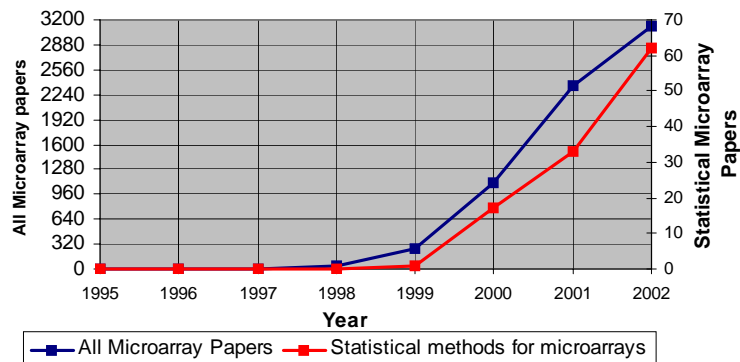
Statistical analysis of microarrays

Grier P Page Ph.D. Assistant Professor
Section on Statistical Genetics
Department of Biostatistics
School of Public Health

Now Playing: rmhigh 351 kbps 0:01 / 42:02

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

Microarray Articles in PubMed



From Mehta, Tanik, & Allison .

A Perspective on Statistics

- We study:
- We wish to obtain knowledge about:

Samples

Data

Populations

Nature

Things Statisticians Do:

Develop Design & Analysis Procedures to Facilitate:

- **Measurement** – (e.g., produce a variable Y' that represents Y).
- **Prediction** – (e.g., 'impute' unobserved values of X using observed Y).
- **Estimation** – (e.g., estimate $\Delta = \mu_1 - \mu_2$).
- **Inference** – (e.g., conclude whether $\delta = 0$).
- **Classification** – (e.g., for $j = 1$ to k , sort the Y_j into $m < k$ groups).

Epistemological Foundations

- Epistemology is the study of how we come to have and what constitutes knowledge.
- 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 also enjoy a solid epistemological foundation?

Method Validation

Epistemologically Valid Frameworks: Induction & Deduction

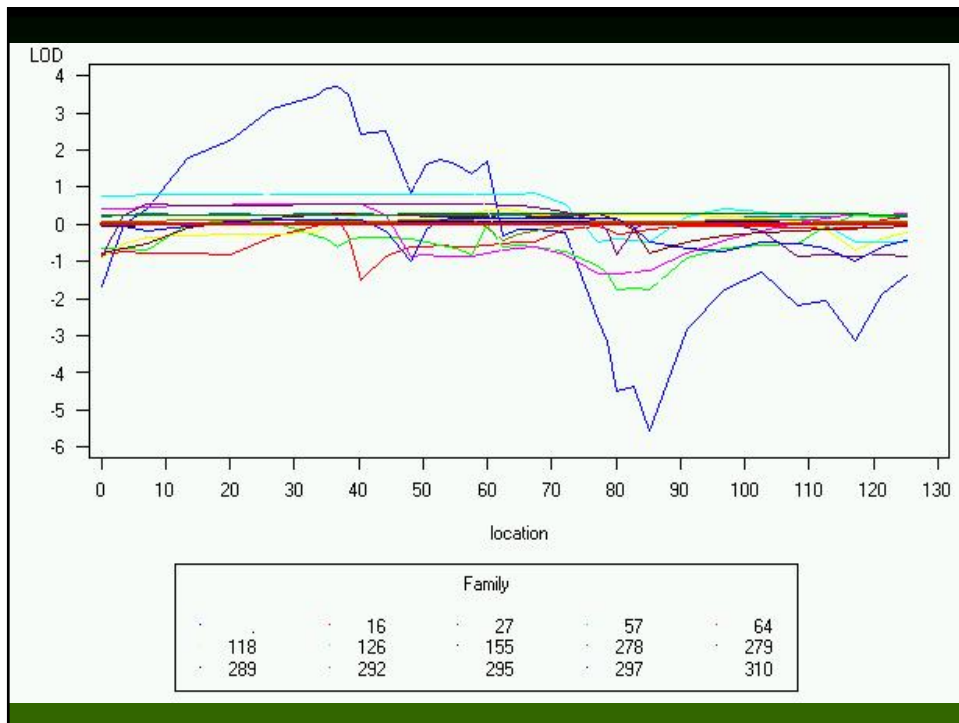
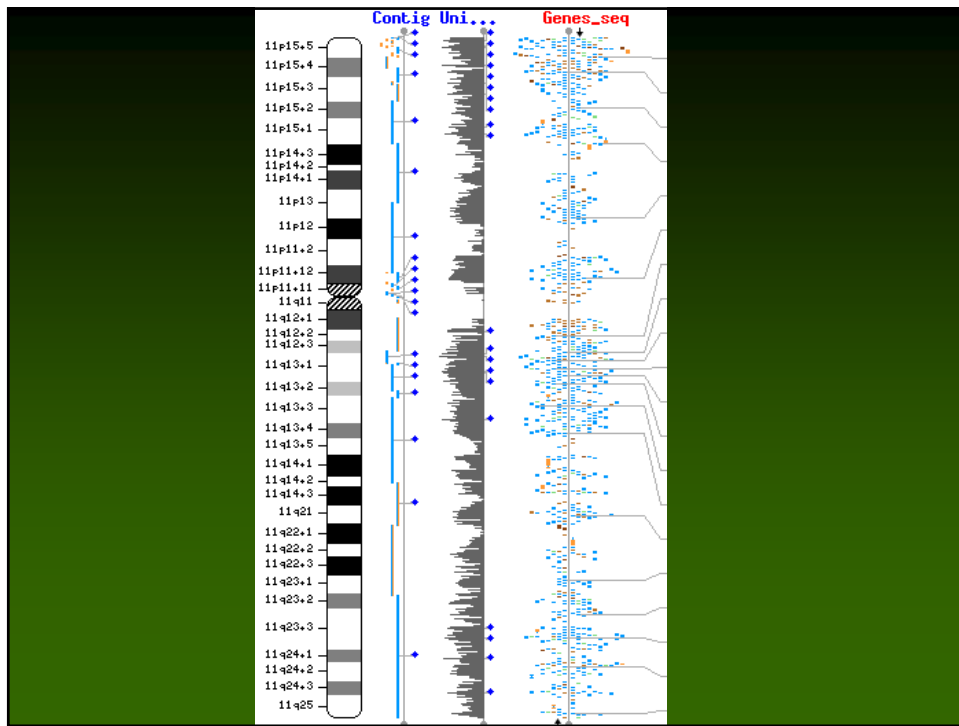
- Deduction: i.e., mathematical proof.
- Induction:
 - Simulations
 - Plasmodes
- Composite Approaches: Application to multiple real data sets of unknown nature with methods of partially known properties.

A Circular & Epistemologically Invalid Framework

- ◆ Application to single real data sets of unknown nature.

What is High Dimensional Biology?

- High Dimensional Biology – is a broad topic covering biological systems where the number of variables is very large.
- Topics that often fall in HDB are microarray, proteomics, linkage, and genomics.
- HDB is also highly collaborative both ‘wet’ and ‘dry’ lab people.



What Do All These Topics Have in Common?

Lots and Lots and Lots of
Numbers !!!

If you have numbers what do you do?

- Statistics (and Design) !
- Or as most of you think Statistics Ugh!

- Most of the statistics used in HDB are identical to statistical methods that have been used for years.
- The thought process that goes into design is also similar to those that have been used for years.

Design

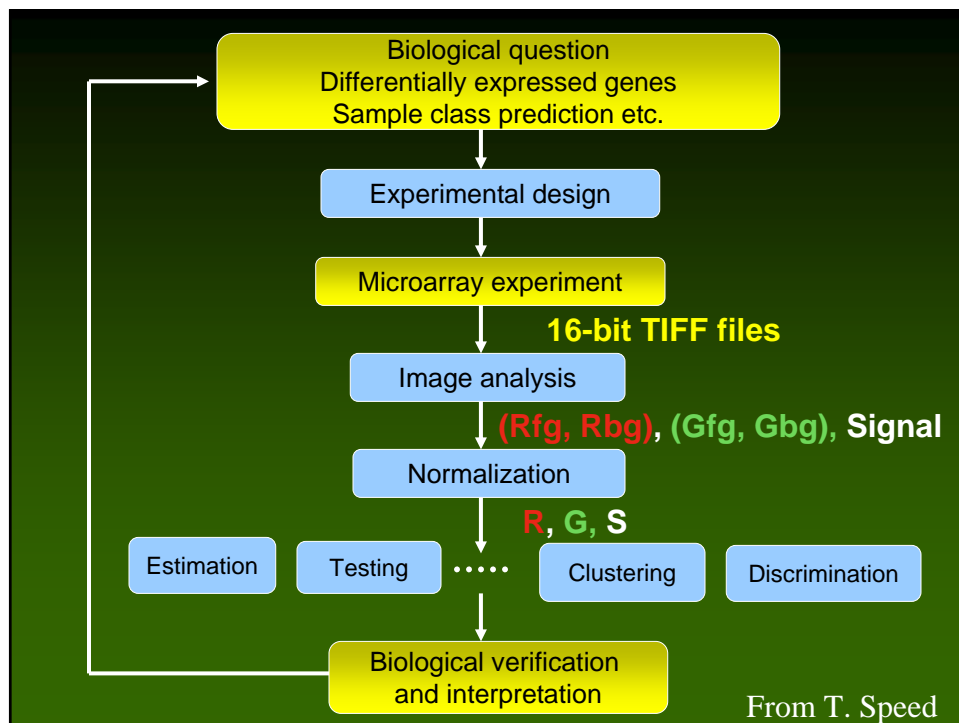
- Design is the art of designing an experiment in such way that the question that is being asked can be easily and unambiguously answered.
- The experimental hypothesis drives the design.

