THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Section ON Statistical Genetics

Statistical Analysis of proteomic data

Grier P Page Ph.D.
Assistant 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

File Edit View Favorites Tools Help

Back Search Favorites Media

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

SCHOOL OF PUBLIC HEALTH
Department of Biostatistics

Section ON Statistical Genetics

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

Home

Activities in News

Janet L. Norwood Award

Short Course on Statistical Genetics

Contact Us

RealPlayer

File View Play Favorites Tools Help

Statistical analysis of microarrays

THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Section ON Statistical Genetics

Statistical analysis of microarrays

Gpage@uab.edu
4-4930
Ryals 317D

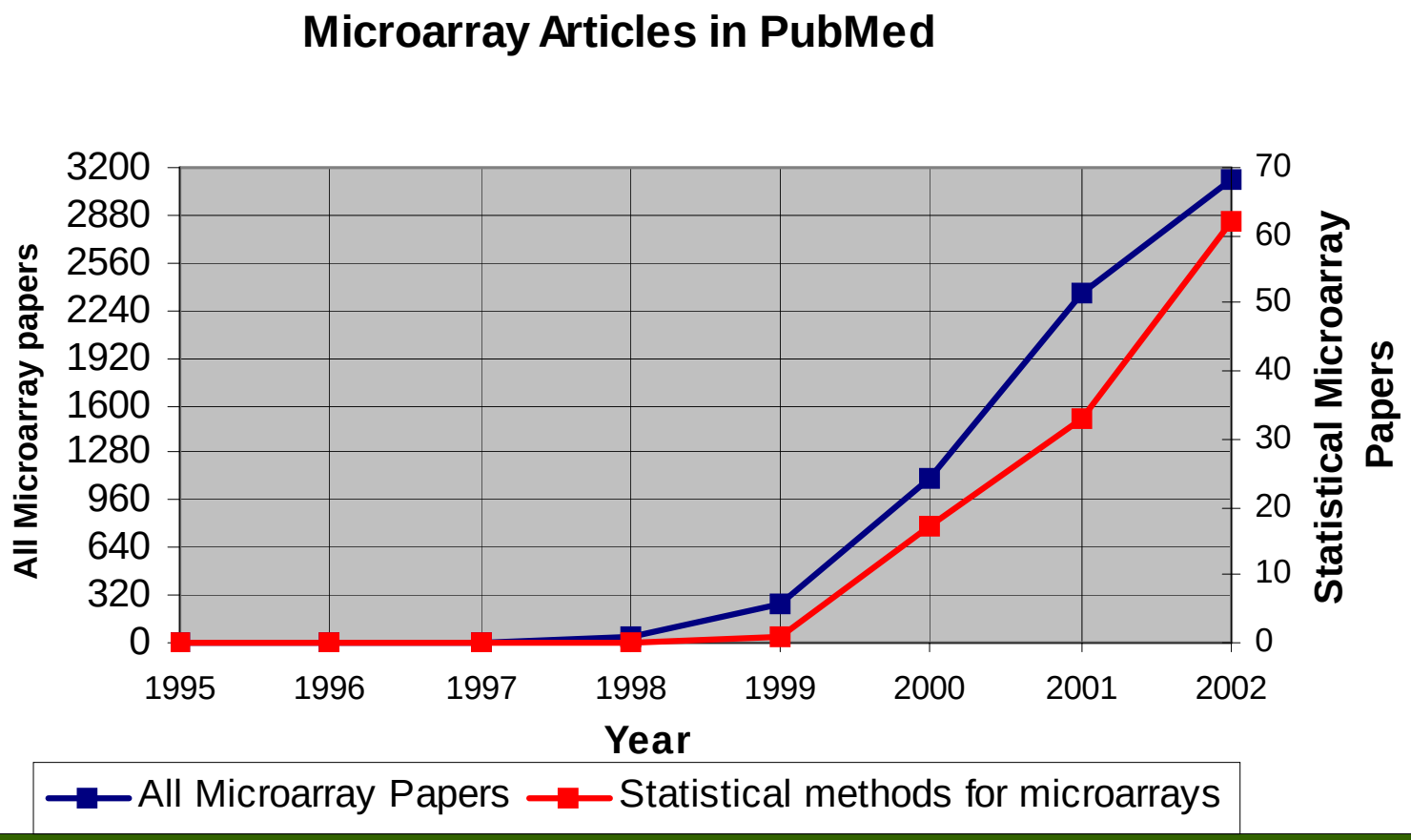
Grier P Page Ph.D.
Assistant Professor
Section on Statistical Genetics
Department of Biostatistics
School of Public Health

Now Playing rnhigh 351 Kbps 0:01 / 42:02

Real Guide Music & My Library Music Store

Done

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



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 are 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.

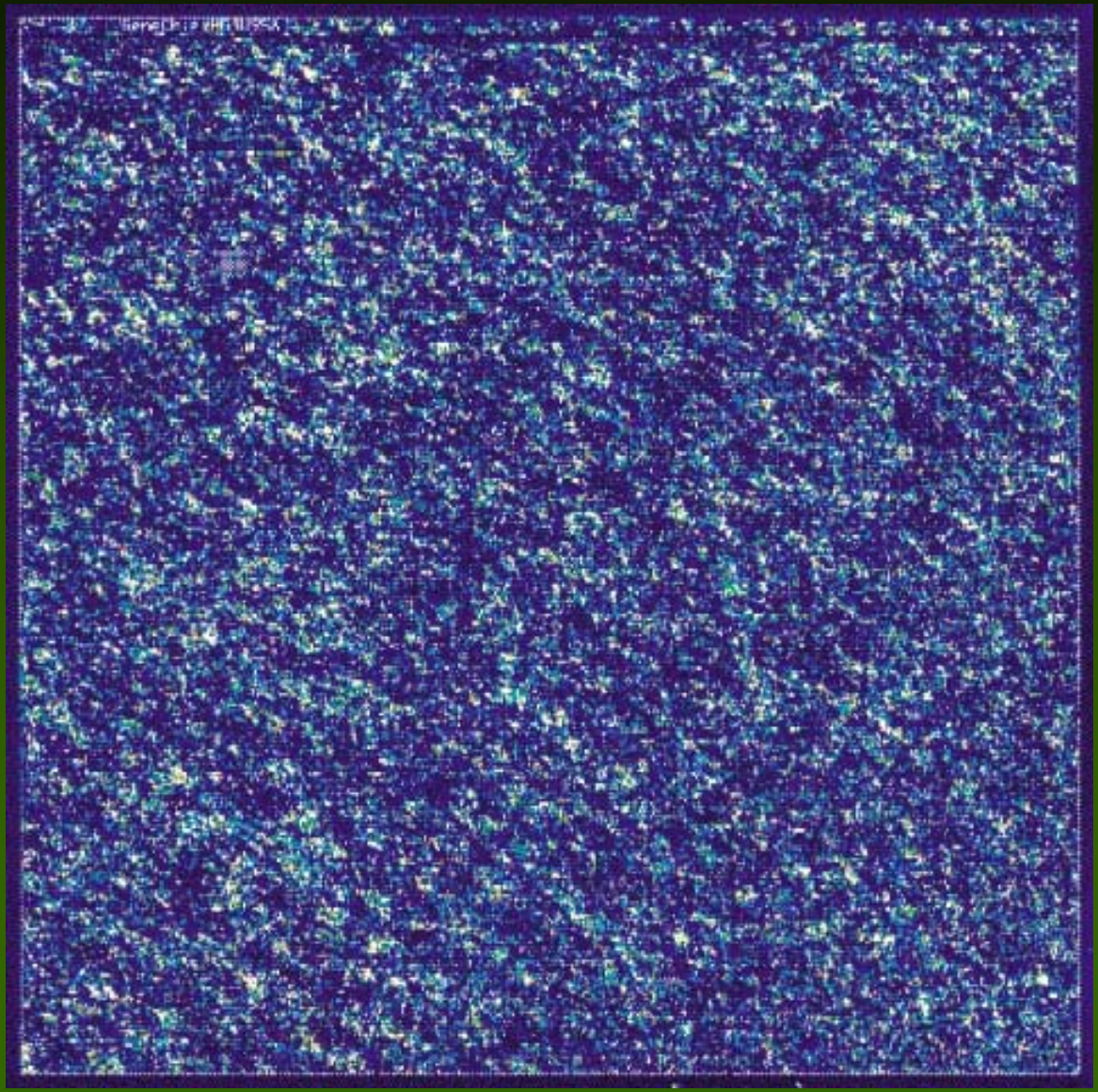
A Circular & Epistemologically Invalid Framework

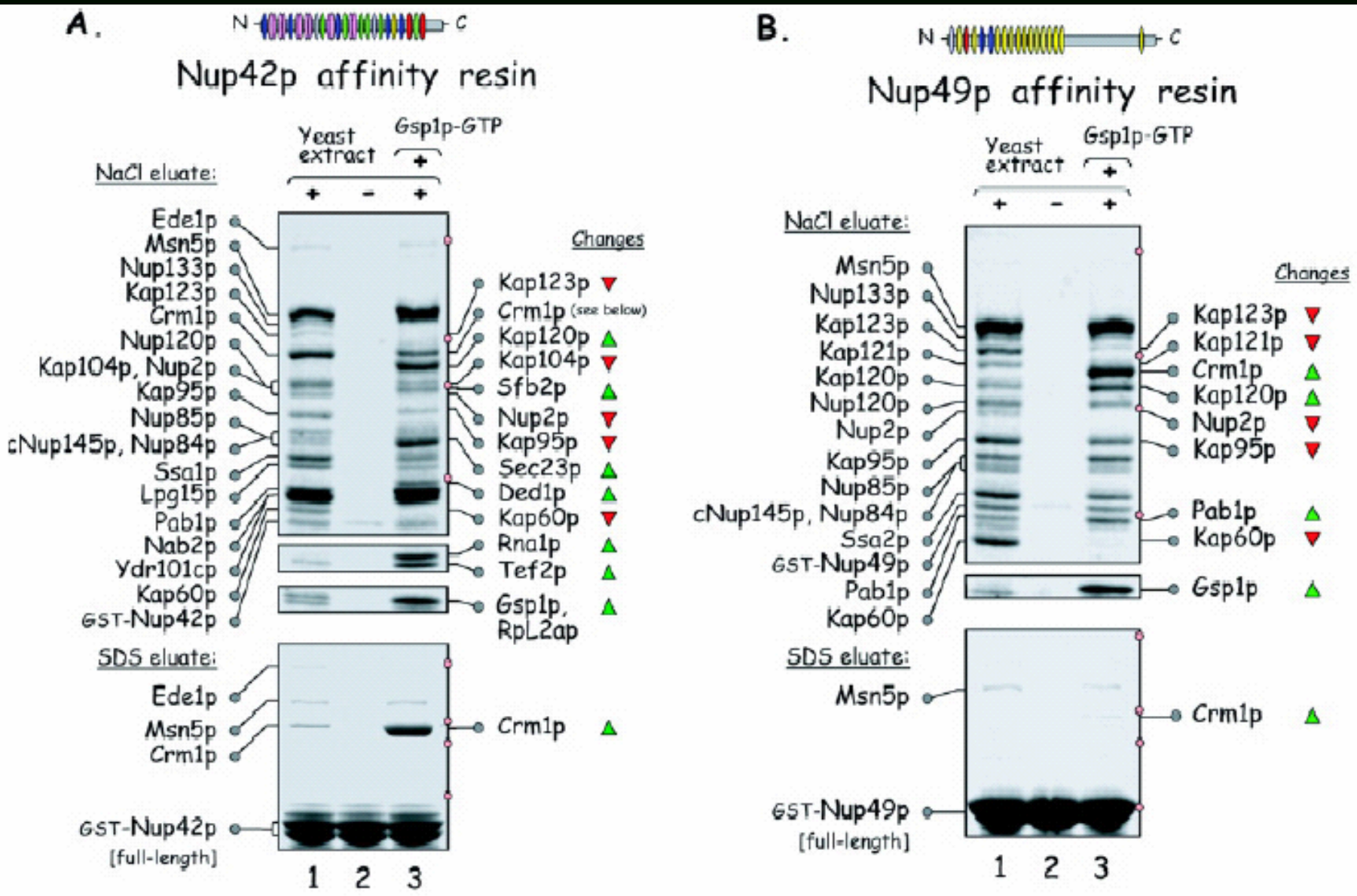
- Application to single real data sets of unknown nature.

What is High Dimensional Biology?

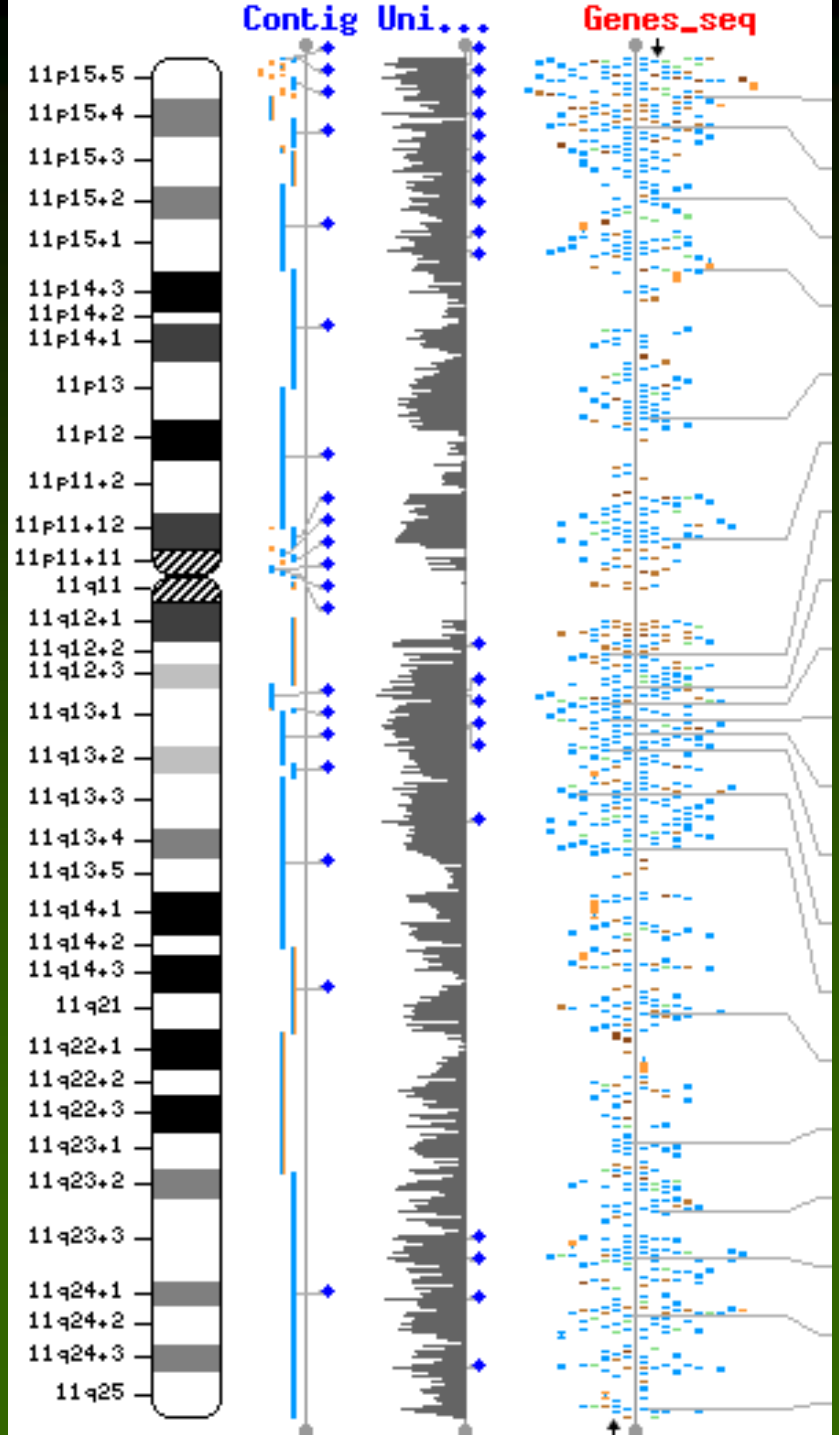
- High Dimensional Biology – is a broad topic covering biological systems where the number of variables is very large.
- Topics that often fall in HDB are microarray, proteomics, linkage, and genomics.
- HDB is also highly collaborative both ‘wet’ and ‘dry’ lab people.