Statistics

- Methods for make inferences about a population as a whole by taking a sample.
- Statistics and design work in harmony with the biology, while design and statistical may be the cause of alterations in experiments, the biology is the *sine qua nome*.

What are Statistics and Design?

- The goal of experimental design and statistical analysis is to allow an investigator to answer the question that they would like to ask correctly and efficiently.
- Often statisticians are a reality check. If you can't explain your experiment to a statistician will it make sense in a publication?



Quality Issues - I

- Known sources of non-biological error (not exhaustive) that must be addressed
 - Technician
 - Chip lot
 - Reagent/gel lot
 - Printer tip
 - Time of printing
 - Date
 - Fluidics well/ Scanner/ position on scanner
 - Order of scanning
 - Location
 - Cage/ Field position
 - Far and away the largest issue is labeling
 - Software!

First Steps

- Look at your data/images do they look like they are supposed to?
- Then conduct semi-statistical analyses.

Which one is good or bad?



Distribution of the geography index (GEODEX) for individual chips

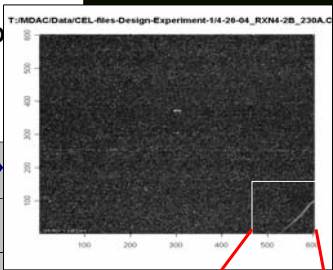
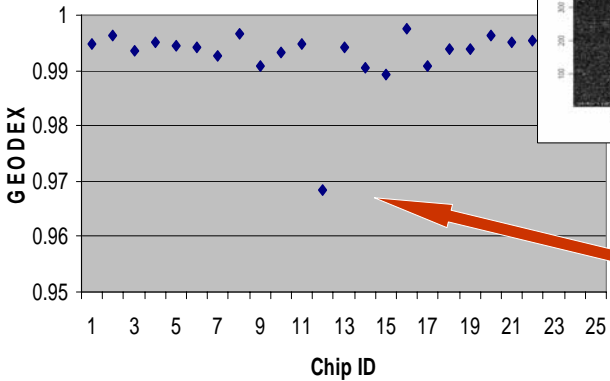


Image of chip 12

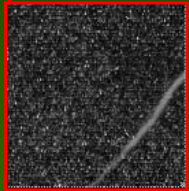
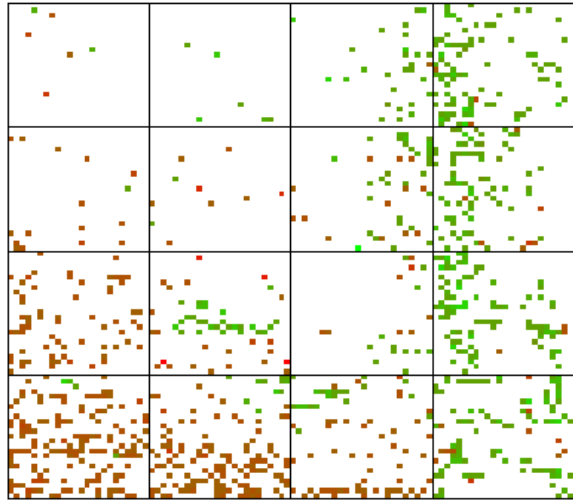
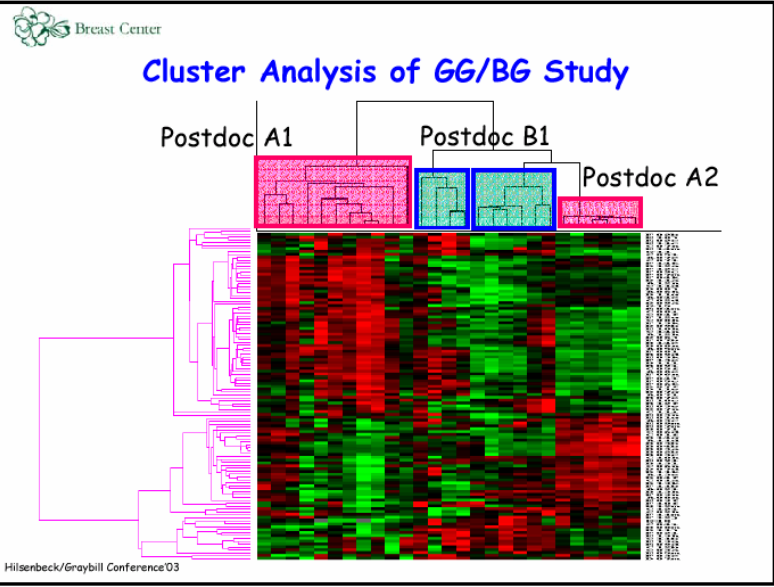


Image of the right corner on the bottom.

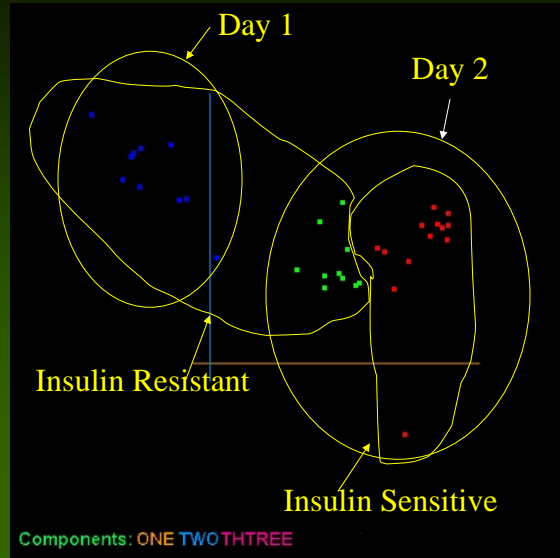


Top 2.5% of ratios red, bottom 2.5% of ratios green

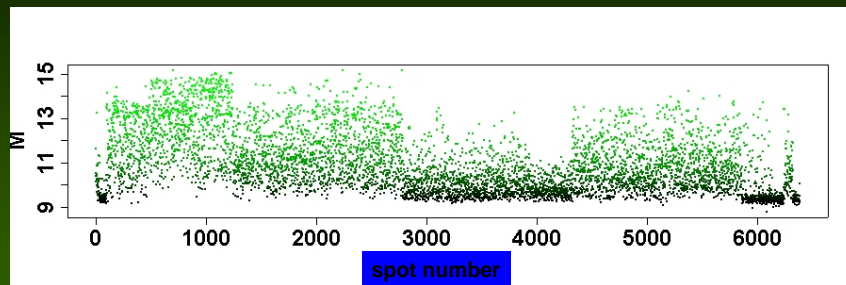


From Susan Hilsenbeck with permission

UMSA Analysis



Time of printing effects



Green channel intensities ($\log_2 G$). Printing over 4.5 days.
The previous slide depicts a slide from this print run.
From T. Speed/H Yang

Quality Issues – II

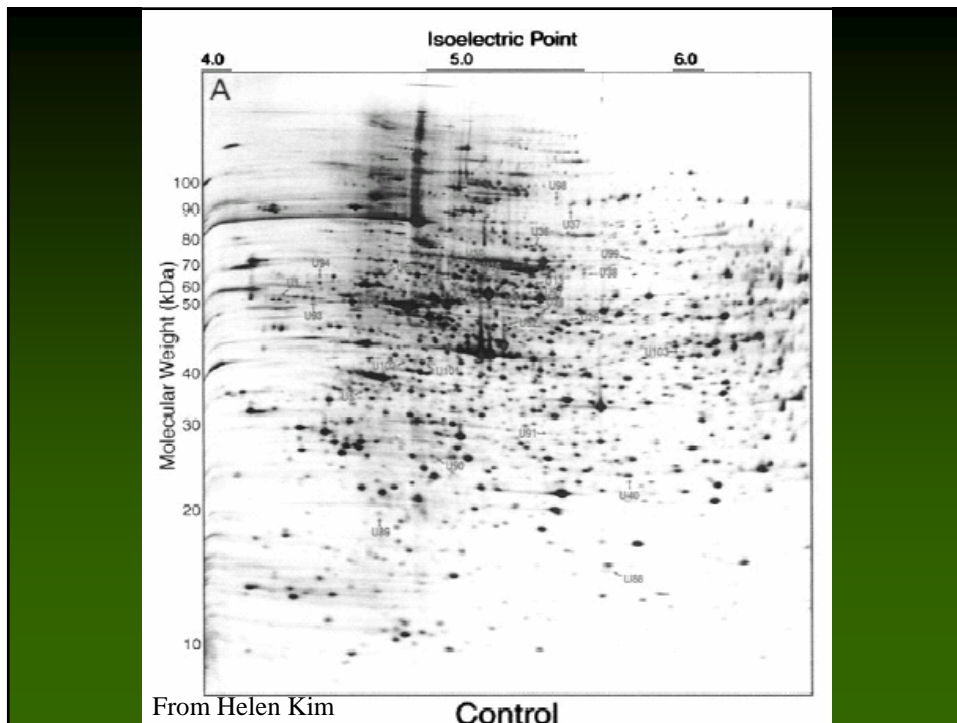
- How to address these issues
 - Make the experiment as uniform as possible
 - Agree on exactly what defines the tissue to be used, use same technician, same chip lot, same reagents (always buy a little too much), same scanner, do sample extraction, labeling and hybridization on one day if possible, establish quality control
 - Randomize when uniformity is not possible
 - Don't do all of condition 1 on day 1 and condition 2 on day 2
 - Randomize the time a chips sits waiting to be scanned
 - Randomize animal cage/plant field position
- Microarrays generate such a huge volume of data that it is possible to detect these issues, I suspect that northern, Southern, RT-PCR, westerns, and more have similar problems.

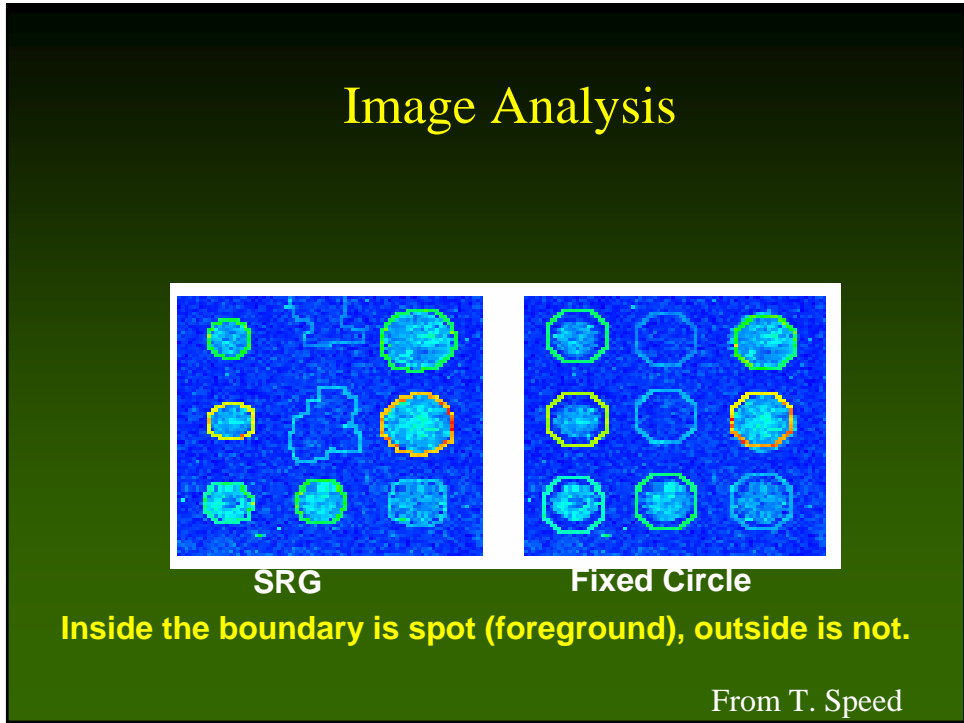
Elements of Statistics

- Power – the probability of detecting something if it is there. Usually a function of sample size and size of difference to be detected
- Image Analysis
- Quality Control- normalization/transformation
- Normalization
- Statistical Analysis
 - Class discrimination
 - Class prediction
 - Class differentiation
- Annotation
- Bioinformatics issue

Image Analysis

- How do you go from an image to a number?





Different backgrounds estimates

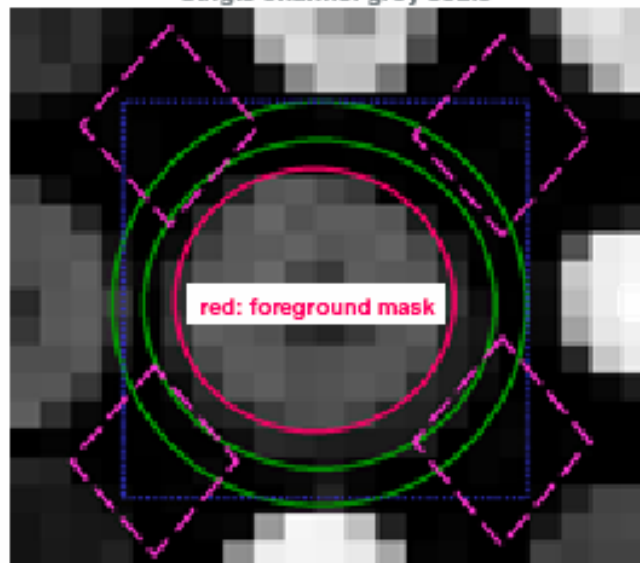
(which one is best?)

Single channel grey scale

green: QuantArray

blue: ScanAlyze

pink: Spot valley
(~GenePix)



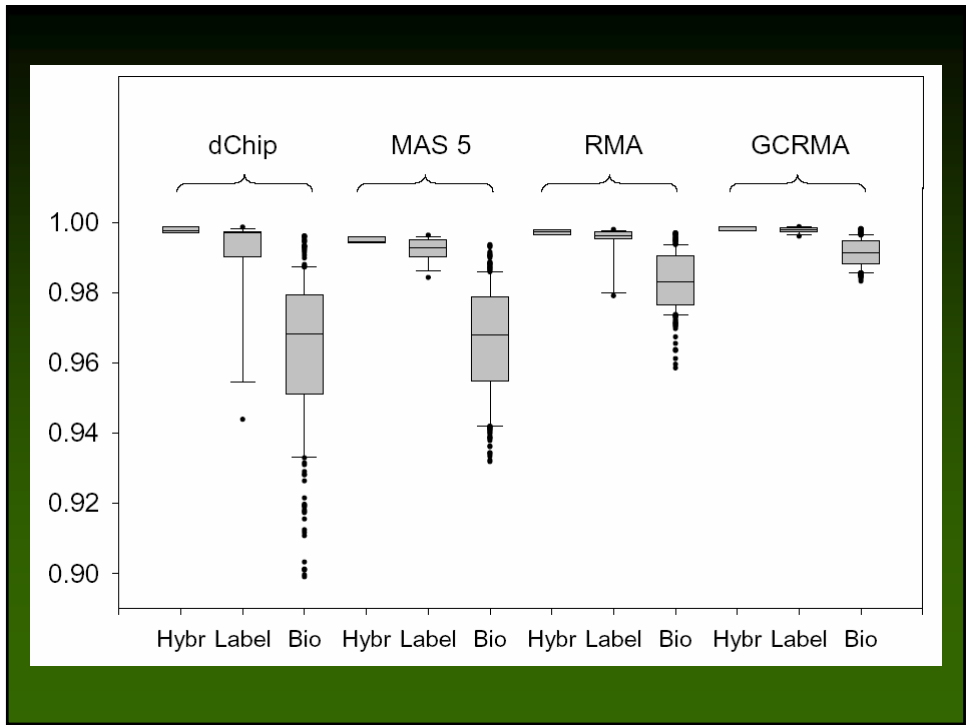
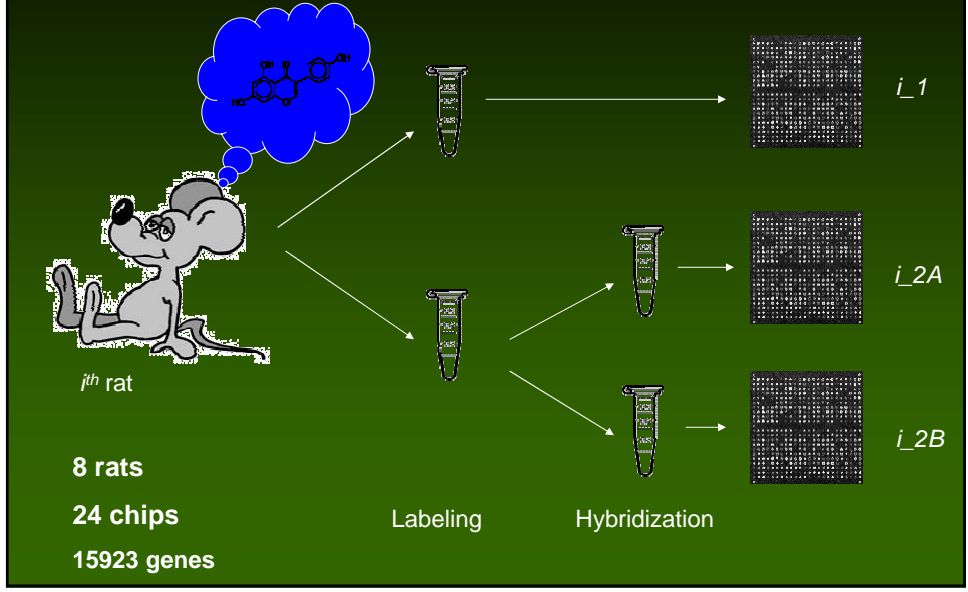
Measurement Properties of Microarrays

Zakharkin,S.O., Kim,k., Mehat,T., Chen,L., Barnes,S., Scheier,K., Parrish,R., Allison,DB.
and Page,GP. (2005) Sources of Variation in Affymetrix Microarray Experiments.
BMC Bioinformatics Aug 29;6:214.