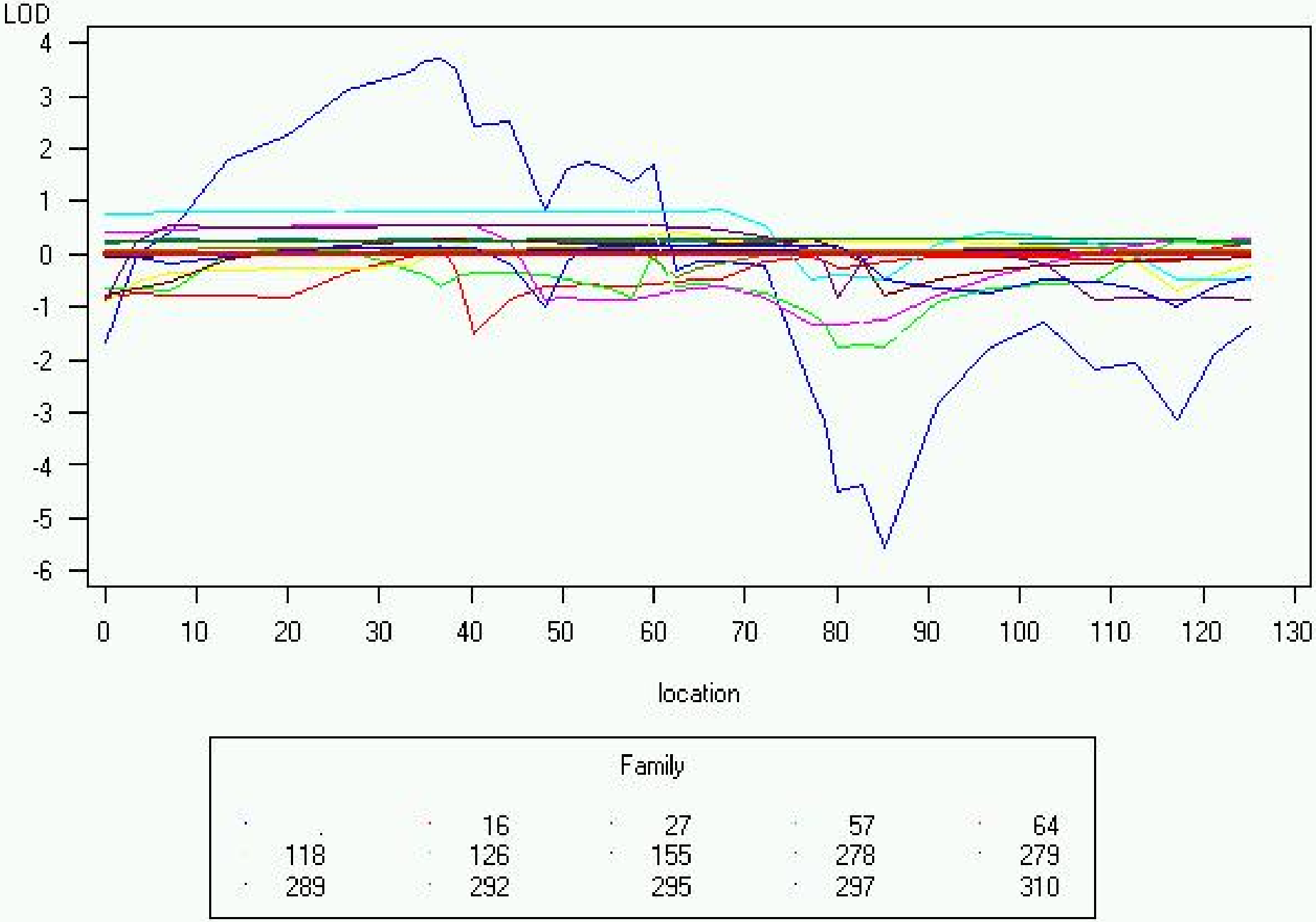
Affymetrix type array





Barnes class 02-04-03


Contig Uni...
Genes_seq
11p15.5
11p15.4
11p15.3
11p15.2
11p15.1
11p14.3
11p14.2
11p14.1
11p13
11p12
11p11.2
11p11.12
11p11.11
11q11
11q12.1
11q12.2
11q12.3
11q13.1
11q13.2
11q13.3
11q13.4
11q13.5
11q14.1
11q14.2
11q14.3
11q21
11q22.1
11q22.2
11q22.3
11q23.1
11q23.2
11q23.3
11q24.1
11q24.2
11q24.3
11q25


LOD
4
3
2
1
0
-1
-2
-3
-4
-5
-6
0
10
20
30
40
50
60
70
80
90
100
110
120
130
location
Family
.
16
27
57
64
118
126
155
278
279
289
292
295
297
310

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 away that the question that is being asked can be easily and unambiguously answered.
- The experimental hypothesis drives the design.

Statistics

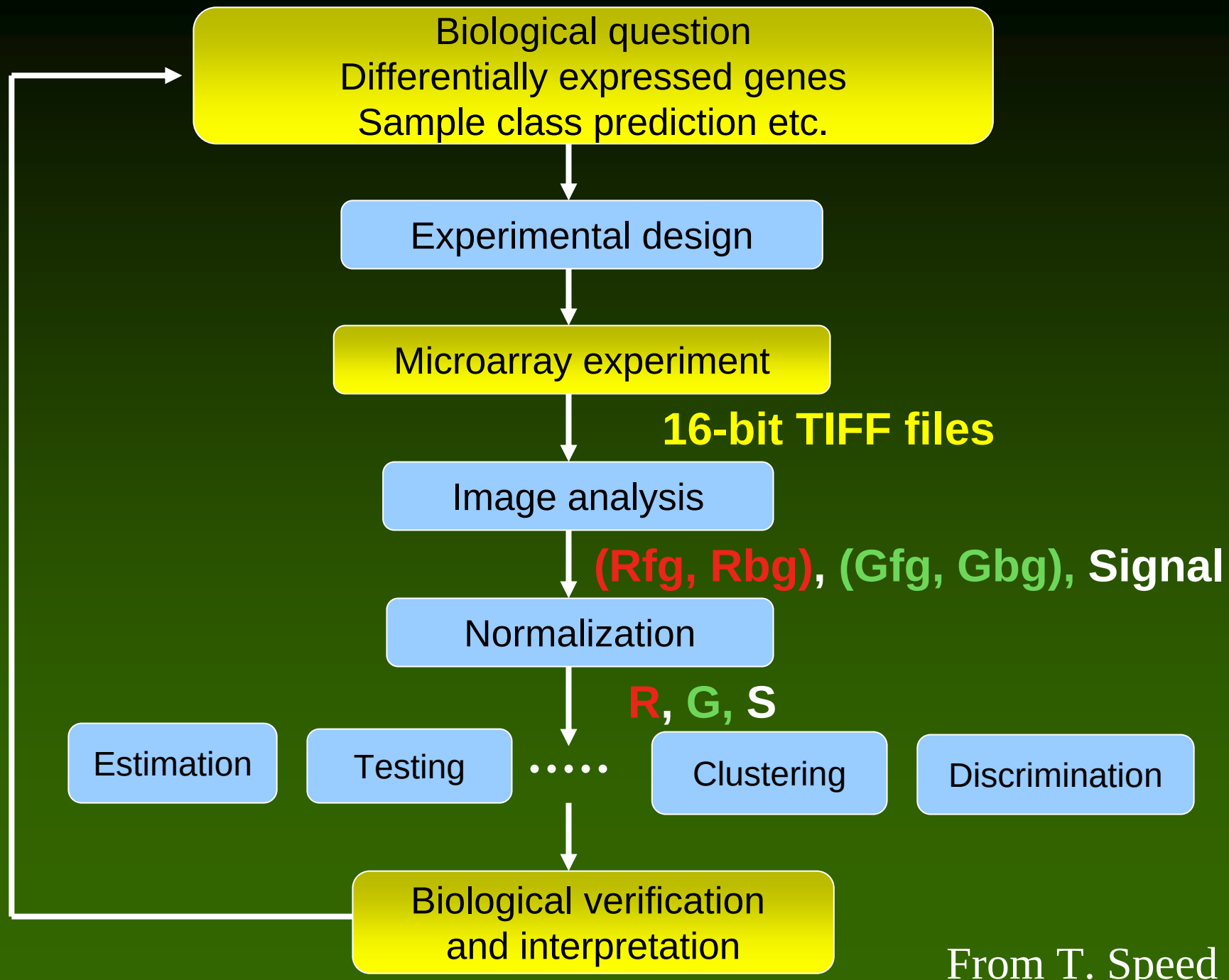
- Methods for make inferences about a population as a whole by taking a sample.

- 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?


Biological question
Differentially expressed genes
Sample class prediction etc.
Experimental design
Microarray experiment
16-bit TIFF files
Image analysis
(Rfg, Rbg), (Gfg, Gbg), Signal
Normalization
R, G, S
Estimation
Testing
.....
Clustering
Discrimination
Biological verification
and interpretation

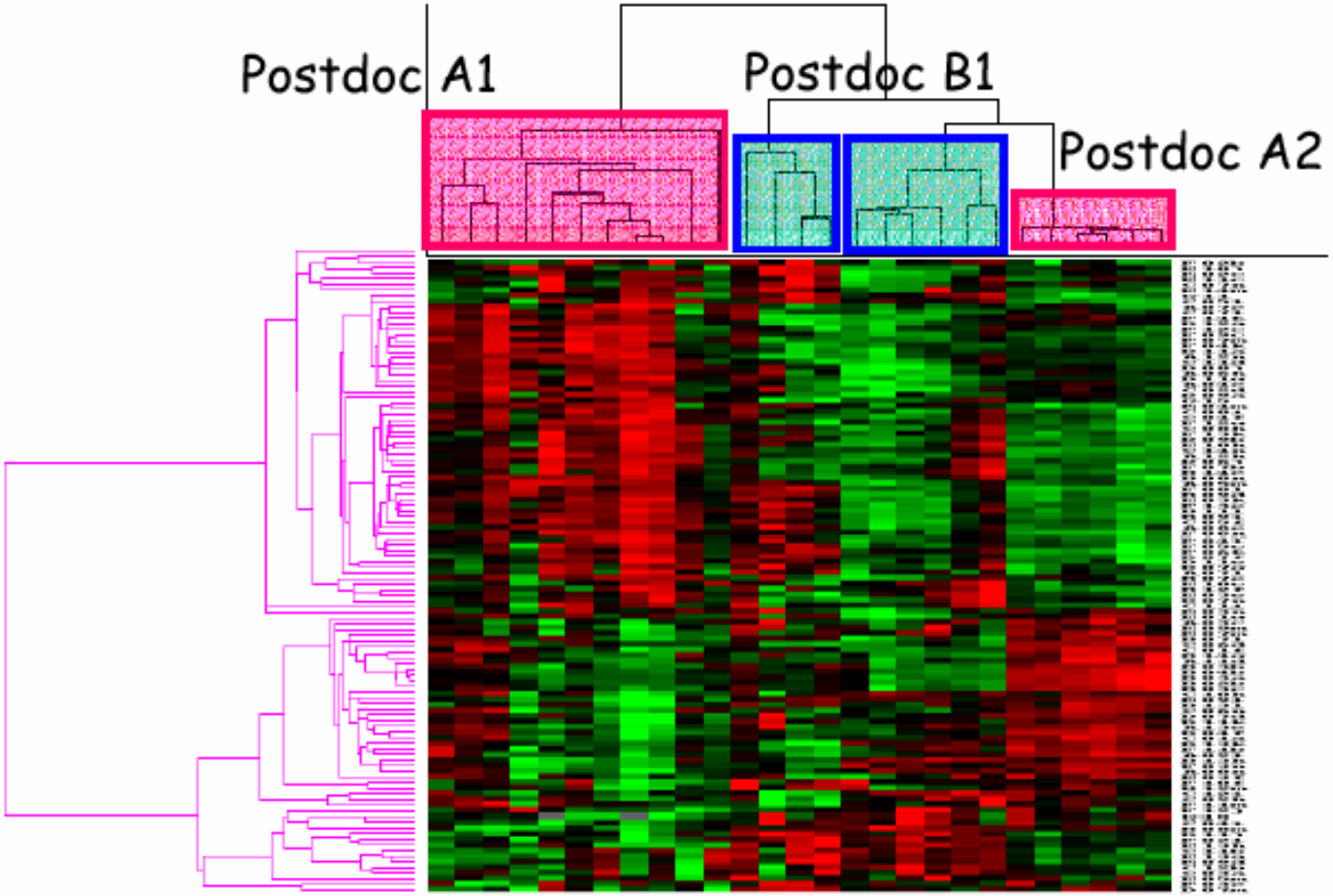
From T. Speed

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/ positionon scanner
  - Order of scanning
  - Location
  - Cage/ Field position
  - Far and away the largest issue is labeling

Breast Center

Cluster Analysis of GG/BG Study



Hilsenbeck/Graybill Conference'03

From Susan Hilsenbeck with permission

Quality Issues – II

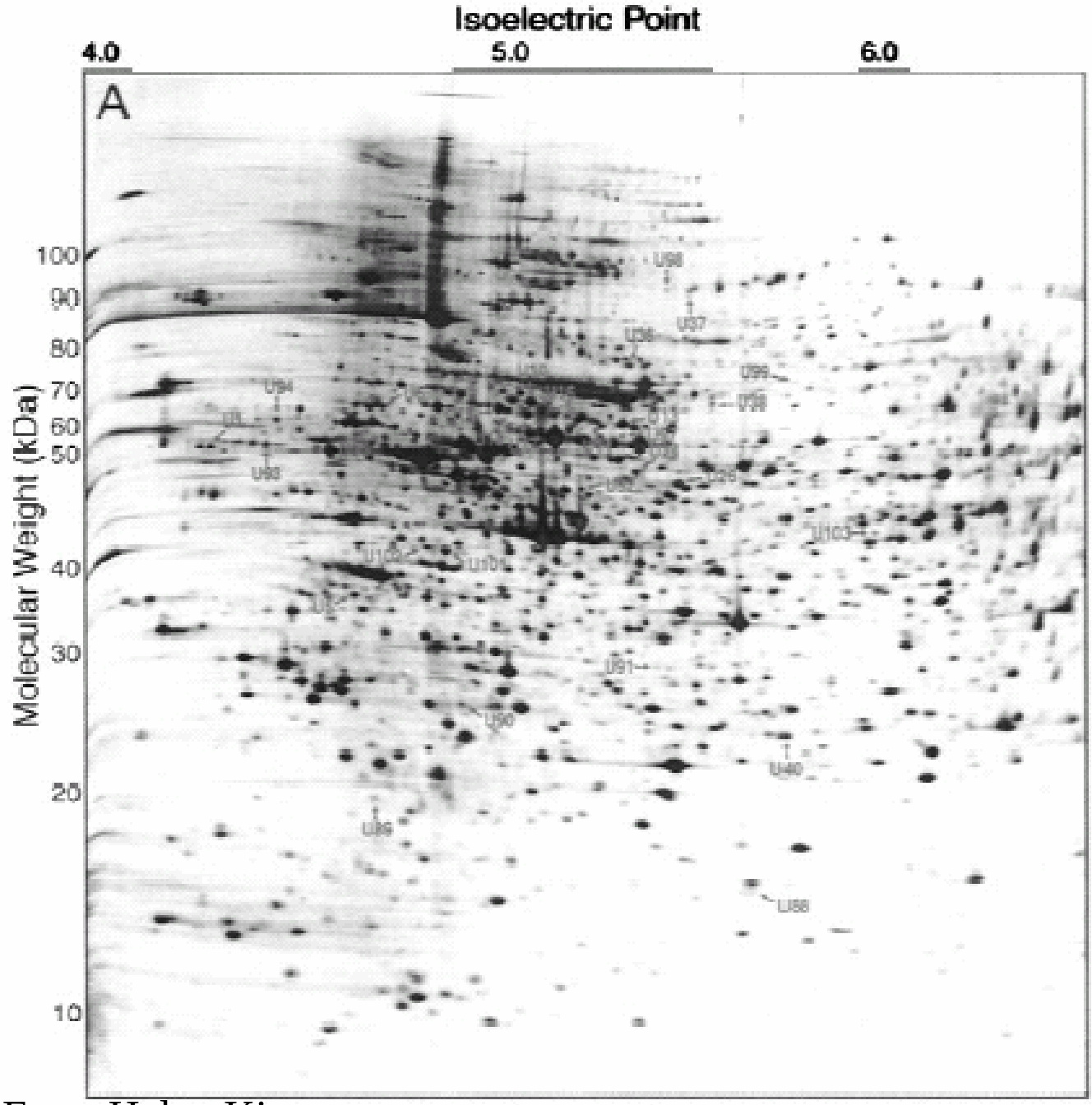
- How to address these issues
  - Make the experiment as uniform as possible
    - Agree on exactly what defines the tissue to be used, use same technician, same chip lot, same reagents (always buy a little too much), same scanner, do sample extraction, labeling and hybridization on one day if possible, 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 is is possible to detect these issues, I suspect that northerns, Southerns, 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?



From Helen Kim

Control

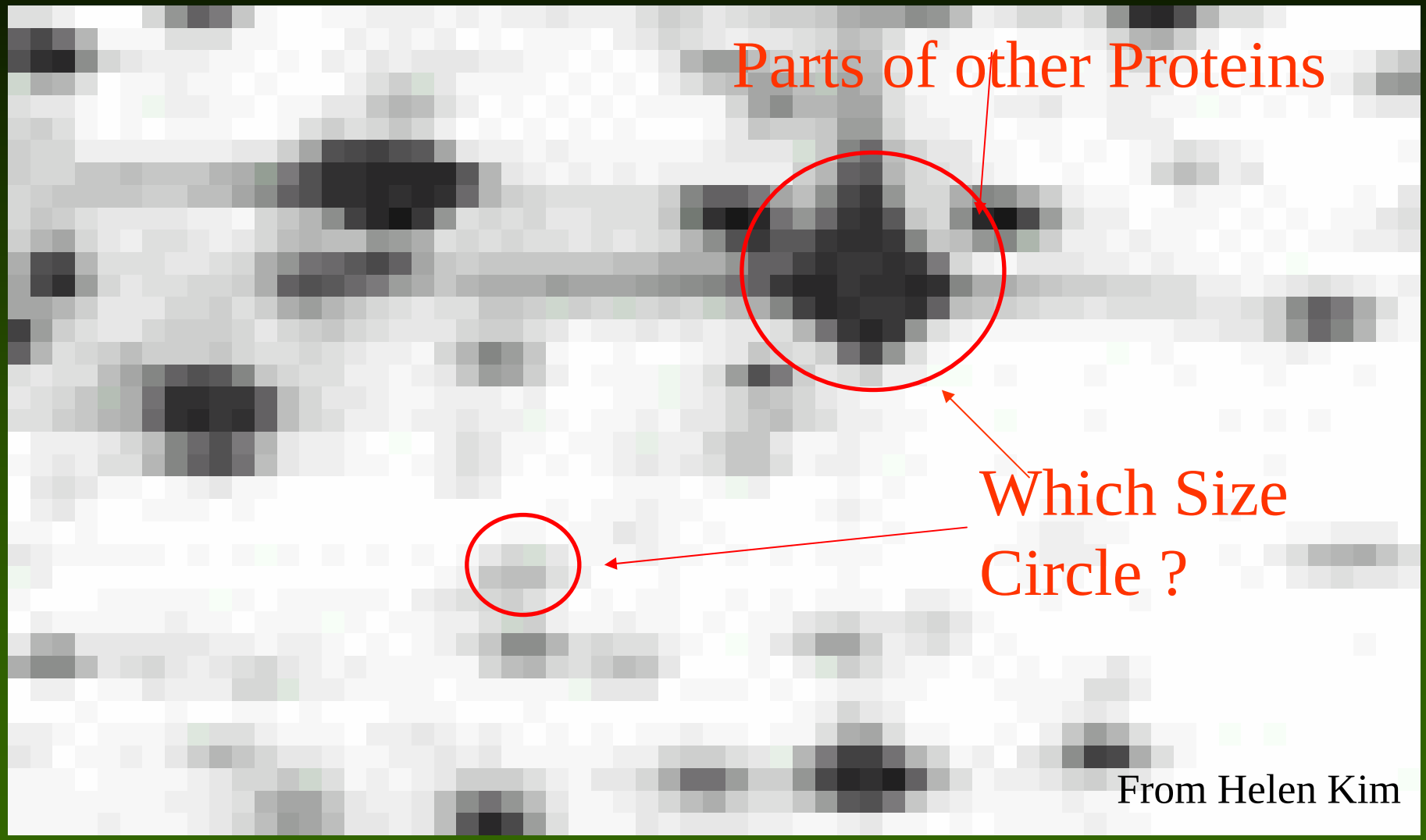
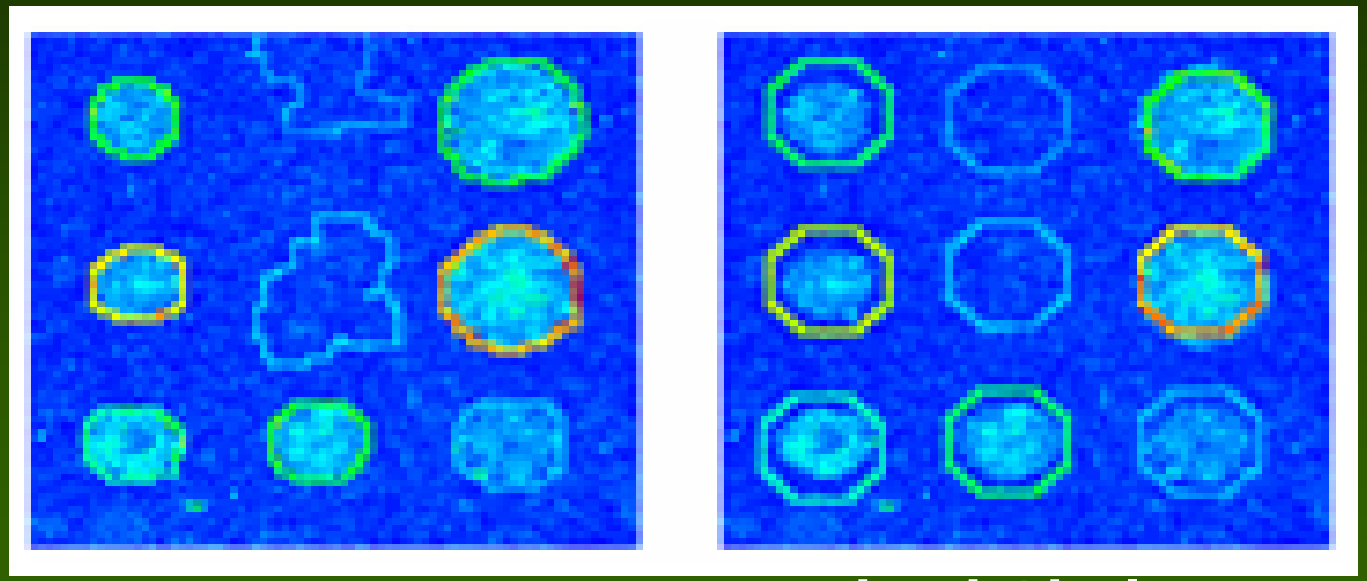

Parts of other Proteins
Which Size
Circle ?
From Helen Kim

Image Analysis



SRG **Fixed Circle**

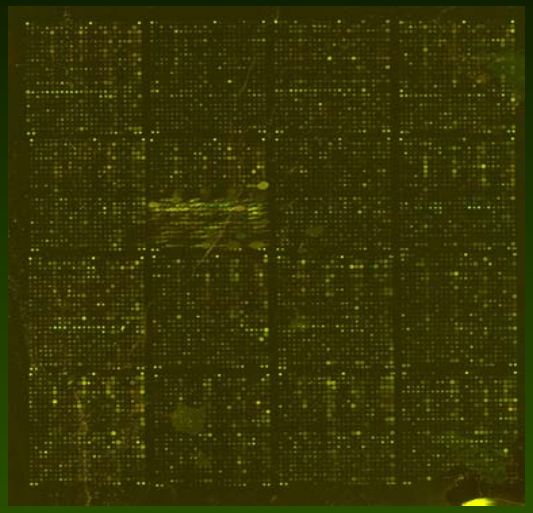
Inside the boundary is spot (foreground), outside is not.

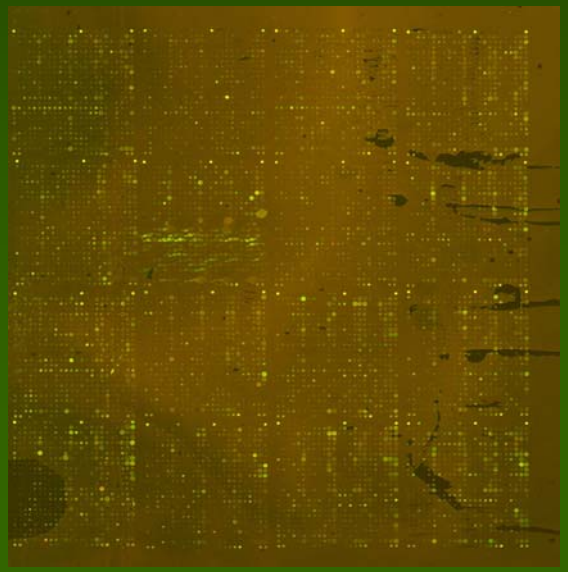
From T. Speed

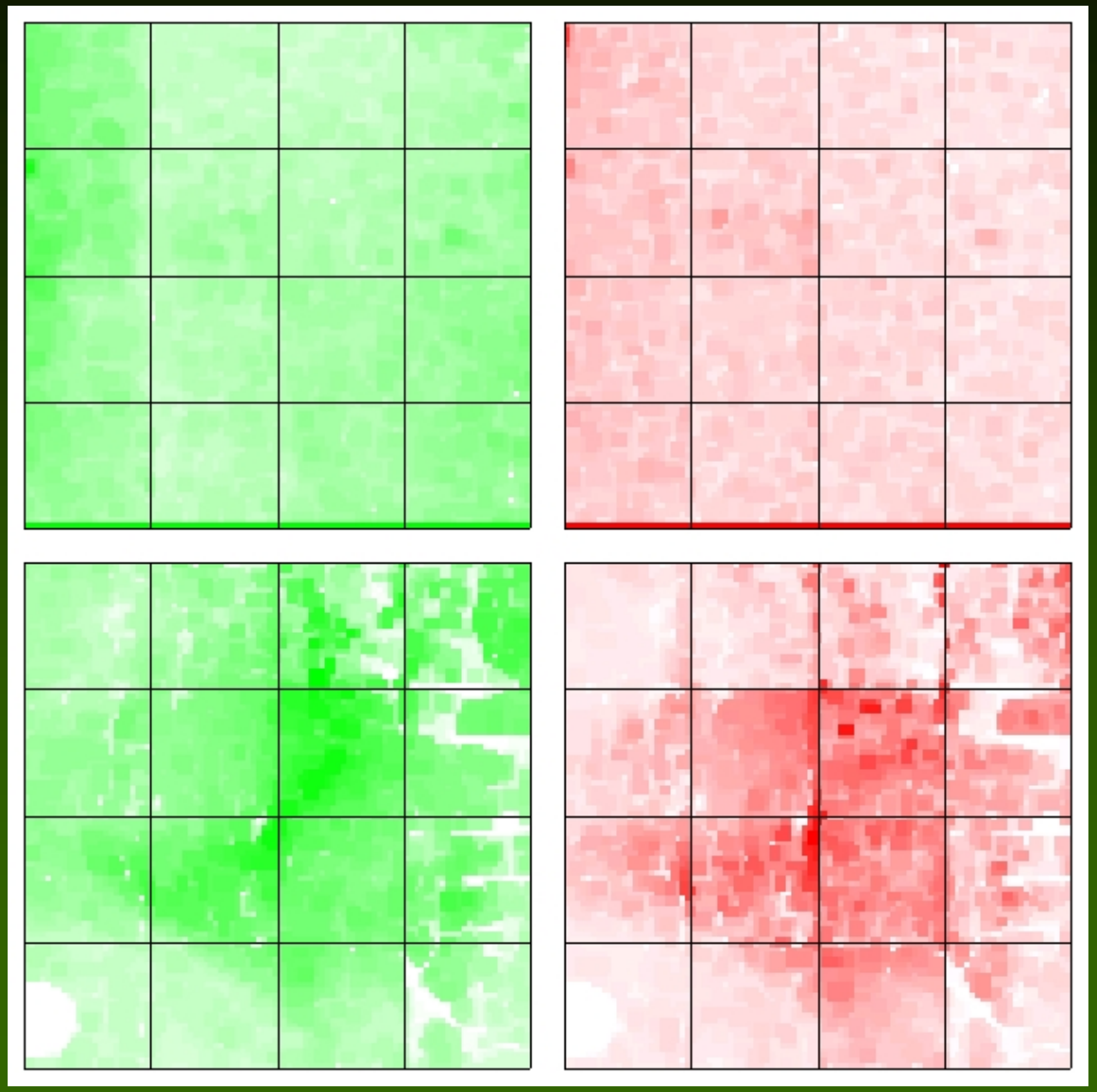
Quality Control/Normalization

- 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

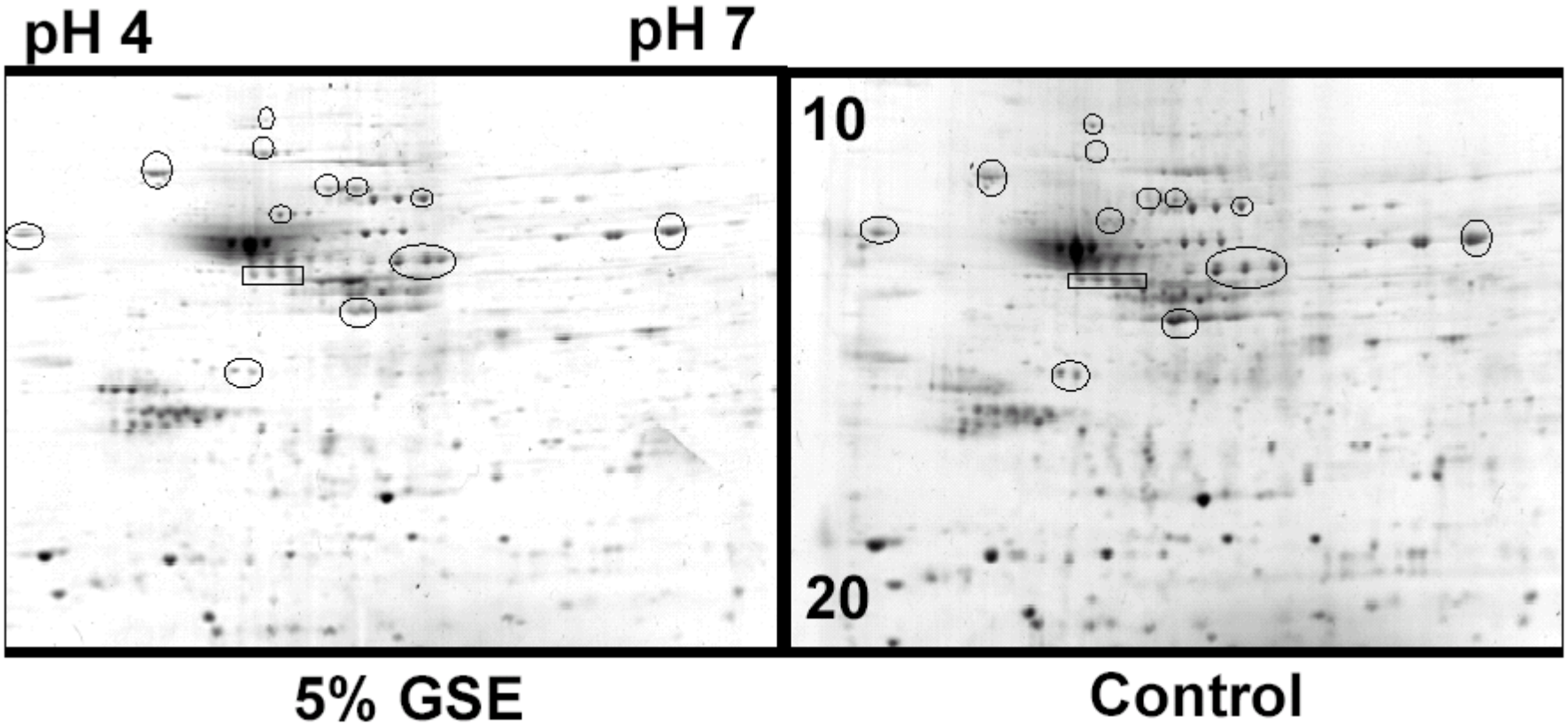
Spatial plots: background from the two slides