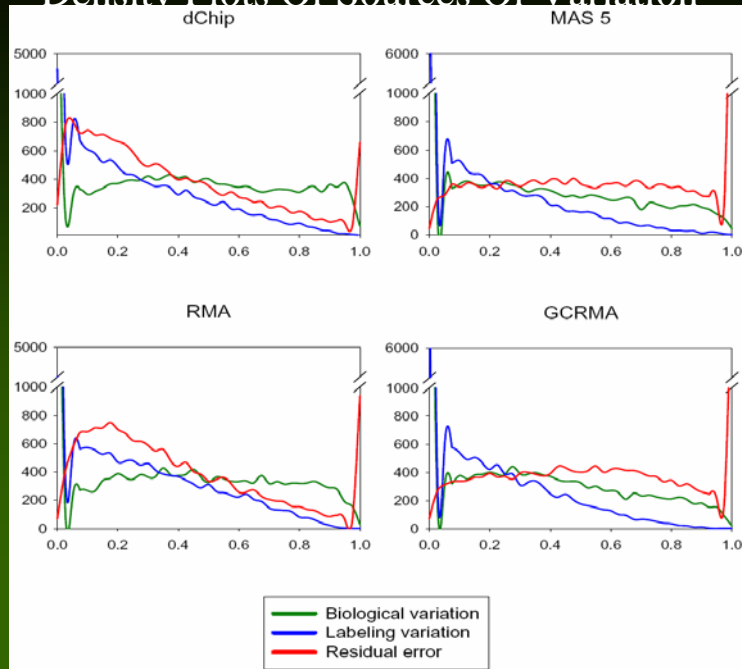
Highly accessed

According to Biomed Central over 2000 accessions in first 6 weeks

Experimental design



Density Plots Of Sources Of Variation

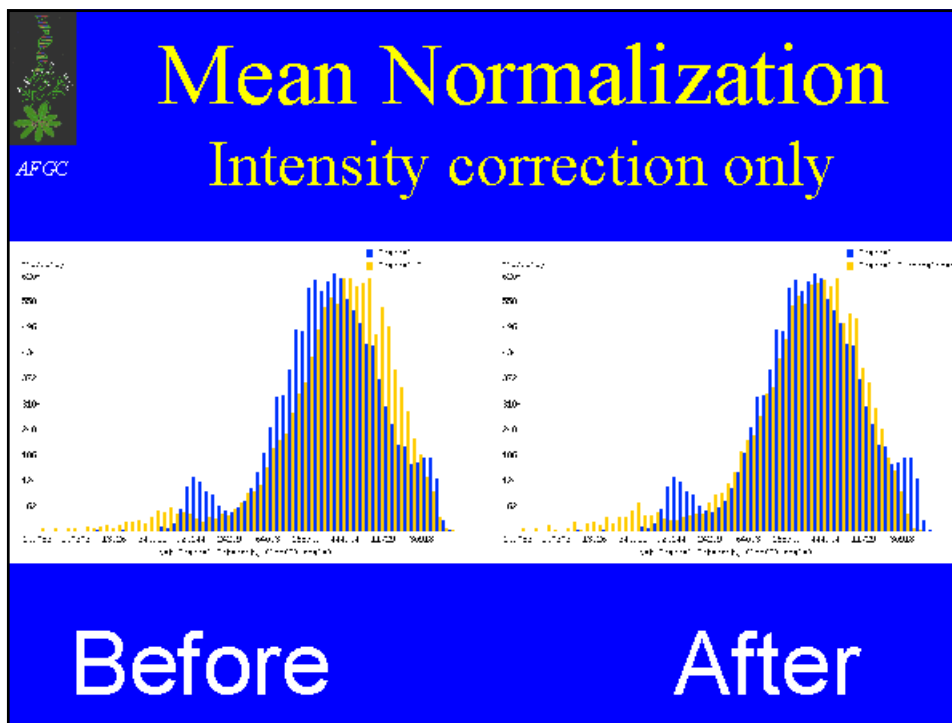


Quality Control/Normalization

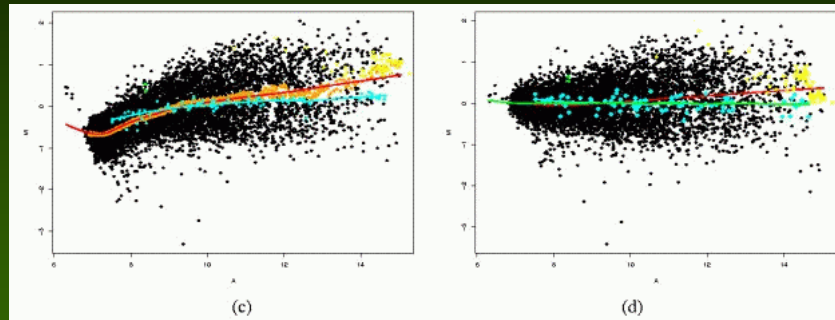
- Not all gels, chips, sequencing runs, etc are perfect
- Some are so bad they should be dropped
- Other can be fixed
 - Identify problem values/ areas
 - Fix them – adjustments and normalization

Types of Normalization

- Changes in mean
 - Add or multiple every value on a chip to take to some consistent value across chips.
 - Log transform
 - Quantile
- Changes in position
 - Lowess
 - Variance adjustments



Composite normalization



Before and after composite normalization

From T. Speed

-MSP lowess curve
-Global lowess curve
-Composite lowess curve
(Other colours control spots)

Statistical Analysis

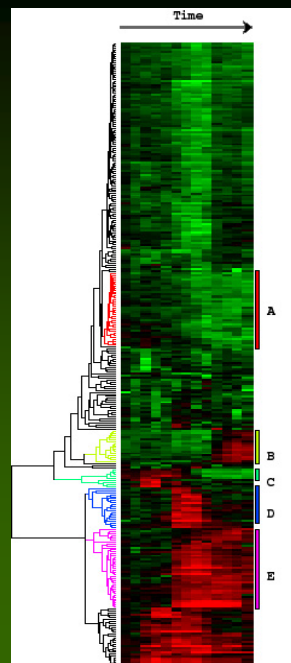
- Statistical Analysis
 - Class discrimination
 - Class prediction
 - Class differentiation

Class Discovery

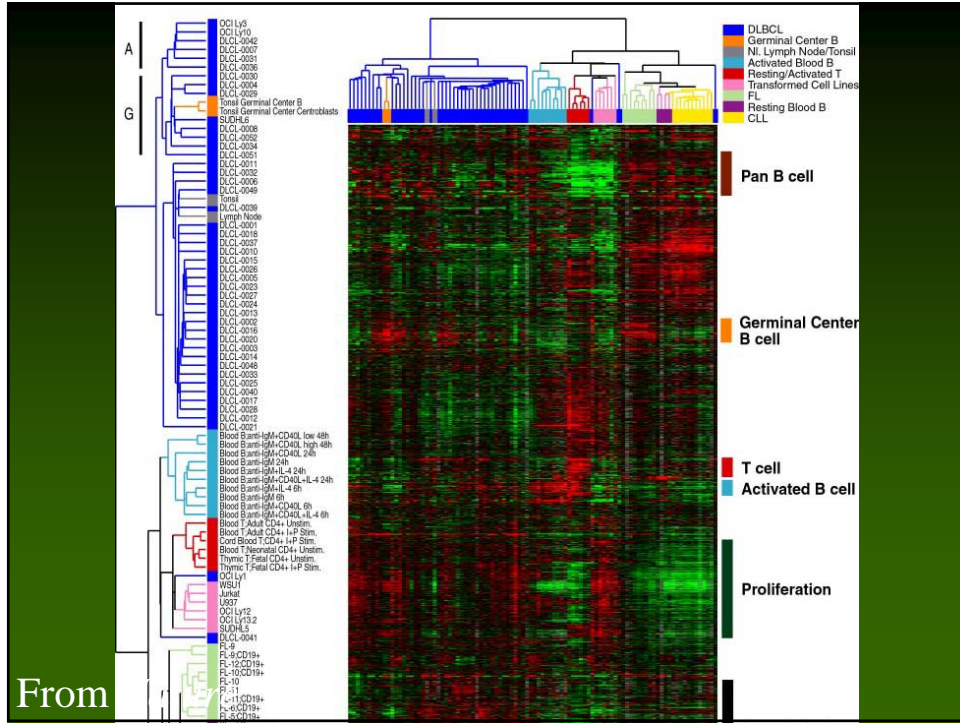
- Data visualization
- Cluster analysis
 - Clustering
 - Self organizing maps
- Multidimensional scaling
- Similarity searching

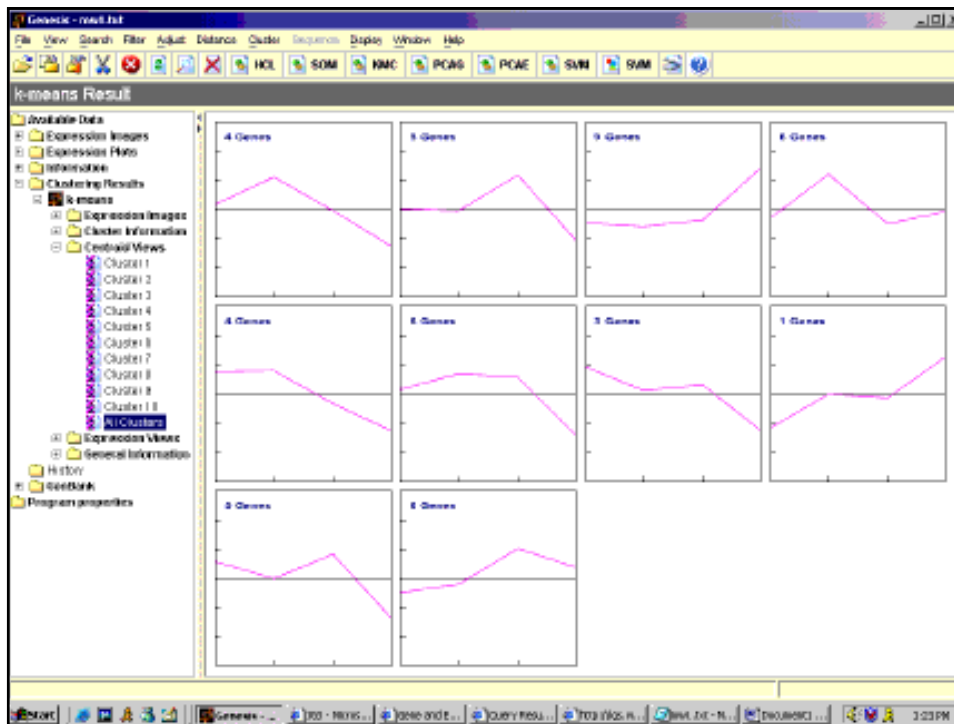
Clustering

- There are a large number of clustering algorithms.
 - Hierarchical
 - Non-hierarchical
 - Different weights
 - All will give different answers.
 - None are statistical tests

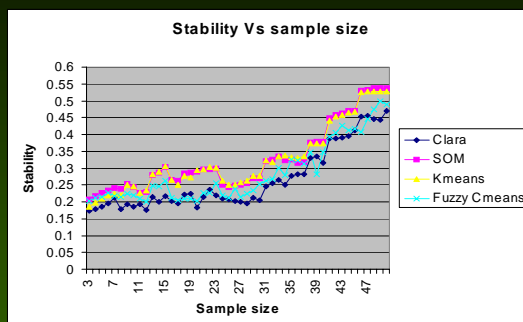


From *Nature*





Stability Results



Findings:

- Low stability (~55%) achieved for all four clustering algorithms even at the elevated sample sizes of n=50.

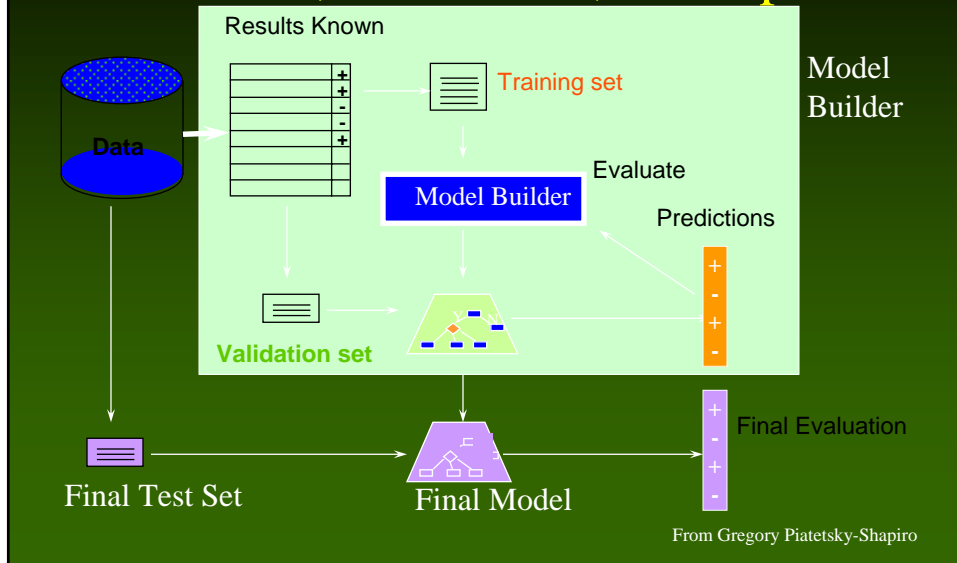
Class Prediction

- Discriminate Analysis
 - Build a predictive model for future data based upon previous data.
 - Each new sample is assigned the probability that it will fall into one the classes.
- Assign new samples to one of several groups
 - e.g. is a new tumor adenoma or squamous cell carcinoma

Class Prediction II

- Methods
 - Genes are selected with a discriminant function. Many exist, all very similar.
- Methods
 - Modeling building and test dataset are needed.
 - Often a form of cross validation such as 10 fold or leave one out are used.

Classification: Train, Validation, Test split



Cross-validation

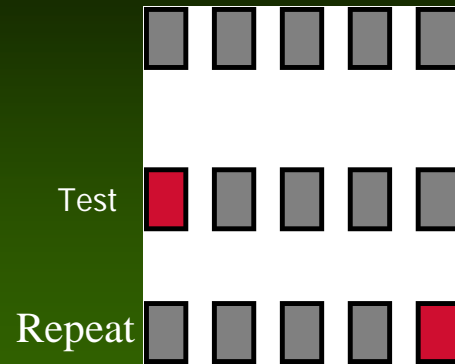
- *Cross-validation* avoids overlapping test sets
 - First step: data is split into k subsets of equal size
 - Second step: each subset in turn is used for testing and the remainder for training
- This is called *k-fold cross-validation*
- Often the subsets are stratified before the cross-validation is performed
- The error estimates are averaged to yield an overall error estimate

From Gregory Piatetsky-Shapiro

Cross-validation example:

— Break up data into groups of the same size

— Hold aside one group for testing
and use the rest to build model



From Gregory Piatetsky-Shapiro

More on cross-validation

- Standard method for evaluation: stratified ten-fold cross-validation
- Why ten? Extensive experiments have shown that this is the best choice to get an accurate estimate
- Stratification reduces the estimate's variance
- Even better: repeated stratified cross-validation
 - E.g. ten-fold cross-validation is repeated ten times and results are averaged (reduces the variance)

From Gregory Piatetsky-Shapiro

Leave-One-Out cross-validation

- Leave-One-Out:
a particular form of cross-validation:
 - Set number of folds to number of training instances
 - I.e., for n training instances, build classifier n times
- Makes best use of the data
- Involves no random subsampling
- Very computationally expensive
 - (exception: NN)

From Gregory Piatetsky-Shapiro

Class Differentiation

- Supervised Analysis
- What genes are most different between two or more groups


What is 'Significant'

- There is no such thing as significant.
- P-values are a continuum of evidence
- RA Fisher thought a $p < 0.05$ merited further study.
- Has been abused toward significant.


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 HR, Andersen B, Brandenhoff P, Guttler F (1987):
What do doctors know about statistics? *Statistics in Medicine* 6:3-10



Inference
Requires
Knowledge
of Variation



“There are other experiments, however, which cannot easily be repeated very often; in such cases it is sometimes necessary to judge the certainty of the results from a very small sample, which itself affords the only indication of the variability.”

-- Student (1908)

Types of Statistical Tests and Approaches

Type of Dependent Data	One Sample (focus usually on estimation)	Type of Independent Data					
		Categorical				Continuous	
		Two Samples		Multiple Samples		Single	Multiple
		Independent	Matched	Independent	Repeated Measures		
Categorical (dichotomous)	1 Estimate proportion (and confidence limits)	2 Chi-Square Test	3 McNemar Test	4 Chi Square Test	5 Generalized Estimating Equations (GEE)	6 Logistic Regression	7 Logistic Regression
Continuous	8 Estimate mean (and confidence limit)	9 Independent t-test	10 Paired t-test	11 Analysis of Variance	12 Multivariate Analysis of Variance	13 Simple linear regression & correlation coefficient	14 Multiple Regression
Right Censored (survival)	15 Kaplan Meier Survival	16 Kaplan Meier Survival for both curves, with tests of difference by Wilcoxon or log-rank test	17 Very unusual	18 Kaplan-Meier Survival for each group, with tests by Wilcoxon or Generalized Log Rank	19 Very unusual	20 Proportional Hazards analysis	21 Proportional Hazards analysis

After G. Howard

What should I use for 2-group testing?