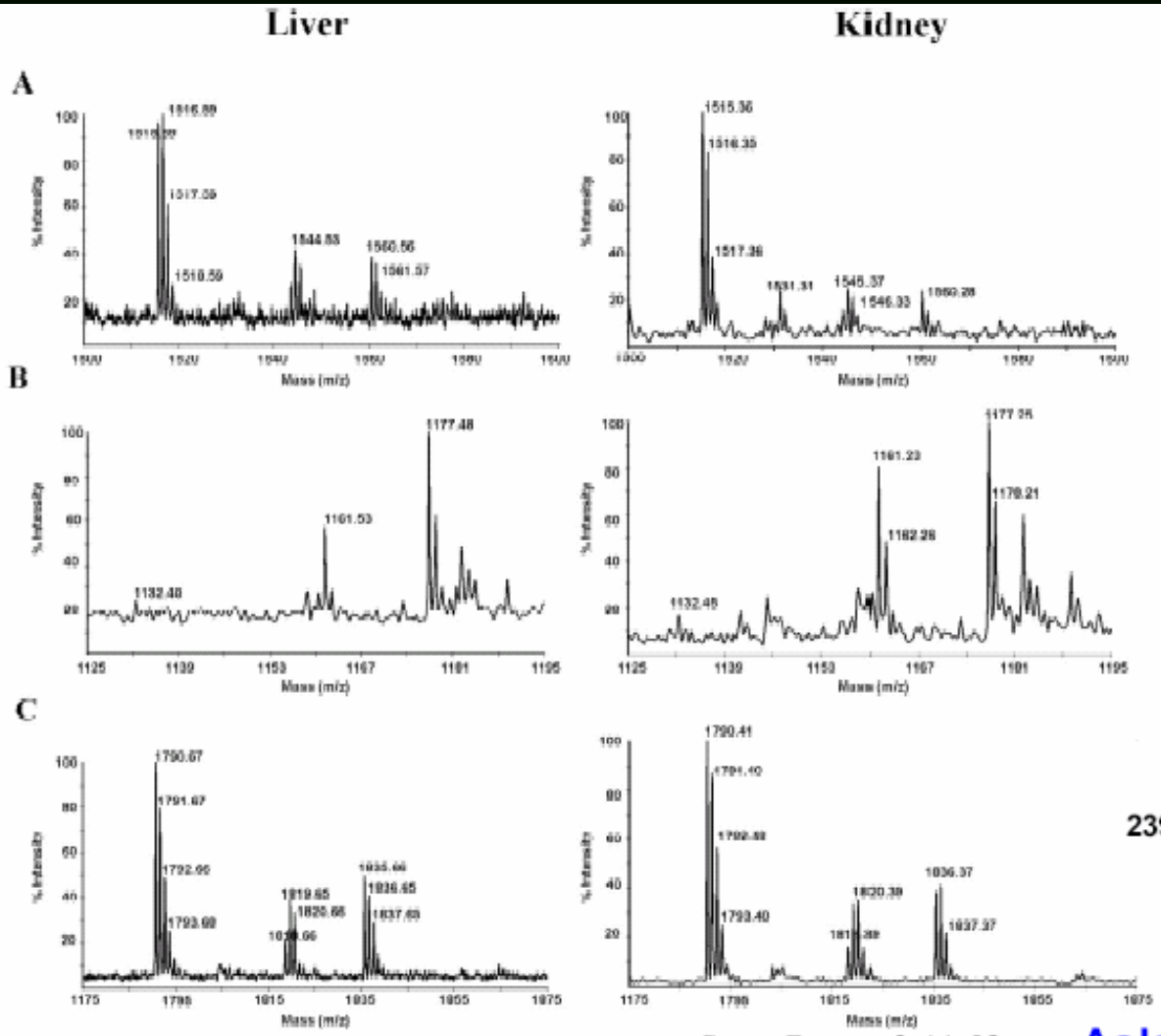




From T. Speed



From Helen Kim



^{85}IWHHTFYNELR^{95}

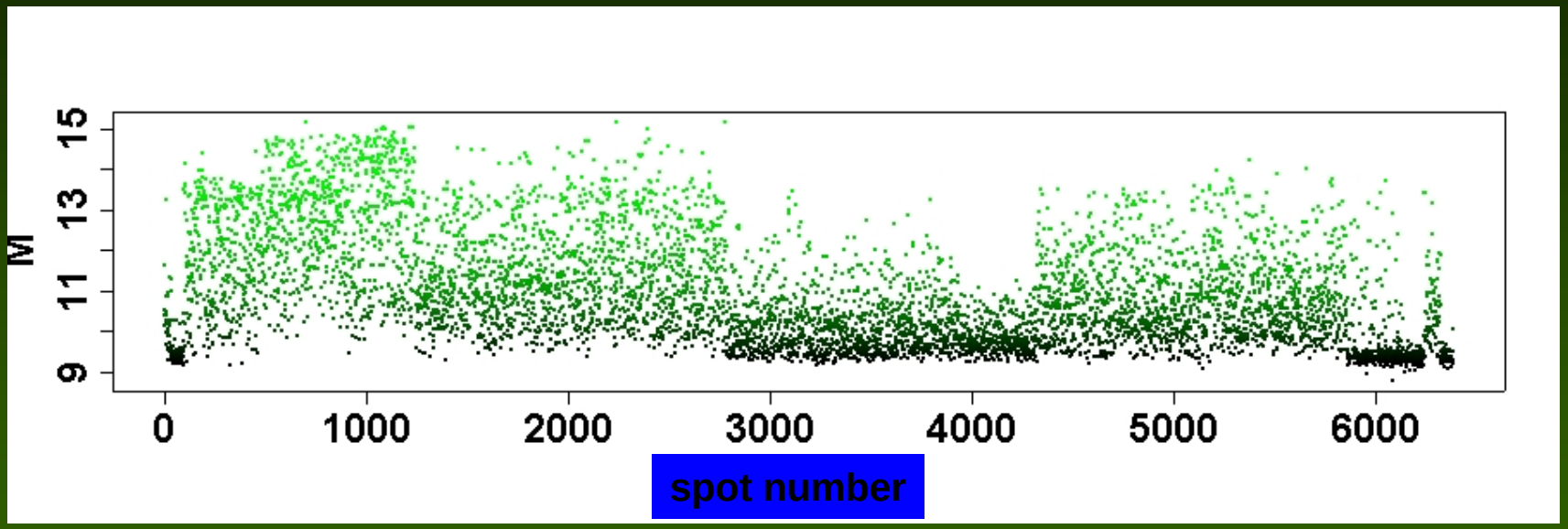
^{197}GYSFTTTAER^{206}

^{239}SYELPDGQVITIGNER^{254}

Steve Barnes 2-11-03

Aslan et al., JBC 278:4194

Time of printing effects

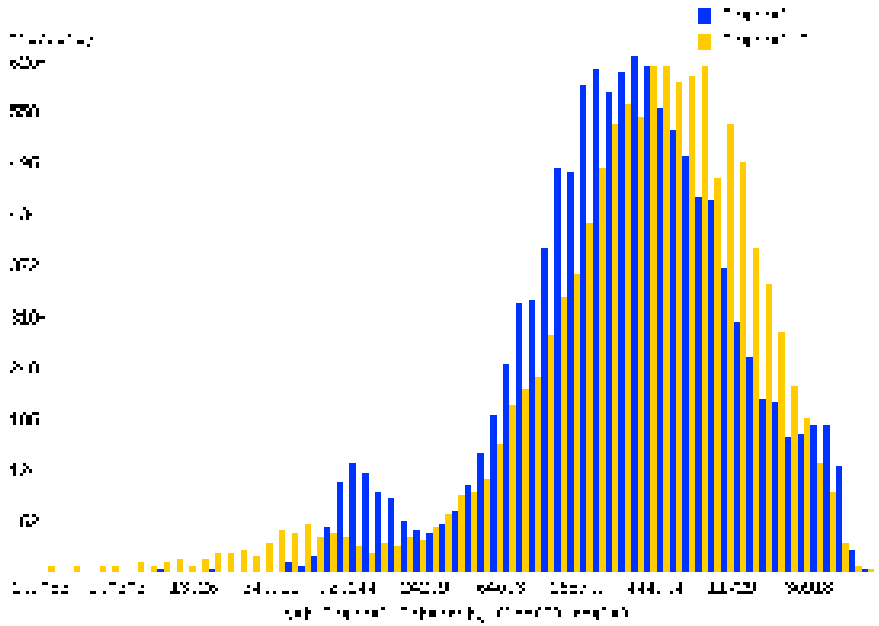


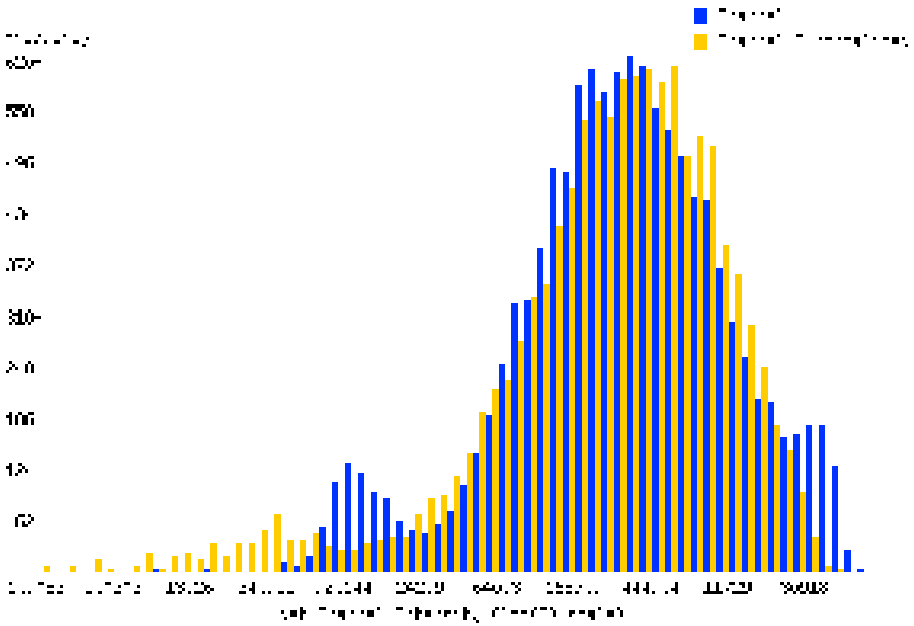
Green channel intensities ($\log_2 G$). Printing over 4.5 days.
The previous slide depicts a slide from this print run.
From T. Speed/H Yang

AFGC

Mean Normalization

Intensity correction only

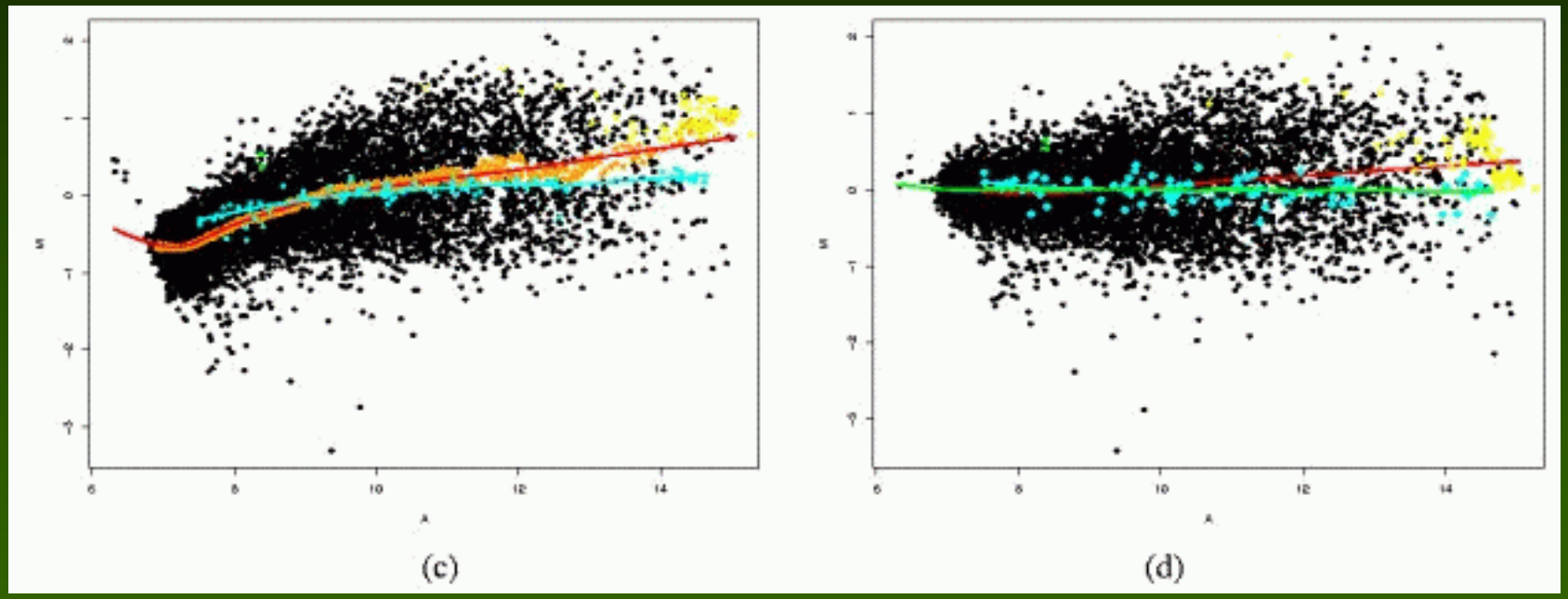




Before

After

Composite normalization



Before and after composite normalization

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

From T. Speed

Statistical Analysis

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

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

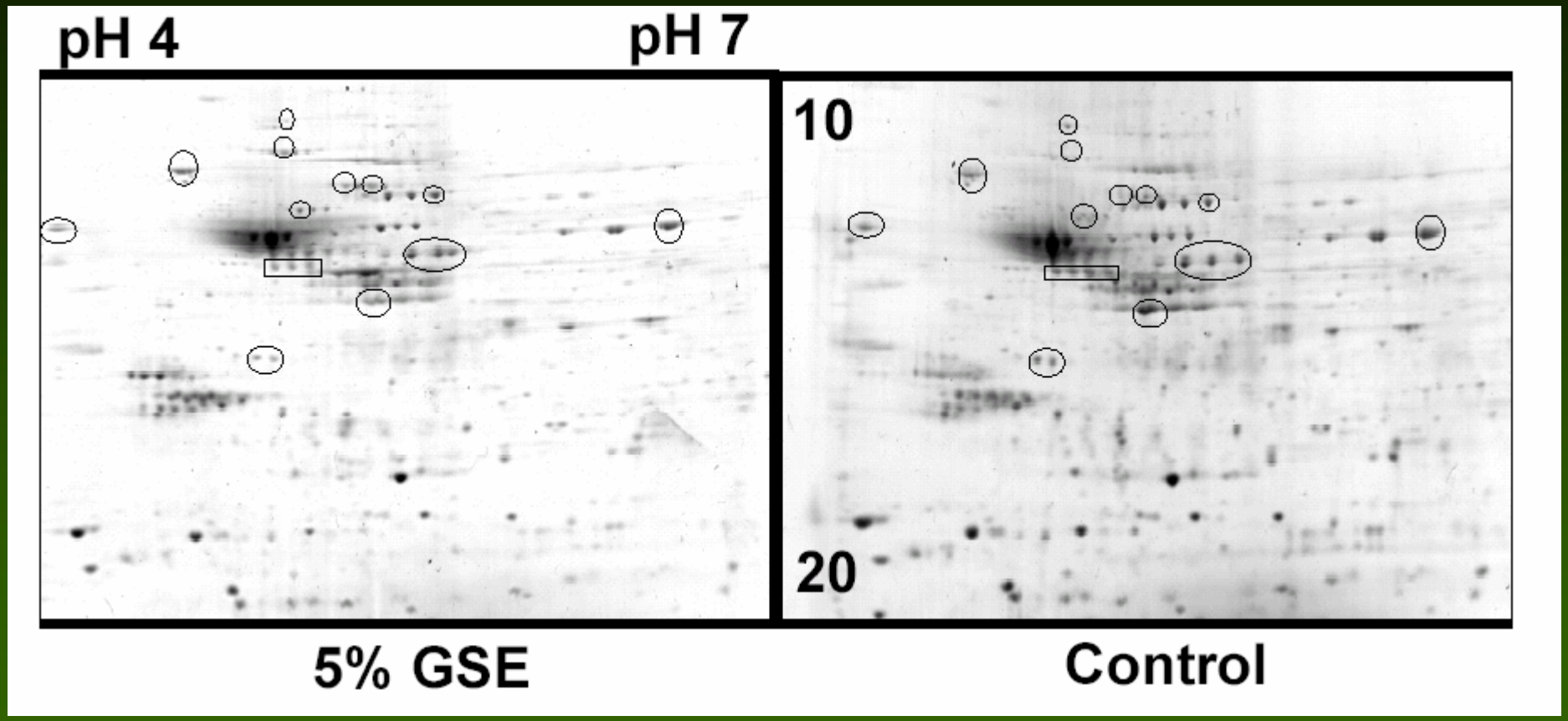
a) There is less than a 5% chance that the results are due to chance.