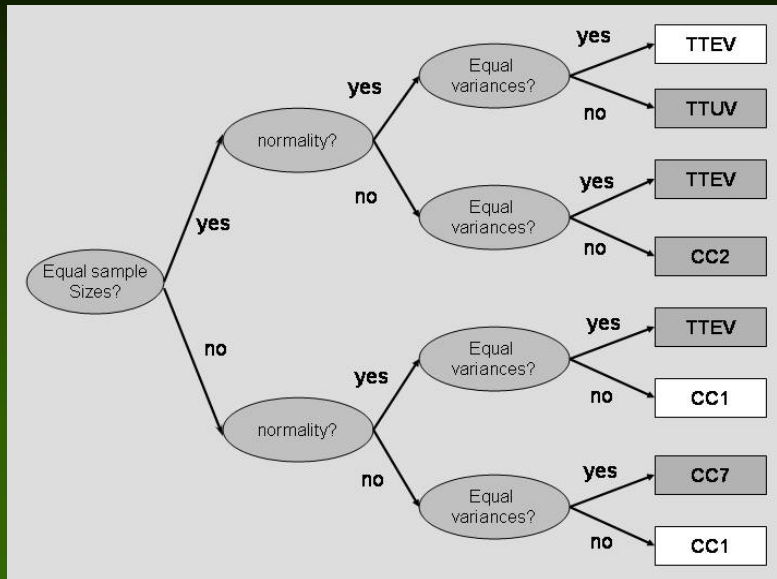
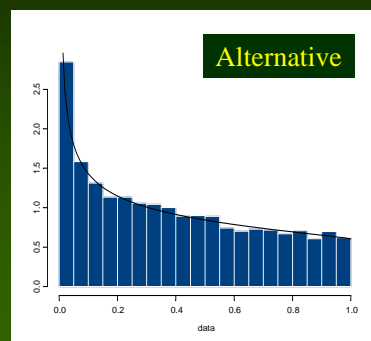
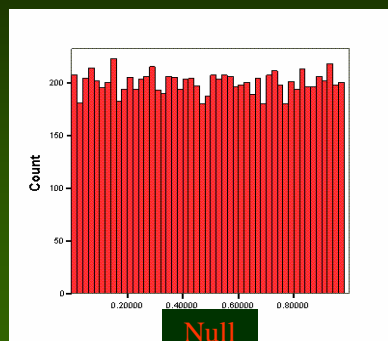


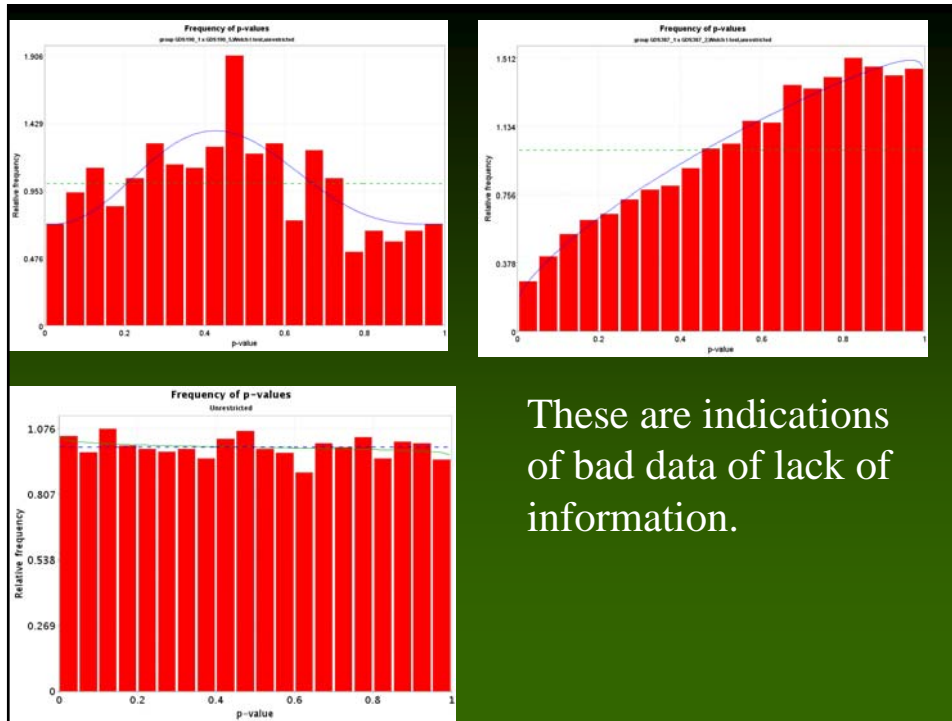
Figure 3. Mixture Model Approach from Allison et al. (2002).

Similar to Story et al (2002) and Pounds (2003)

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.



These are indications of bad data of lack of information.

Adjustments for multiple testing

- All HDB studies involve many tests
- Given the definition of a p-value an adjustment is needed when many tests are conducted.

TESTING DEFINED

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

c = type 1 error (alpha) – false positive
b = type 2 error (beta) – false negative

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

FDR - False Discovery Rate

- When many hypotheses are tested the sample size required for a Bonferroni corrected $p < 0.05$ were prohibitive in most contexts.
- Some attempts were made for intermediate adjustments
 - Lander and Botstein (1989) for linkage data
- Benjamini and Hochberg 1995 pulled together several streams of research on adjusting for multiple testing.
 - Developed method for setting an adjusted p-value that controlled for type I error
 - Like many statistical methods it has been ‘extended’ and abuse to a FDR estimating procedure
- Methods were developed for epidemiology and genetic studies, but were adapted for HDB studies

Family Wise Error Rate vs. False Discovery Rate

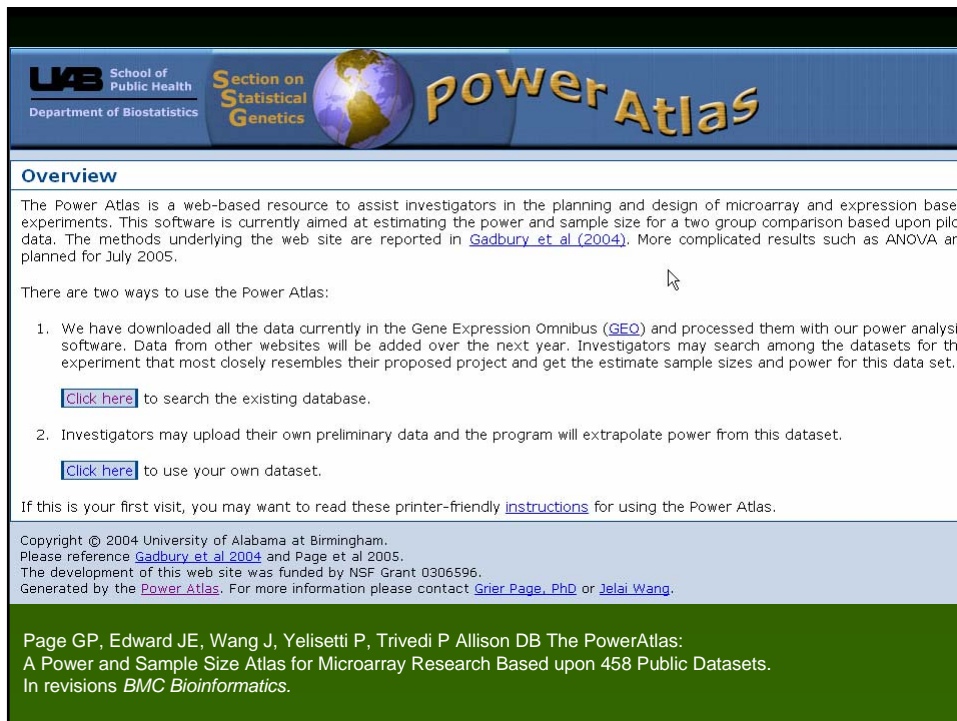
- Traditional FWER
 - Bonferroni $\alpha^* = \alpha/n$
 - Sidak $(1-(1-\alpha)^n)$
 - Very conservative
 - Minimize False discovery rates
 - Assume independence
- False Discovery Rate
 - Designed to estimate the rate of error

Power and Sample Size

- This is where microarray experiments get the most criticism.
- Experiments performed without replication
- Impression that arrays much more expensive than they are now
- Belief that microarrays are not liable to the same experimental error that experiments are
- There also has not been a good way to calculate sample size

Power

- All power and sample size calculations require an estimate of population variability
- For microarrays we use a pilot project
- Based upon the posterior probability that a gene is differentially expressed its test statistic may be increased as a function of proposed increase in sample size



The screenshot shows the 'Power Atlas' website. At the top, there is a header with the University of Alabama at Birmingham (UAB) School of Public Health, Department of Biostatistics, and Section on Statistical Genetics logos. A globe icon is positioned next to the 'powerAtlas' title. Below the header is an 'Overview' section. The text describes the Power Atlas as a web-based resource for planning and designing microarray and expression-based experiments. It mentions that the software is currently aimed at estimating power and sample size for two-group comparisons based on pilot data. Two methods are listed: 1. Searching an existing database (with a 'Click here' link) and 2. Uploading preliminary data (with another 'Click here' link). A footer section contains copyright information (© 2004 University of Alabama at Birmingham) and references to Gadbury et al. (2004) and Page et al. (2005). The footer also mentions funding by NSF Grant 0306596 and provides contact information for Grier Page, PhD and Jelai Wang.