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



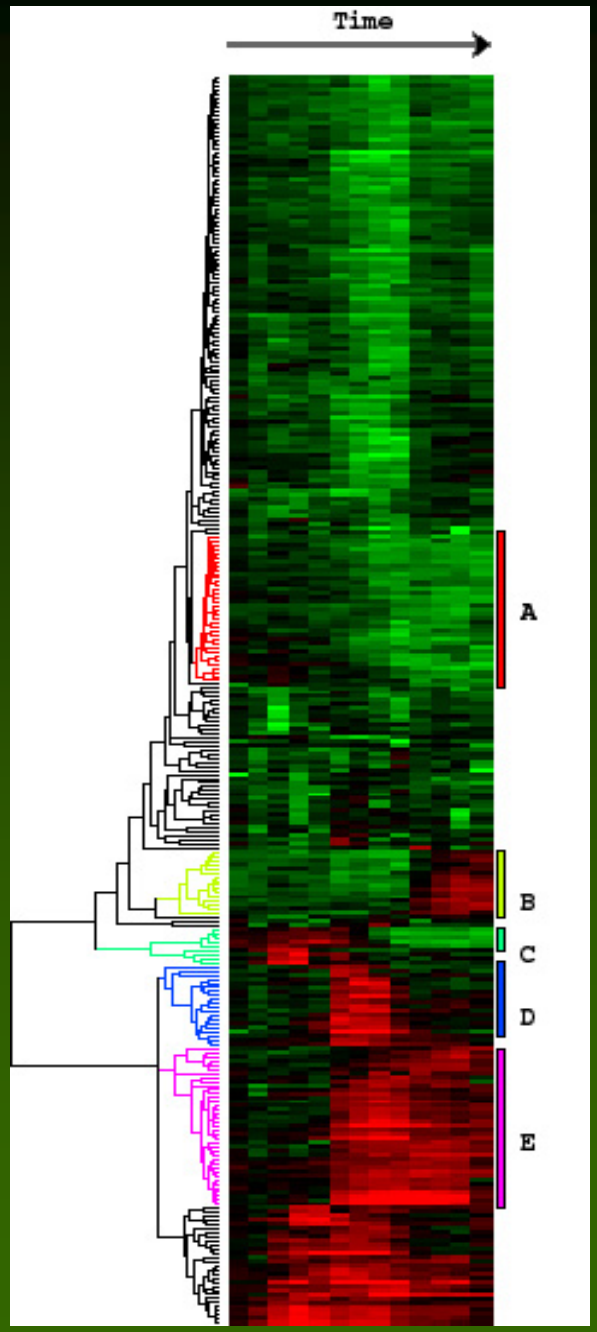
From H Kim

Class Discovery

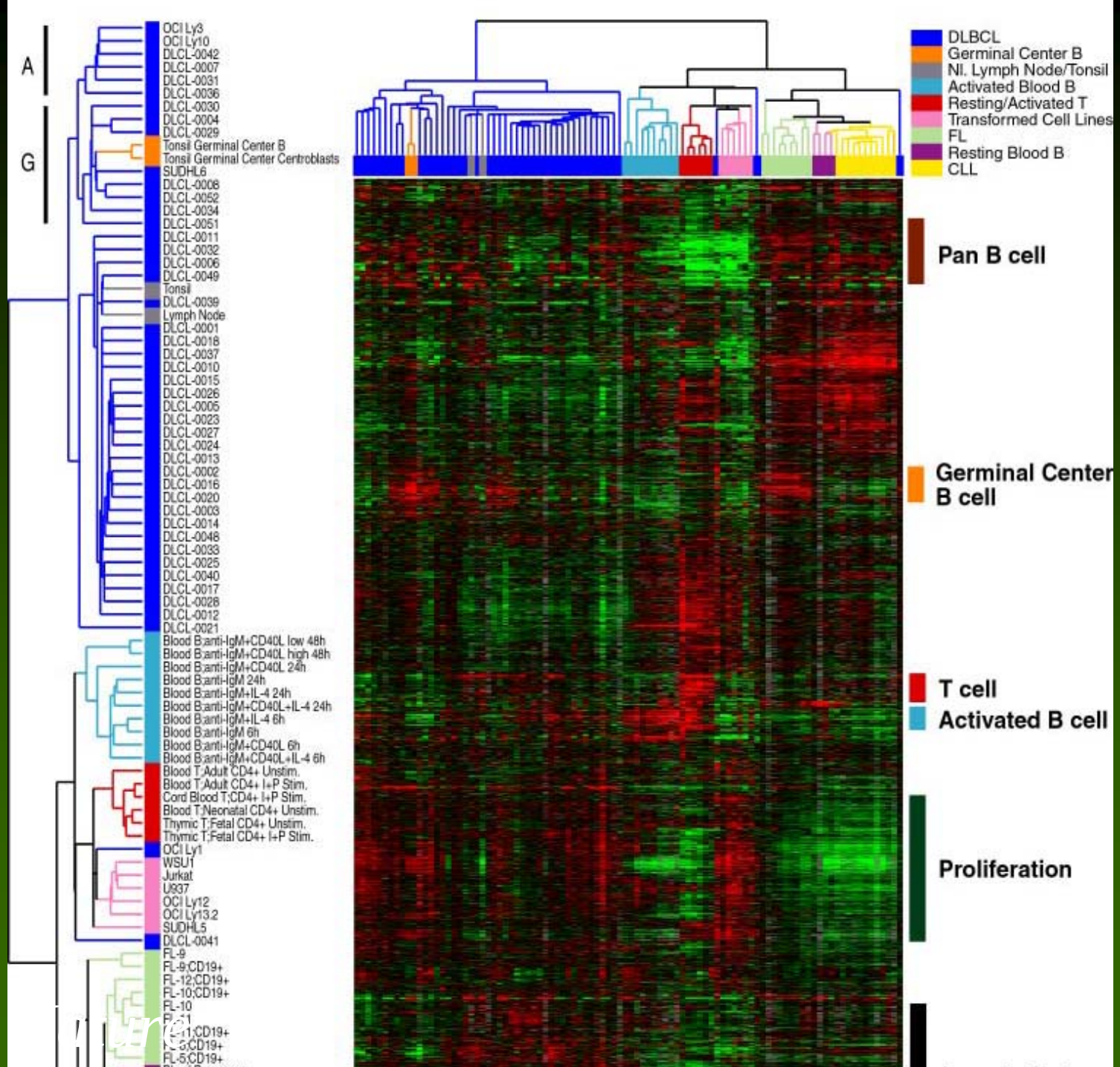
- Data visualization
- Cluster analysis
  - Clustering
  - Self organizing maps
- Multidimensional scaling
- Similarity searching

Clustering

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

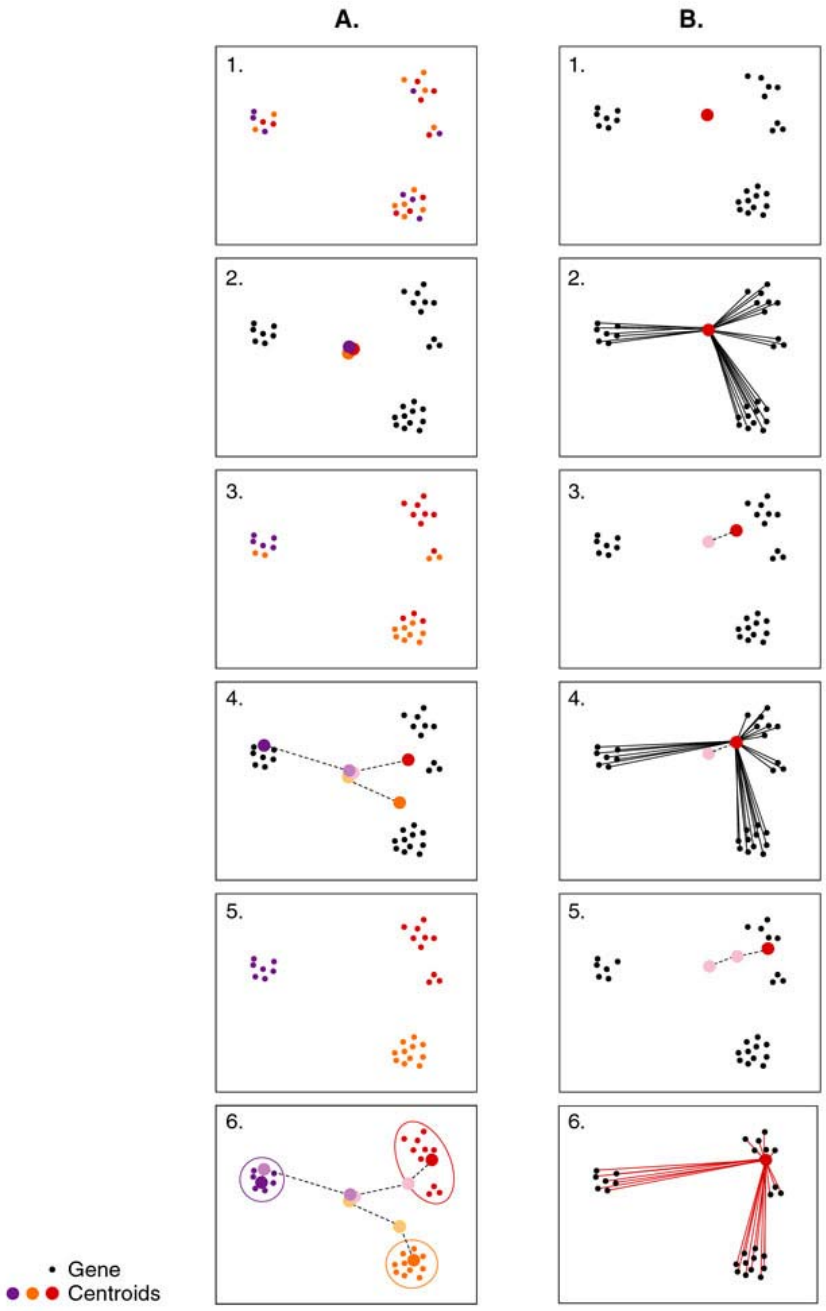


From *Nature*

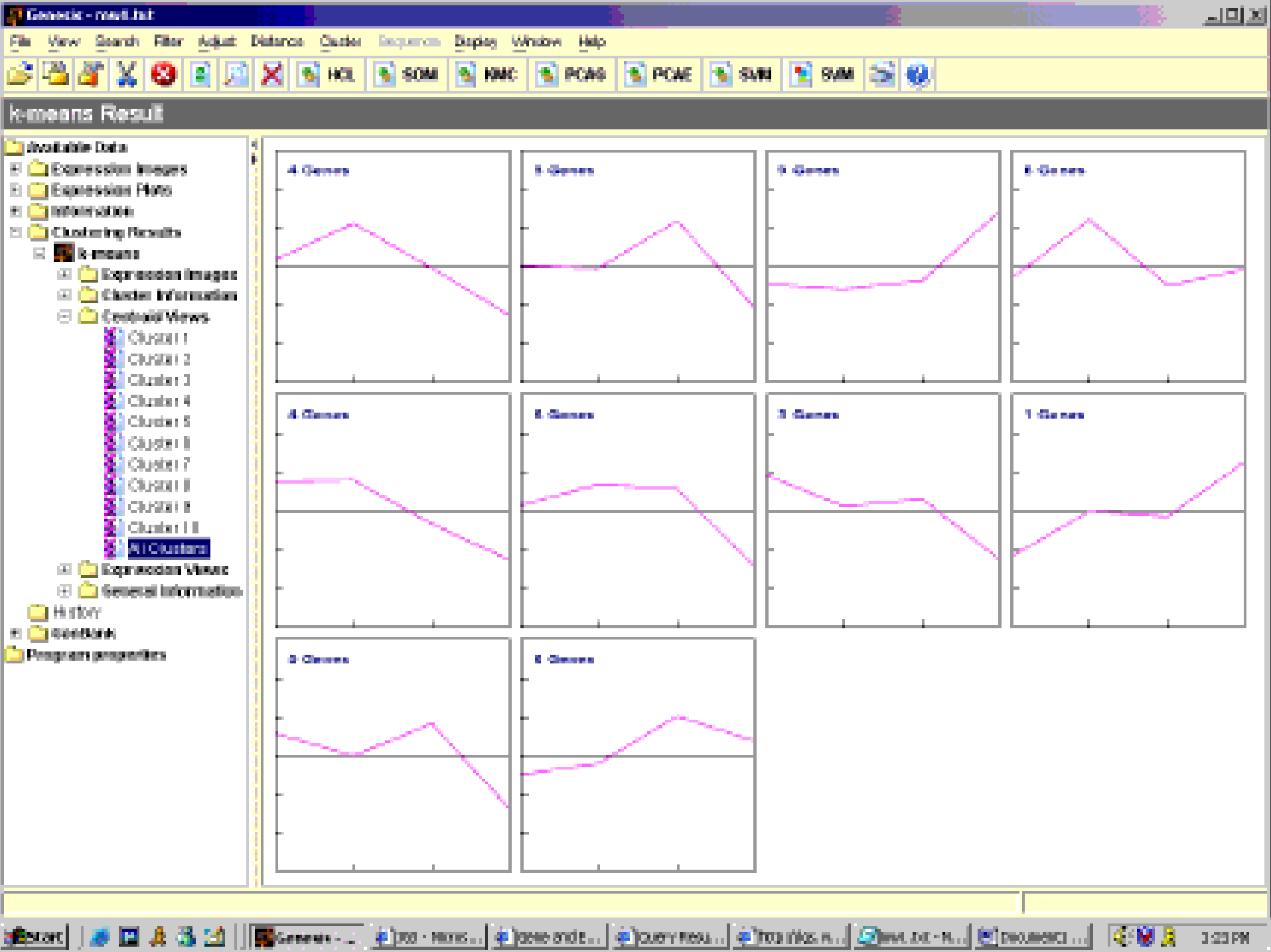

A
G
OCI Ly3
OCI Ly10
DLCL-0042
DLCL-0007
DLCL-0031
DLCL-0036
DLCL-0030
DLCL-0004
DLCL-0029
Tonsil Germinal Center B
Tonsil Germinal Center Centroblasts
SUDHL6
DLCL-0008
DLCL-0052
DLCL-0034
DLCL-0051
DLCL-0011
DLCL-0032
DLCL-0006
DLCL-0049
Tonsil
DLCL-0039
Lymph Node
DLCL-0001
DLCL-0018
DLCL-0037
DLCL-0010
DLCL-0015
DLCL-0026
DLCL-0005
DLCL-0023
DLCL-0027
DLCL-0024
DLCL-0013
DLCL-0002
DLCL-0016
DLCL-0020
DLCL-0003
DLCL-0014
DLCL-0048
DLCL-0033
DLCL-0025
DLCL-0040
DLCL-0017
DLCL-0028
DLCL-0012
DLCL-0021
Blood B;anti-IgM+CD40L low 48h
Blood B;anti-IgM+CD40L high 48h
Blood B;anti-IgM+CD40L 24h
Blood B;anti-IgM 24h
Blood B;anti-IgM+IL-4 24h
Blood B;anti-IgM+CD40L+IL-4 24h
Blood B;anti-IgM+IL-4 6h
Blood B;anti-IgM 6h
Blood B;anti-IgM+CD40L 6h
Blood B;anti-IgM+CD40L+IL-4 6h
Blood T;Adult CD4+ Unstim.
Blood T;Adult CD4+ I+P Stim.
Cord Blood T;CD4+ I+P Stim.
Blood T;Neonatal CD4+ Unstim.
Thymic T;Fetal CD4+ Unstim.
Thymic T;Fetal CD4+ I+P Stim.
OCI Ly1
WSU1
Jurkat
U937
OCI Ly12
OCI Ly13.2
SUDHL5
DLCL-0041
FL-9
FL-9;CD19+
FL-12;CD19+
FL-10;CD19+
FL-10
DLBCL
Germinal Center B
NI. Lymph Node/Tonsil
Activated Blood B
Resting/Activated T
Transformed Cell Lines
FL
Resting Blood B
CLL
Pan B cell
Germinal Center B cell
T cell
Activated B cell
Proliferation