UAB School of Public Health
Department of Biostatistics

Section on Statistical Genetics

powerAtlas

Overview

The Power Atlas is a web-based resource to assist investigators in the planning and design of microarray and expression based experiments. This software is currently aimed at estimating the power and sample size for a two group comparison based upon pilot data. The methods underlying the web site are reported in [Gadbury et al \(2004\)](#). More complicated results such as ANOVA are planned for July 2005.

There are two ways to use the Power Atlas:

1. We have downloaded all the data currently in the Gene Expression Omnibus ([GEO](#)) and processed them with our power analysis software. Data from other websites will be added over the next year. Investigators may search among the datasets for the experiment that most closely resembles their proposed project and get the estimate sample sizes and power for this data set.
[Click here](#) to search the existing database.
2. Investigators may upload their own preliminary data and the program will extrapolate power from this dataset.
[Click here](#) to use your own dataset.

If this is your first visit, you may want to read these printer-friendly [instructions](#) for using the Power Atlas.

Copyright © 2004 University of Alabama at Birmingham.
Please reference [Gadbury et al 2004](#) and Page et al 2005.
The development of this web site was funded by NSF Grant 0306596.
Generated by the [Power Atlas](#). For more information please contact [Grier Page, PhD](#) or [Jelai Wang](#).

Page GP, Edward JE, Wang J, Yeliseti P, Trivedi P Allison DB The PowerAtlas:
A Power and Sample Size Atlas for Microarray Research Based upon 458 Public Datasets.
In revisions *BMC Bioinformatics*.

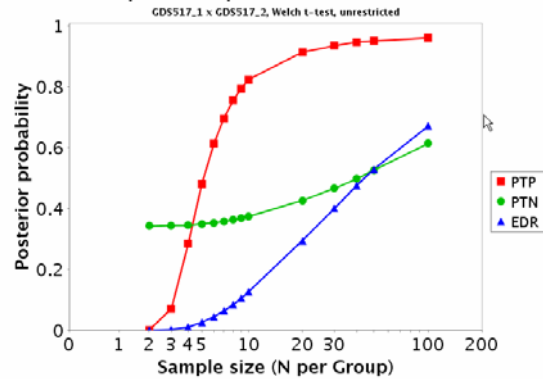
Summary Report for GDS517

Comparison of gene expression in vegetative rosettes of wild type and yda-2 (YODA MAPKK kinase) plants. YODA promotes extra-embryonic cell fates in basal lineage. Results suggest YODA is involved with a novel developmental signal transduction pathway.

Data from four biological replications representing tissue grown and harvested at different times was collected. The average signals of all eight experiments were scaled to the target value of 500.

[GSI 71 Affymetrix GeneChip Arabidopsis Genome Array AtGenome1](#)

Combined posterior probabilities at threshold 0.05



The graph uses experiment GSE919 as pilot data to extrapolate out the expected discovery rates, posterior true positive (PTP) rates, and posterior true negative (PTN) rates for sample sizes of 4 biological replicates per group at a alpha level of 0.05. Graphs for alternative alpha levels are in the zip file containing the complete results.

The EDR is the proportion of genes that are truly different that will be called significant at the alpha level chosen. This is analogous to average power. The smaller the alpha level the lower the EDR and increasing the sample size increases the EDR. The PTP is the proportion of genes that are called significant that will be truly different between the conditions. In general, the PTP will increase as sample size and alpha level are increased. The PTN is the proportion of genes that are called non-significant that will truly not be different between the conditions. In general, the PTN will increase as sample size increases, but will decrease as the alpha level

Data Interpretation

- The most time consuming portion of a HDB experiment is the interpretation
- Many databases and resources exist
 - Dr. Loraine talked about these in great detail

a posteriori vs. *a Priori* data interpretation

- Many people get the data and then stare at it and tell a story based on their subjective observations about the data.
- *A posteriori* observations are highly biased
- *A priori* observations require knowledge of pathway, gene family, etc. There can be a large number of classes.

Global/Meta Analytical Tests of Pathways

Premise: We can learn something additional and/or test with more power if we consider the fact that genes may exist within ‘families.’ Several Tests –

- Fisher’s meta analytical tests – combine the individual p-values from n genes $\sim \chi^2_{(2n-2)}$
- Vote Counting methods
 - Onto-express
 - GSEA
- Normalize all the data to Z scores and compare the expression levels
- Issues even under H_0 if genes in a pathway are correlated there will be an increase in type I error
- Address FEWR vs FDR per group

Gene Family-Based Hypothesis Testing: What people say they are testing vs what they are testing.

Which Null?

1. None of the genes in family c are differentially expressed.
2. The proportion of genes in family c that are differentially expressed is equal to the proportion of genes in the remainder of the genome that are differentially expressed.
3. The correlation matrix among the expression levels of the genes in family c is an identity matrix.
4. The correlation matrix among the expression levels of the genes in family c is the same across experimental conditions.
5. The intersection of #1 and #3.

Mootha et al (2003). "We introduce an analytical strategy, Gene Set Enrichment Analysis, designed to detect modest but coordinate changes in the expression of groups of functionally related genes."

This implies that the null of interest is #1, but the test appears to be the intersection of #2 and #3.

Global/Meta Analysis

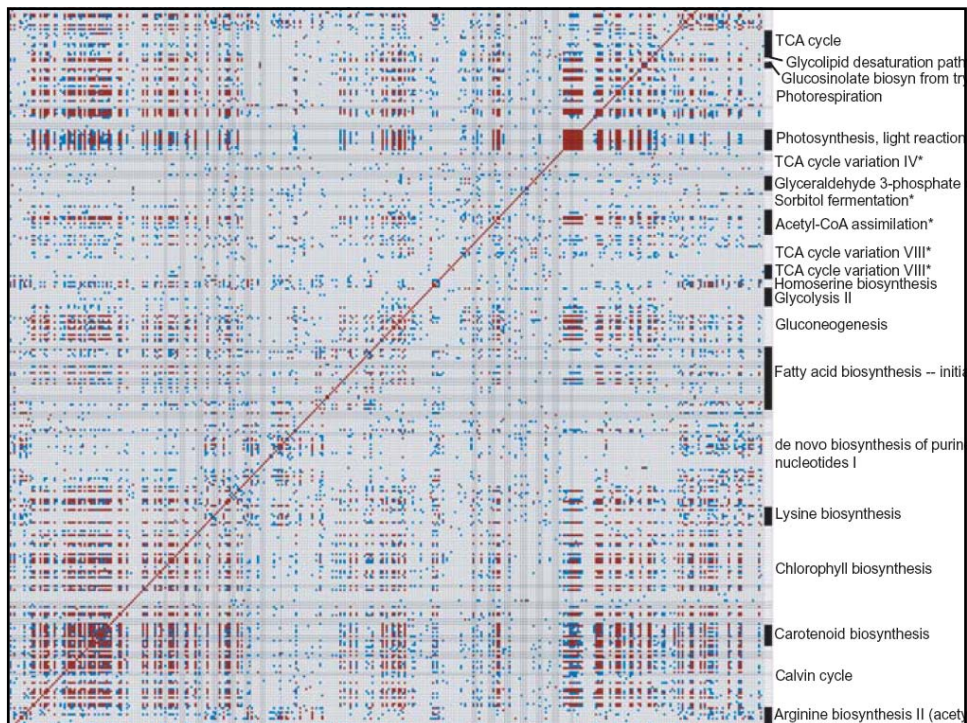
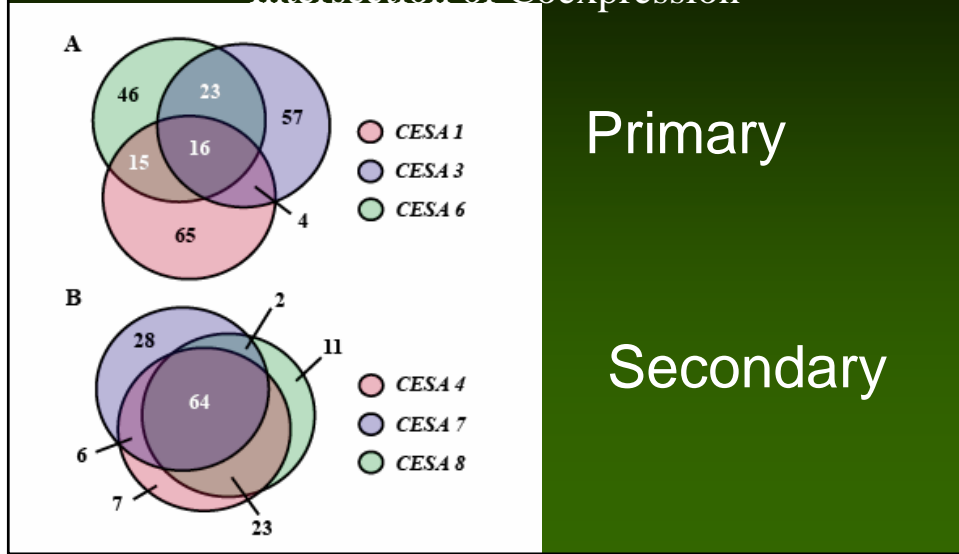
Biological Process				
<i>Function Name</i>	<i>Total</i>	<i>P-Value</i>	<i>FDR</i>	<i>Bonferroni</i>
inflammatory response	71	1.11E-16	4.72E-14	4.72E-14
immune response	95	8.44E-15	1.79E-12	3.59E-12
epidermal differentiation	38	1.65E-11	2.34E-09	7.02E-09
cell-cell signaling	100	3.14E-10	3.34E-08	1.34E-07
cell adhesion	77	5.72E-09	4.86E-07	2.43E-06
chemotaxis	43	8.73E-09	6.18E-07	3.71E-06
cellular defense response	40	1.74E-08	1.06E-06	7.39E-06
development	80	3.44E-08	1.83E-06	1.46E-05
antimicrobial humoral response	45	9.90E-08	4.68E-06	4.21E-05
response to viruses	18	7.16E-07	3.04E-05	3.04E-04
cell surface receptor linked signal transduction	54	3.29E-06	1.27E-04	1.40E-03
cell motility	47	3.55E-06	1.26E-04	1.51E-03
cell proliferation	79	1.81E-05	5.90E-04	7.67E-03
protein biosynthesis	6	1.81E-05	5.49E-04	7.69E-03
skeletal development	36	2.59E-05	7.34E-04	1.10E-02

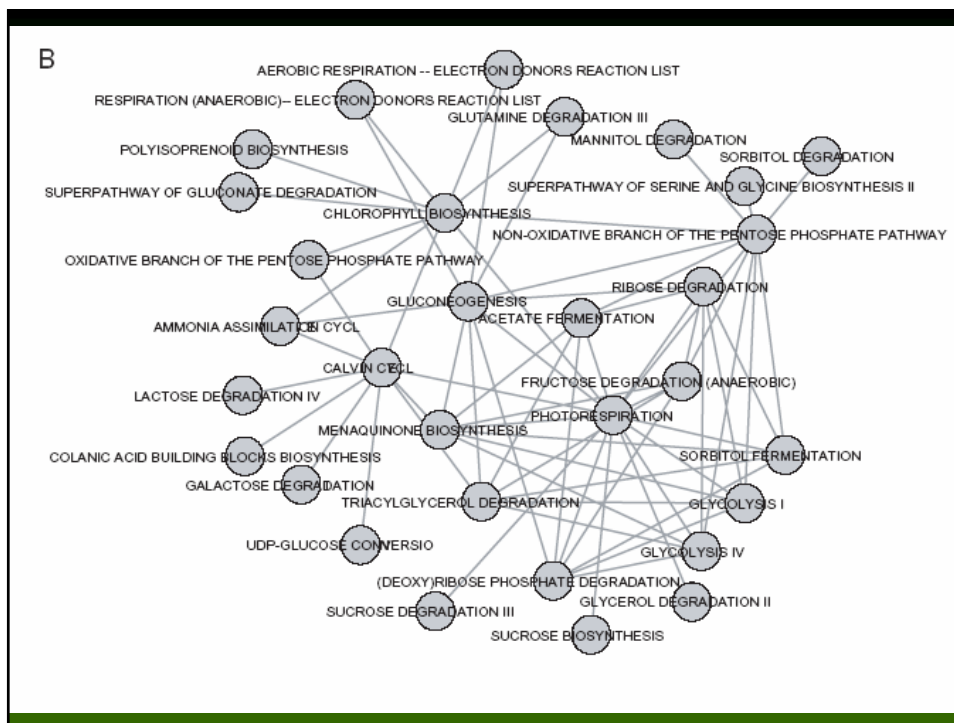
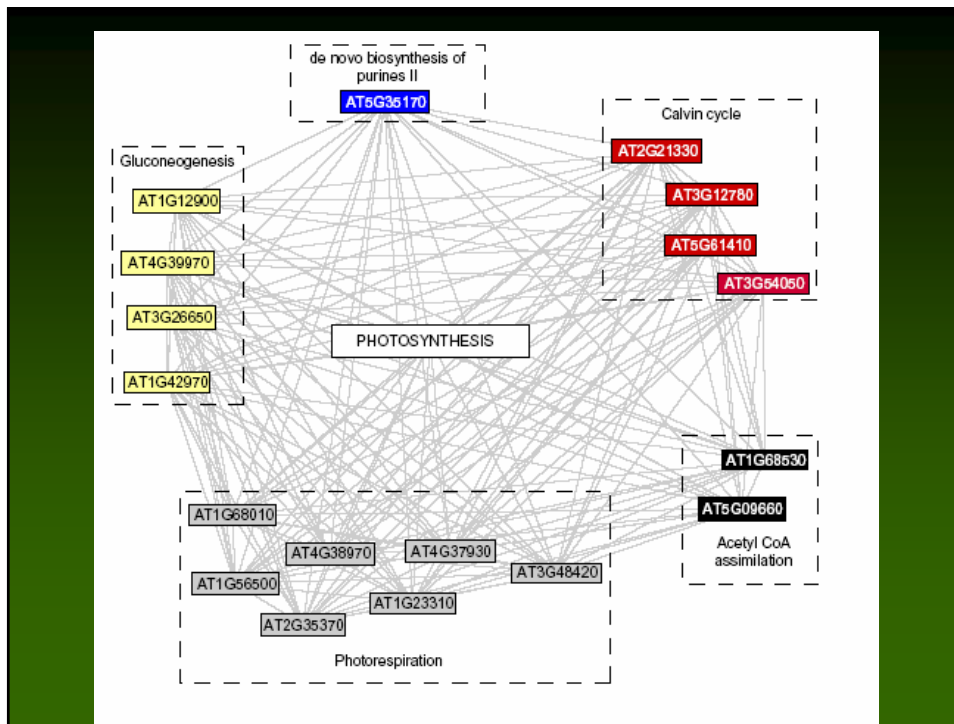
Use of FDR for Union-Intersection tests

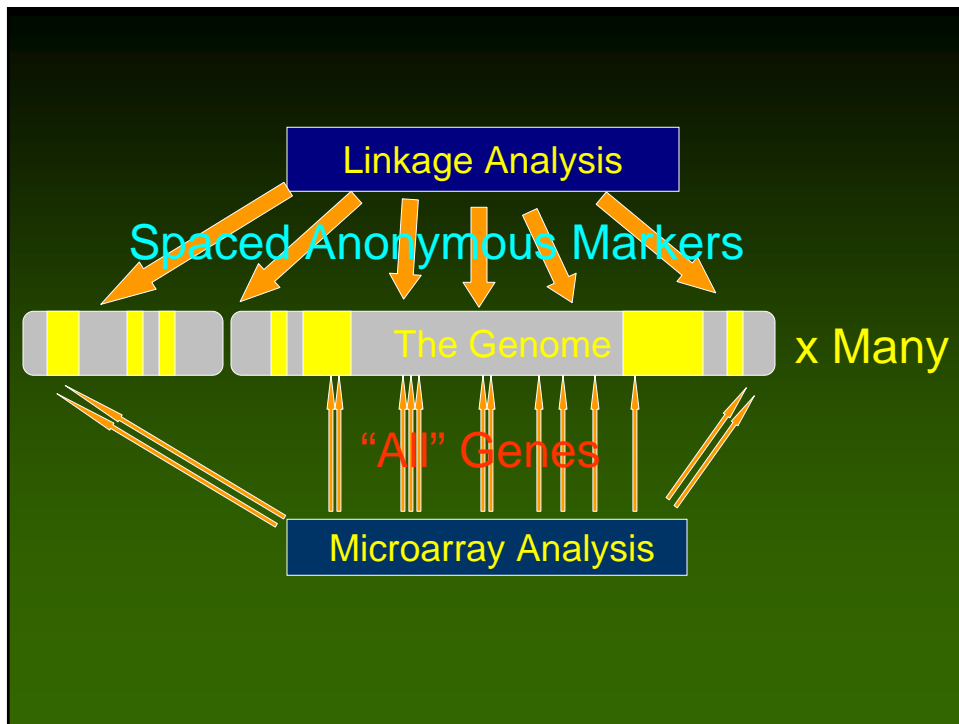
- Traditional
 - The ‘min’ test.
 - Low power
 - Not of definitive size
 - Ignores information (i.e., the p-value for min test is largest p-value for $h_0 \in H_0$ regardless of the value of any other p-values).
- Informational based approaches
 - All p-values are not equal
 - A variety of ways to weight
 - Let’s consider FDR or PTP –these are equal across datasets
 - Can conduct simple product of FDR.



Venn Diagram of 100 gene more Significantly associated with known CESA genes: ICE Intersection of Coexpression







Bioinformatics Issues

- HDB studies generate a huge amount of information.
- Storage and handling of the data can be difficult.
- Data standards are developing (MIAME for microarrays), proteomics just beginning.