From

K-Means Clustering

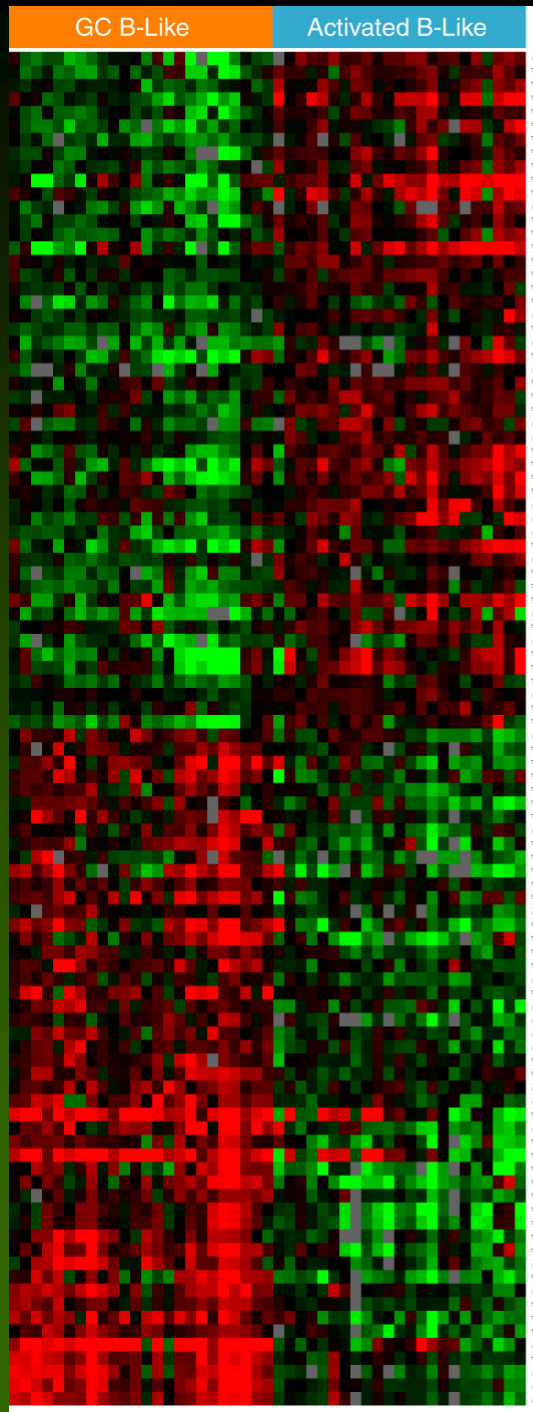


Source Unknown

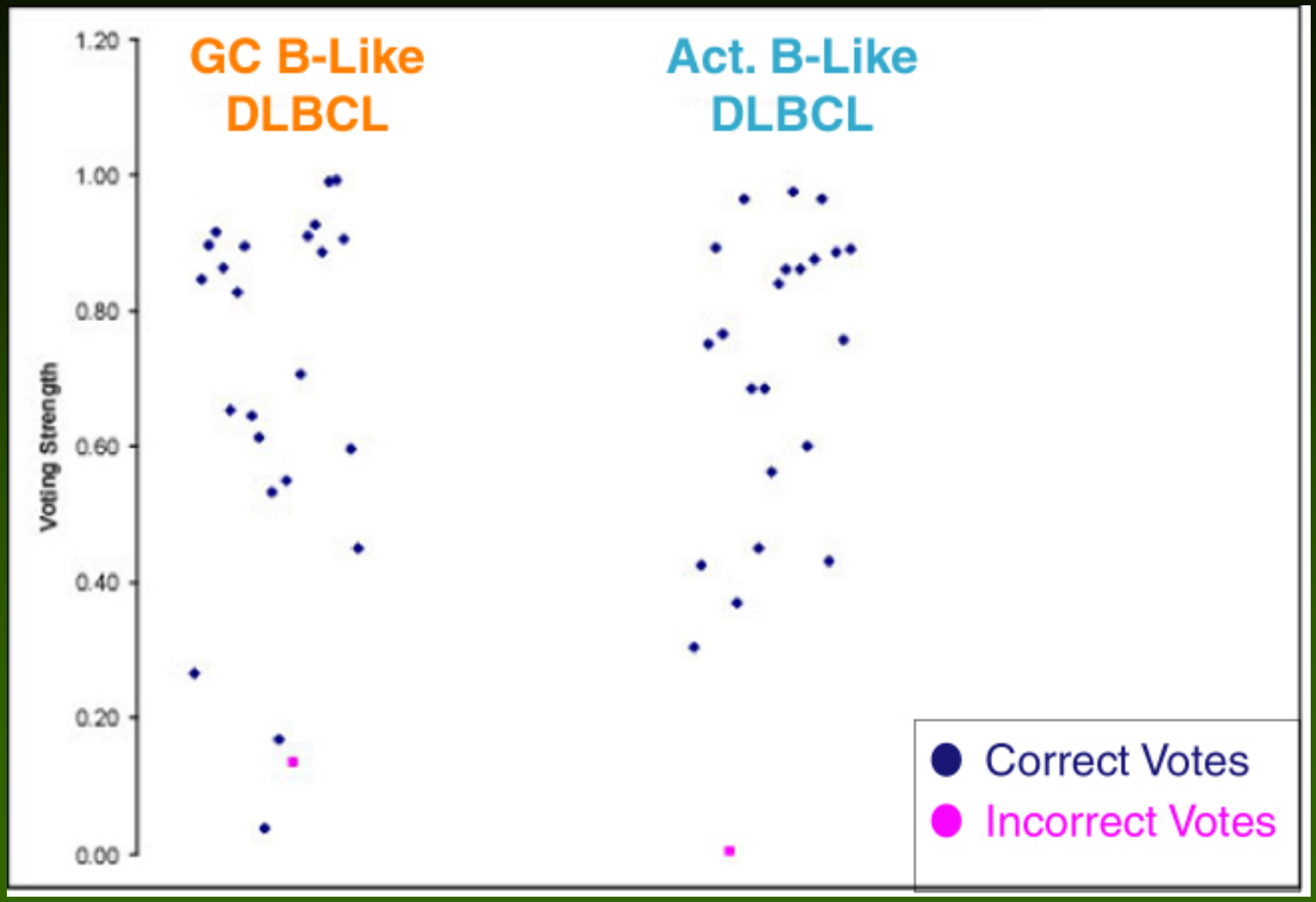

File View Search Filter Adjust Distance Cluster Sequences Display Window Help
HCL
SOM
KMC
PCAG
PCAE
SVM
SVM
k-means Result
Available Data
Expression Images
Expression Plots
Information
Clustering Results
k-means
Expression Images
Cluster Information
Centroid Views
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10
All Clusters
Expression Views
General Information
History
GeneBank
Program properties
4 Genes
5 Genes
9 Genes
6 Genes
4 Genes
6 Genes
3 Genes
1 Genes
3 Genes
6 Genes

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



(Unknown UG Hs.169081 ets variant gene 6 (TEL oncogene); Clone=1355435)
*Deoxycytidylate deaminase; Clone=1302032
*T-cell protein-tyrosine phosphatase=Protein tyrosine phosphatase, non-receptor type 2; Clone=665903
*Cyclin D2/KIAK0002=3' end of KIAK0002 cDNA; Clone=1357360
*Deoxycytidylate deaminase; Clone=1185959
*Potassium voltage-gated channel, shaker-related subfamily, member 3; Clone=1337856
*Unknown; Clone=1350877
*Deoxycytidylate deaminase; Clone=489681
*T-cell protein-tyrosine phosphatase=Protein tyrosine phosphatase, non-receptor type 2; Clone=740402
*IRF-4=LSIRF=Mum1=homologue of Pip=Lymphoid-specific interferon regulatory factor =Multiple myeloma oncogene 1; Clone=270770
*Cyclin D2/KIAK0002=3' end of KIAK0002 cDNA; Clone=366412
(Unknown; Clone=825920)
*Deoxycytidylate deaminase; Clone=489681
*T-cell protein-tyrosine phosphatase=Protein tyrosine phosphatase, non-receptor type 2; Clone=1370148
*IRF-4=LSIRF=Mum1=homologue of Pip=Lymphoid-specific interferon regulatory factor =Multiple myeloma oncogene 1; Clone=1272196
*MCL1=myeloid cell differentiation protein; Clone=711870
*core binding factor alpha1b subunit=CBF alpha1=PEBP2aA1 transcription factor =AML1 Proto-oncogene=translocated in acute myeloid leukemia; Clone=263251
*zinc finger protein 42 MZF-1; Clone=490387
*Unknown; Clone=1372162
(Unknown UG Hs.55947 Homo sapiens mRNA for KIAA0805 protein, partial cds; Clone=1288180)
*PKU-beta=KIAA0137=protein kinase; Clone=825383
(Unknown; Clone=1352715)
*SLAP=src-like adapter protein; Clone=52564
(XE7=B-lymphocyte surface protein; Clone=1339106)
*erk3=extracellular signal-regulated kinase 3; Clone=50506
*PRK=putative serine/threonine protein kinase; Clone=739192
*MAPKAP kinase (3pK); Clone=1336478
(Unknown UG Hs.79937 ESTs; Clone=682976)
(dual specificity phosphatase tyrosine/serine; Clone=291332)
*FLICE-like inhibitory protein long form=I-FLICE=FLAME-1=Casper=MRIT=CASH=cFLIP=CLARP; Clone=711633
*SLAP=src-like adapter protein; Clone=815774
*PTP-1B=phosphotyrosyl-protein phosphatase; Clone=472182
*Pak1=p21-activated protein kinase; Clone=595474
(Protein disulfide isomerase-related protein (PDIR); Clone=703707)
(Unknown UG Hs.143722 ESTs, Moderately similar to !!!! ALU SUBFAMILY SQ WARNING ENTRY !!!! [H.sapiens]; Clone=705272)
(Unknown; Clone=1340742)
*SLAP=src-like adapter protein; Clone=701768
(Smad4=DPC4=Homologue of Mothers Against Decapentaplegic (MAD)=required for TGF beta signaling=tumor suppressor in pancreatic cancer; Clone=774619)
*PKU-beta=KIAA0137=protein kinase; Clone=563451
*BMI-1; Clone=1048586
*PTP-1B=phosphotyrosyl-protein phosphatase; Clone=685177
(EDG-1=endothelial differentiation protein=putative G-protein-coupled receptor; Clone=307325)
*BAK=BCL-2 family member; Clone=1288183
(Unknown UG Hs.59368 ESTs; Clone=1353778)
*CD44=Pgp-1=extracellular matrix receptor-III=Hyaluronate receptor; Clone=713145
*CD44=Pgp-1=extracellular matrix receptor-III=Hyaluronate receptor; Clone=703824
*Transforming growth factor, beta receptor II (70-80kD); Clone=1351378
(CDC37 homolog=subunit of Hsp90; Clone=346753)
*pM5 protein=homology to conserved regions of the collagenase gene family; Clone=1357489
*3' 5'-cyclic AMP phosphodiesterase=rolipram-sensitive cAMP-specific phosphodiesterase (PDE2); Clone=377708
(Unknown UG Hs.86987 ESTs, Highly similar to (defline not available 5059425) [H.sapiens]; Clone=825854)
*Unknown UG Hs.192708 ESTs, Highly similar to A-myb N-terminal region)2341 is 2nd base in codon) [H.sapiens]; Clone=745995
*Unknown UG Hs.28355 ESTs; Clone=703735
*BCL-6; Clone=712395
*TdT = Terminal Deoxynucleotide Transferase; Clone=667782
*KIAA0093=NEDD-4=E3 ubiquitin protein ligase; Clone=135343
*Unknown; Clone=684877
(Unknown; Clone=2020)
*BCL-7A; Clone=1337241
*Unknown UG Hs.125815 ESTs; Clone=1252102
*CD10=CALLA=Neprilysin=enkepalinase; Clone=200814
*Cyclin H; Clone=795296
*BCL-6; Clone=1340526
(Unknown; Clone=1240688)
*CD10=CALLA=Neprilysin=enkepalinase; Clone=1286850
*JAW1=lymphoid-restricted membrane protein; Clone=815539
*Unknown UG Hs.186709 ESTs, Weakly similar to !!!! ALU SUBFAMILY SB WARNING ENTRY !!!! [H.sapiens]; Clone=825852
*Unknown UG Hs.222808 ESTs; Clone=815273
(Similar to intersectin=adaptor protein with two EH and five SH3 domains; Clone=1339781)
*JNK3=Stress-activated protein kinase; Clone=23173
(Unknown UG Hs.219237 ESTs, Highly similar to !!!! ALU SUBFAMILY SX WARNING ENTRY !!!! [H.sapiens]; Clone=1372254)
(Unknown; Clone=1334297)
(Unknown UG Hs.231798 ESTs; Clone=827169)
(Unknown; Clone=1270568)
*RPD3L1=homologue of yeast RPD3 transcription factor; Clone=814080
*DNA (cytosine-5-)-methyltransferase; Clone=1320361
(Unknown UG Hs.163222 ESTs; Clone=1338044)
(Unknown; Clone=2005)
*TTG-2=Rhombotin-2=translocated in t(11;14)(p13;q11) T cell acute lymphocytic leukemia=cysteine rich protein with LIM motif; Clone=685456
(Unknown UG Hs.120245 Homo sapiens mRNA for KIAA1039 protein, partial cds; Clone=1268870)
*FMR2=Fragile X mental retardation 2=putative transcription factor=LAF-4 and AF-4 homologue; Clone=1352112
*TTG-2=Rhombotin-2=translocated in t(11;14)(p13;q11) T cell acute lymphocytic leukemia=cysteine rich protein with LIM motif; Clone=712829
*myb-related gene A=A-myb; Clone=1367994
*JAW1=lymphoid-restricted membrane protein; Clone=815539
*Unknown UG Hs.145058 ESTs; Clone=824754
*Unknown UG Hs.124922 ESTs; Clone=1337653
*Unknown UG Hs.124922 ESTs; Clone=1358244
*Unknown; Clone=1351325
*JAW1=lymphoid-restricted membrane protein; Clone=417502
(Unknown UG Hs.137038 EST; Clone=1338981)
*myb-related gene A=A-myb; Clone=825476
(Unknown UG Hs.208410 EST, Moderately similar to !!!! ALU SUBFAMILY SB WARNING ENTRY !!!! [H.sapiens]; Clone=1353036)
*Unknown UG Hs.105261 EST; Clone=824088
*Unknown; Clone=1353041
*Unknown; Clone=1353015
(Unknown UG Hs.120716 ESTs; Clone=1334260)
*Unknown; Clone=825199
(Unknown UG Hs.224323 ESTs, Moderately similar to alternatively spliced product using exon 13A [H.sapiens]; Clone=1338448)
(Unknown UG Hs.136345 ESTs; Clone=746300)
(Unknown UG Hs.169565 ESTs, Moderately similar to !!!! ALU SUBFAMILY SB WARNING ENTRY !!!! [H.sapiens]; Clone=825217)


GC B-Like
DLBCL
Act. B-Like
DLBCL
Voting Strength
1.20
1.00
0.80
0.60
0.40
0.20
0.00
Correct Votes
Incorrect Votes

From *Nature*

Class Differentiation

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

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			
		Independent	Matched	Independent	Repeated Measures	Single	Multiple
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 generalized 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?

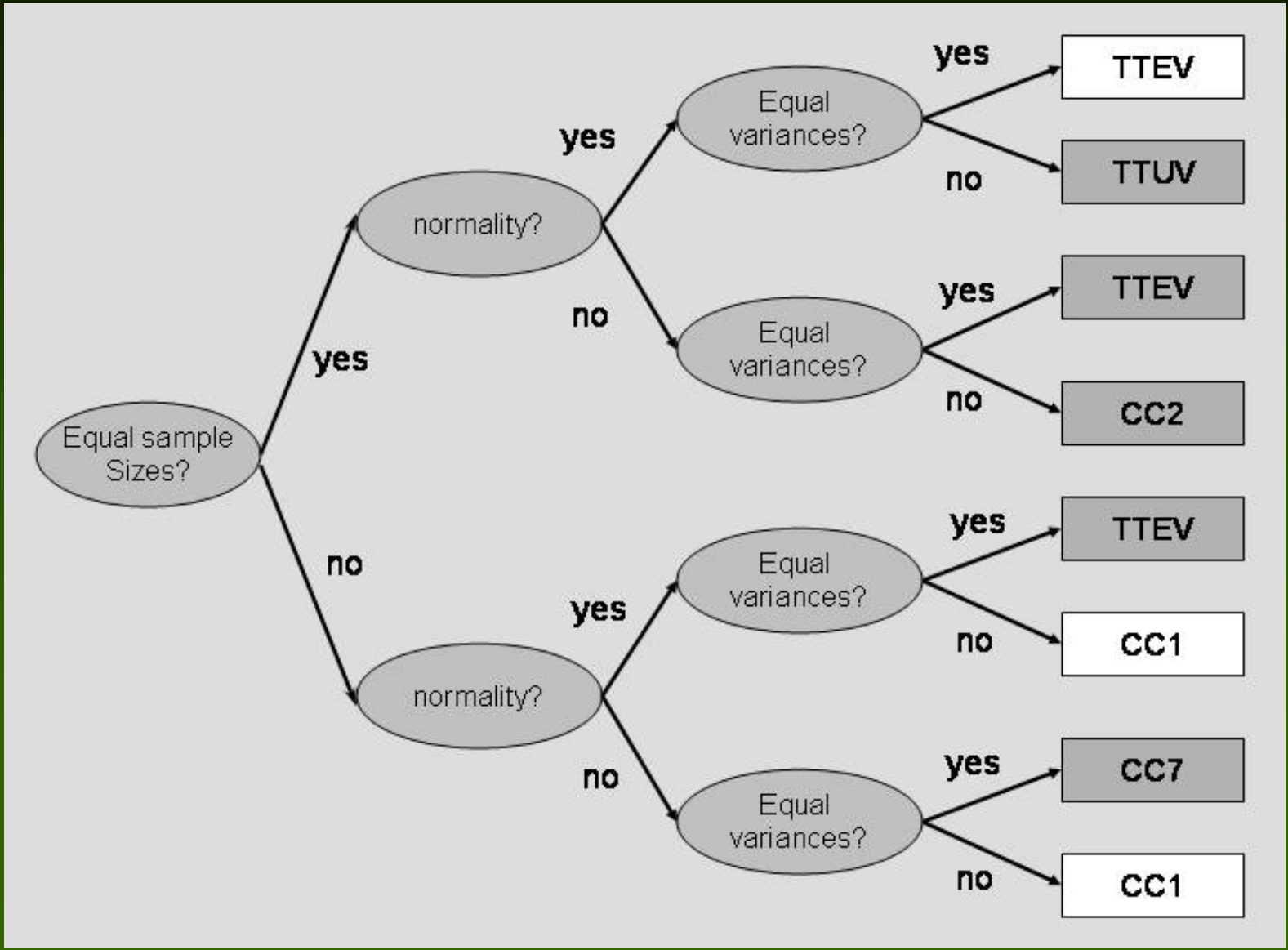
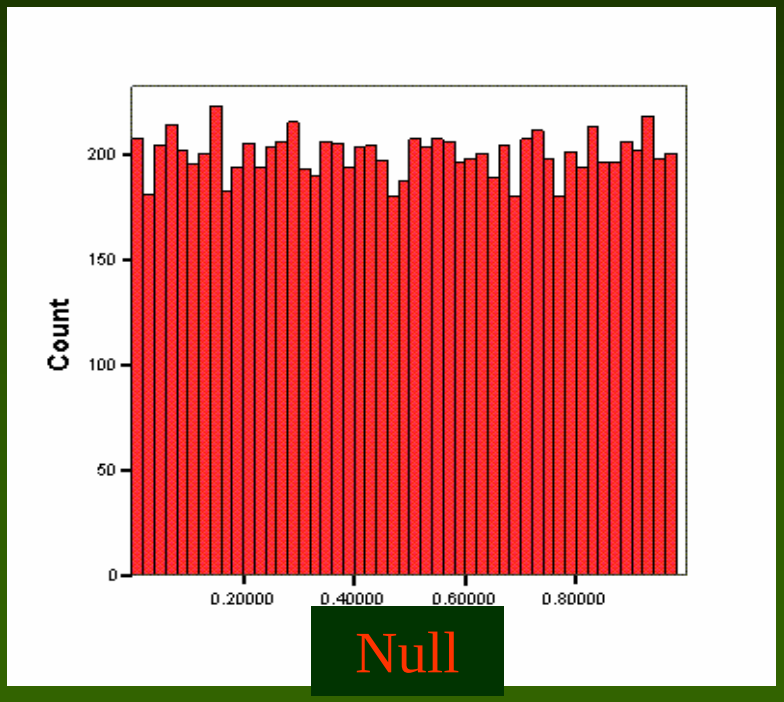
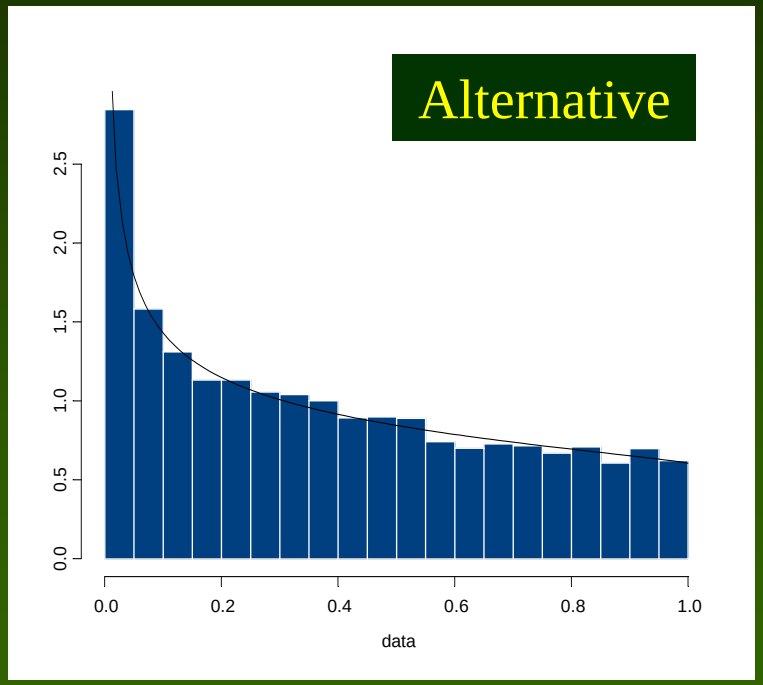


Figure 3. Mixture Model Approach from Allison et al. (2002).

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

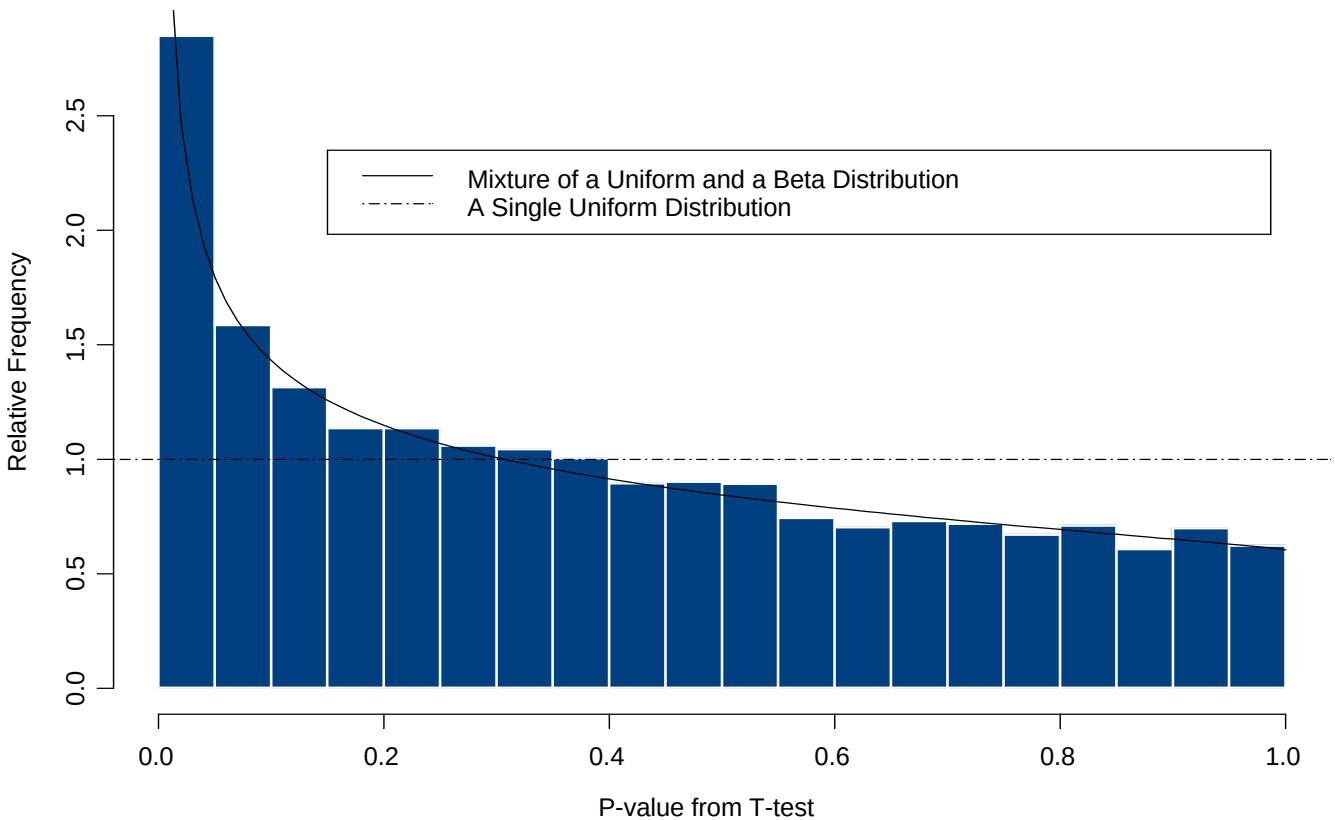
Under the null hypothesis, the distribution of p-values is uniform on the interval [0,1] regardless of the sample size and statistical test used (as long as that test is valid).





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

Fitted mixture model to 12,625 P-values



Testing Defined

Truth

Conclusion

	Null	Alt	
Null	**a**	**b**	K-R
Alt	**c**	**d**	R
	K-M	M	K

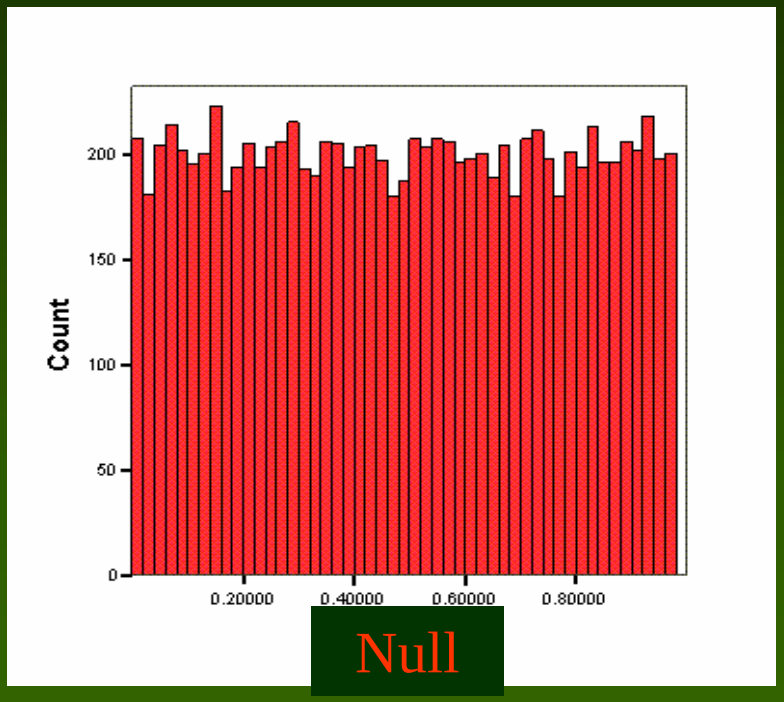
c = type 1 error (alpha) – false positive
b = type 2 error (beta) – false negative

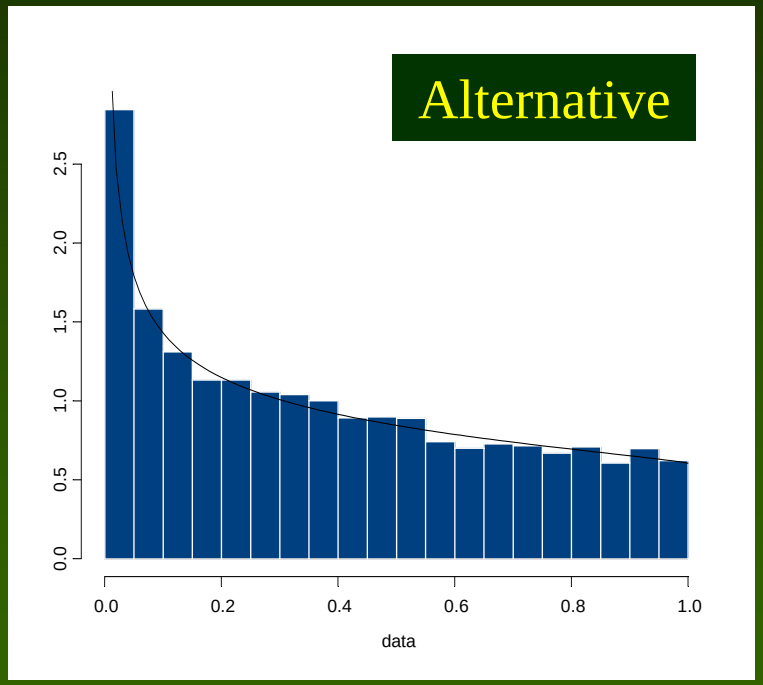
$$FDR = E\left(\frac{c}{c+d}\right)$$

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 proceedure
- Methods were developed for epidemiology and genetic studies, but were adapted for HDB studies

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.

Family Wise Error Rate vs. False Discovery Rate

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

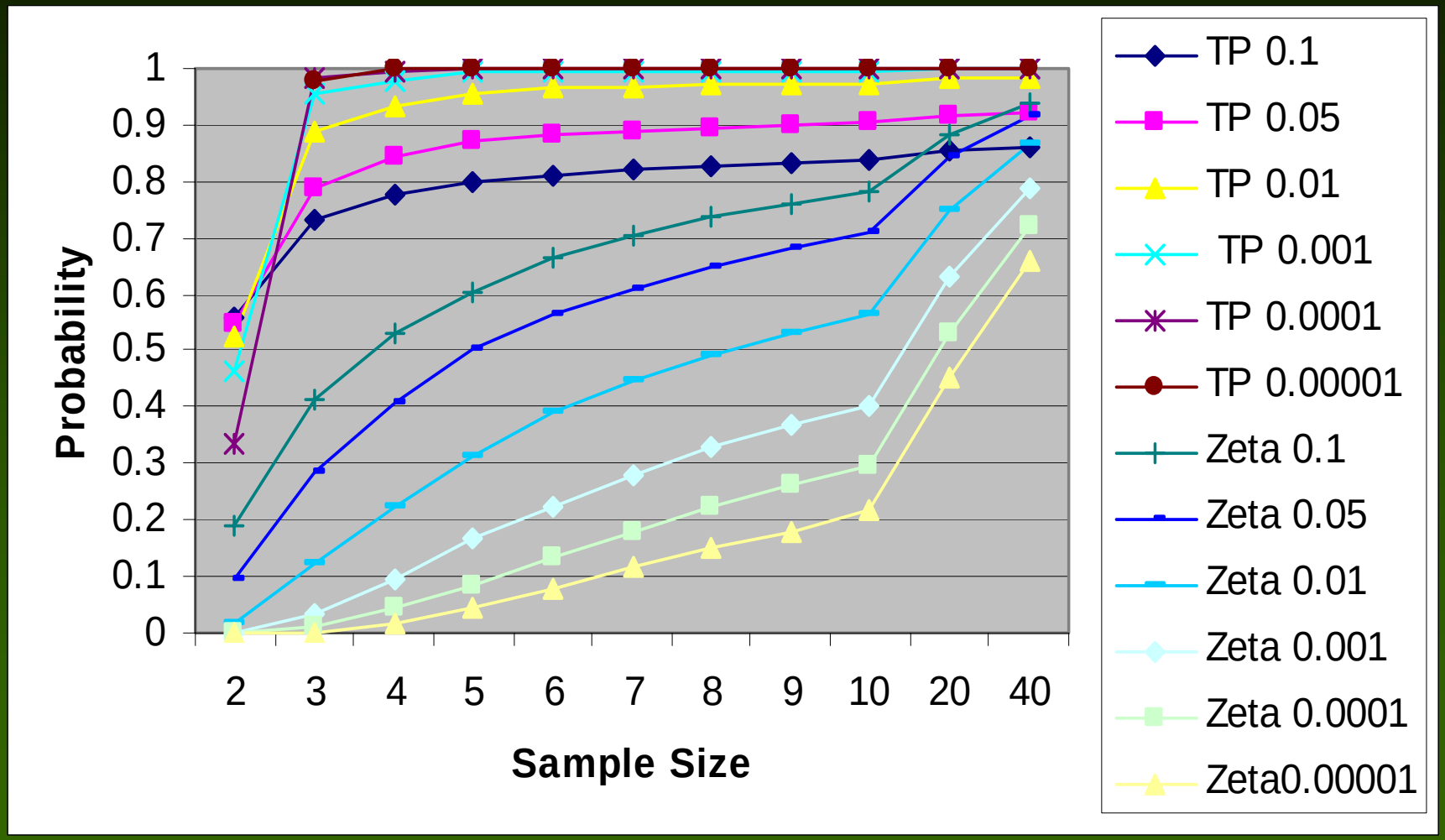
Power and Sample Size

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

Power For Powerful Effect



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 an 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_o if genes in a pathway are correlated there will be an increase in type 1 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?

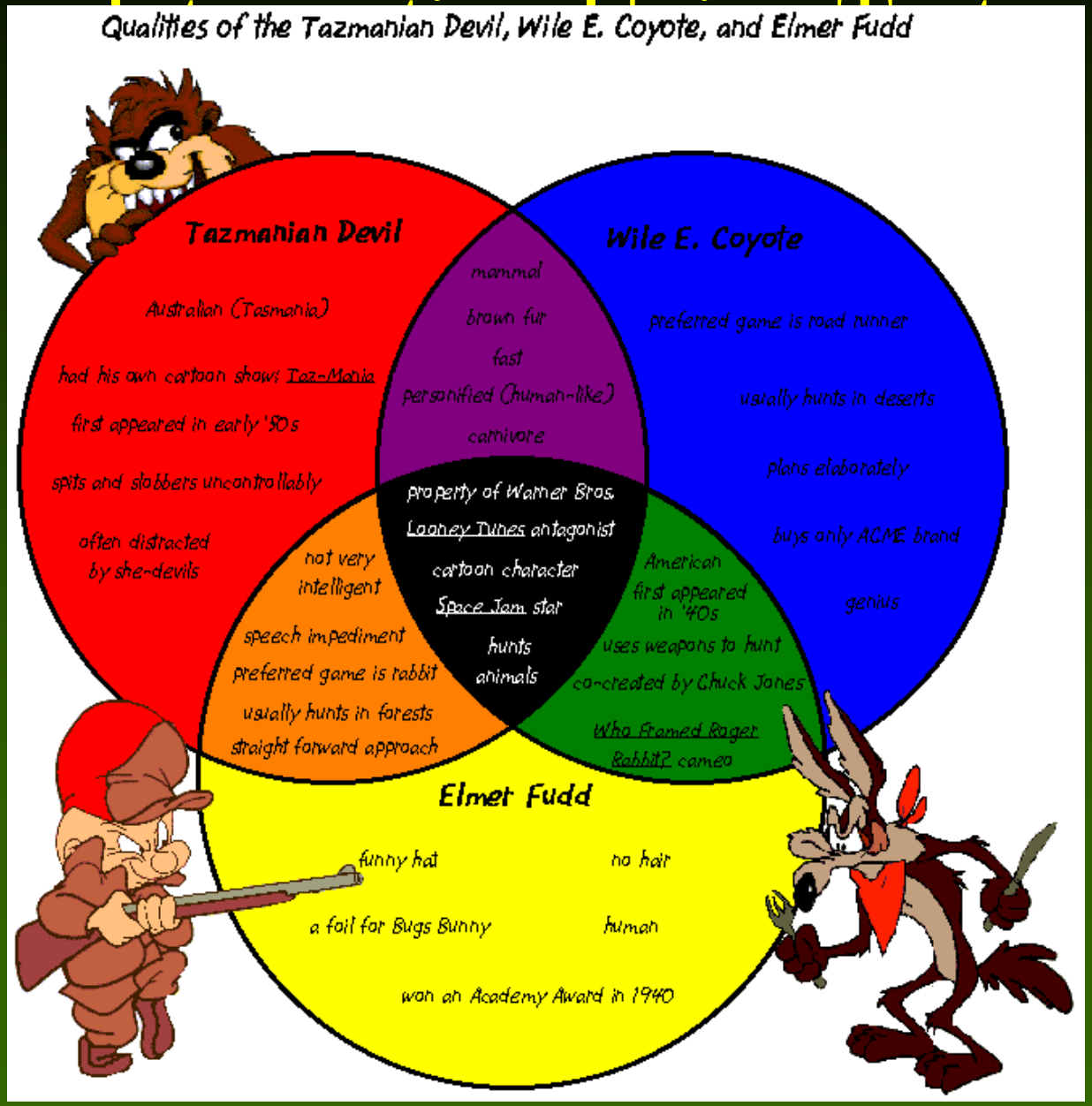
1. None of the genes in family c are differentially expressed.
2. The proportion of genes in family c that are differentially expressed is equal to the proportion of genes in the remainder of the genome that are differentially expressed.
3. The correlation matrix among the expression levels of the genes 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.

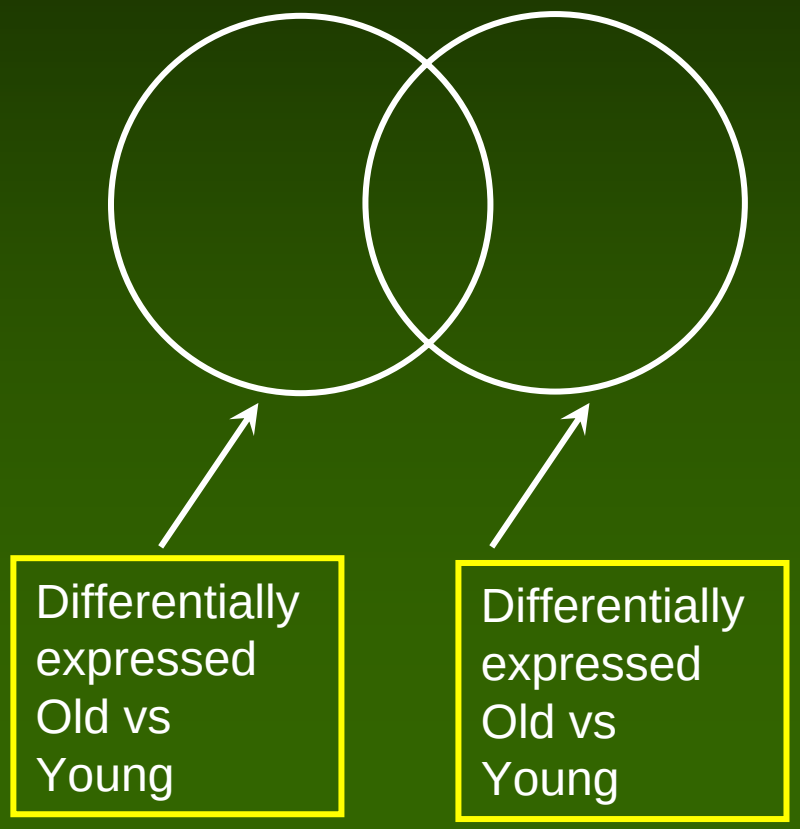
Global/Meta Analysis

Biological Process				
Function Name	*Total*	*P-Value*	*FDR*	*Bonferroni*
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


Qualities of the Tazmanian Devil, Wile E. Coyote, and Elmer Fudd
Tazmanian Devil
Australian (Tasmania)
had his own cartoon show: Taz-Mania
first appeared in early '90s
spits and slobbers uncontrollably
often distracted by she-devils
Wile E. Coyote
preferred game is road runner
usually hunts in deserts
plans elaborately
buys only ACME brand
genius
mammal
brown fur
fast
personified (human-like)
carnivore
property of Warner Bros.
Looney Tunes antagonist
cartoon character
Space Jam star
hunts animals
not very intelligent
speech impediment
preferred game is rabbit
usually hunts in forests
straight forward approach
American
first appeared in '40s
uses weapons to hunt
co-created by Chuck Jones
Who Framed Roger Rabbit? cameo
Elmer Fudd
funny hat
no hair
a foil for Bugs Bunny
human
won an Academy Award in 1940

Kyng KJ, May A, Kolvraa S, Bohr VA. Gene expression profiling in Werner syndrome closely resembles that of normal aging. Proc Natl Acad Sci U S A. 2003 Oct 14;100(21):12259-64.

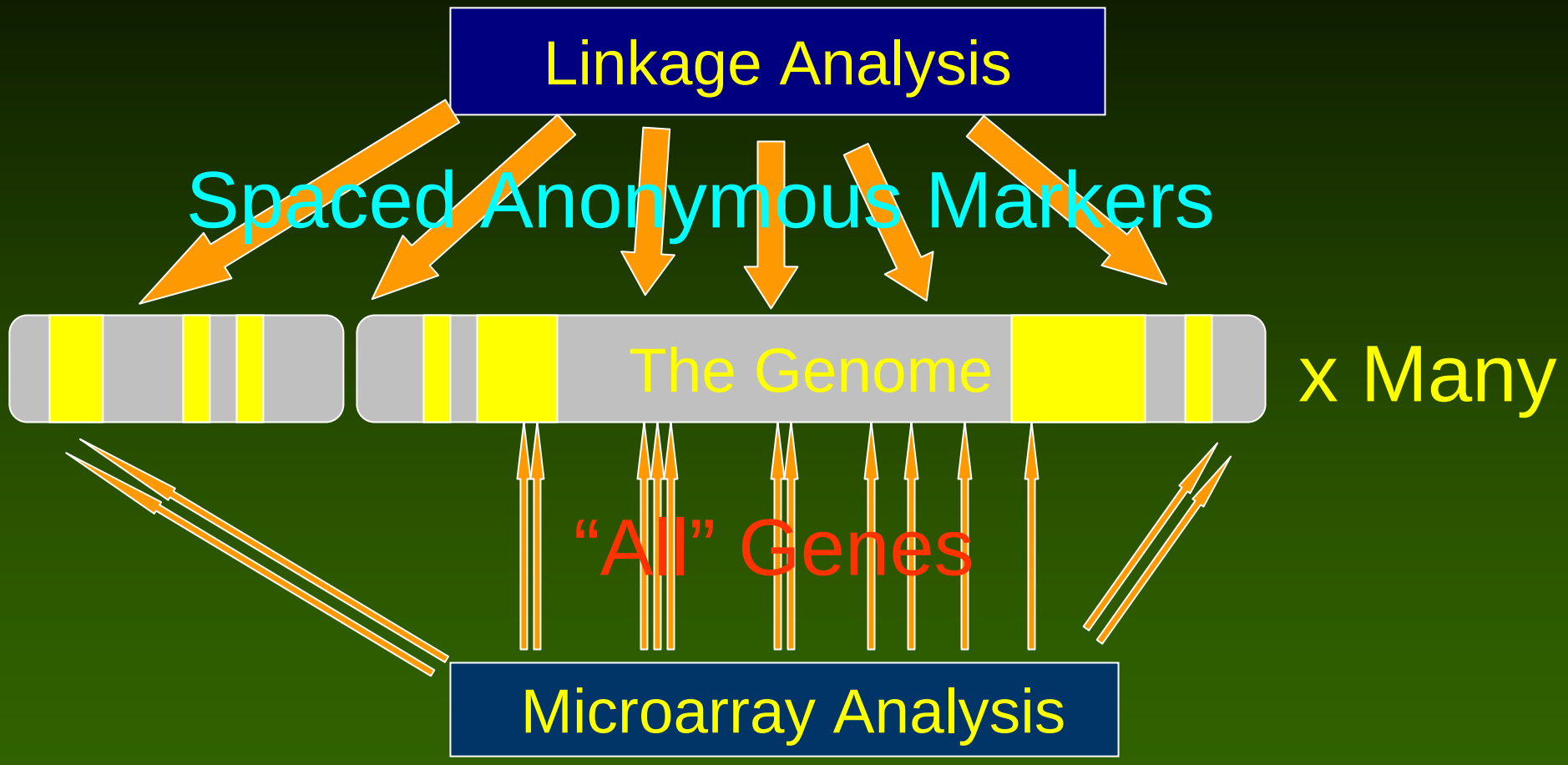
"Transcription alterations in WS were strikingly similar to those in normal aging: 91% of annotated genes displayed similar expression changes in WS and in normal aging, 3% were unique to WS, and 6% were unique to normal aging. "



Yet, by chance alone, (A-B) will generally be correlated with (A-C). Simulating their data as closely as possible suggest a 25% overlap by chance alone.

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.

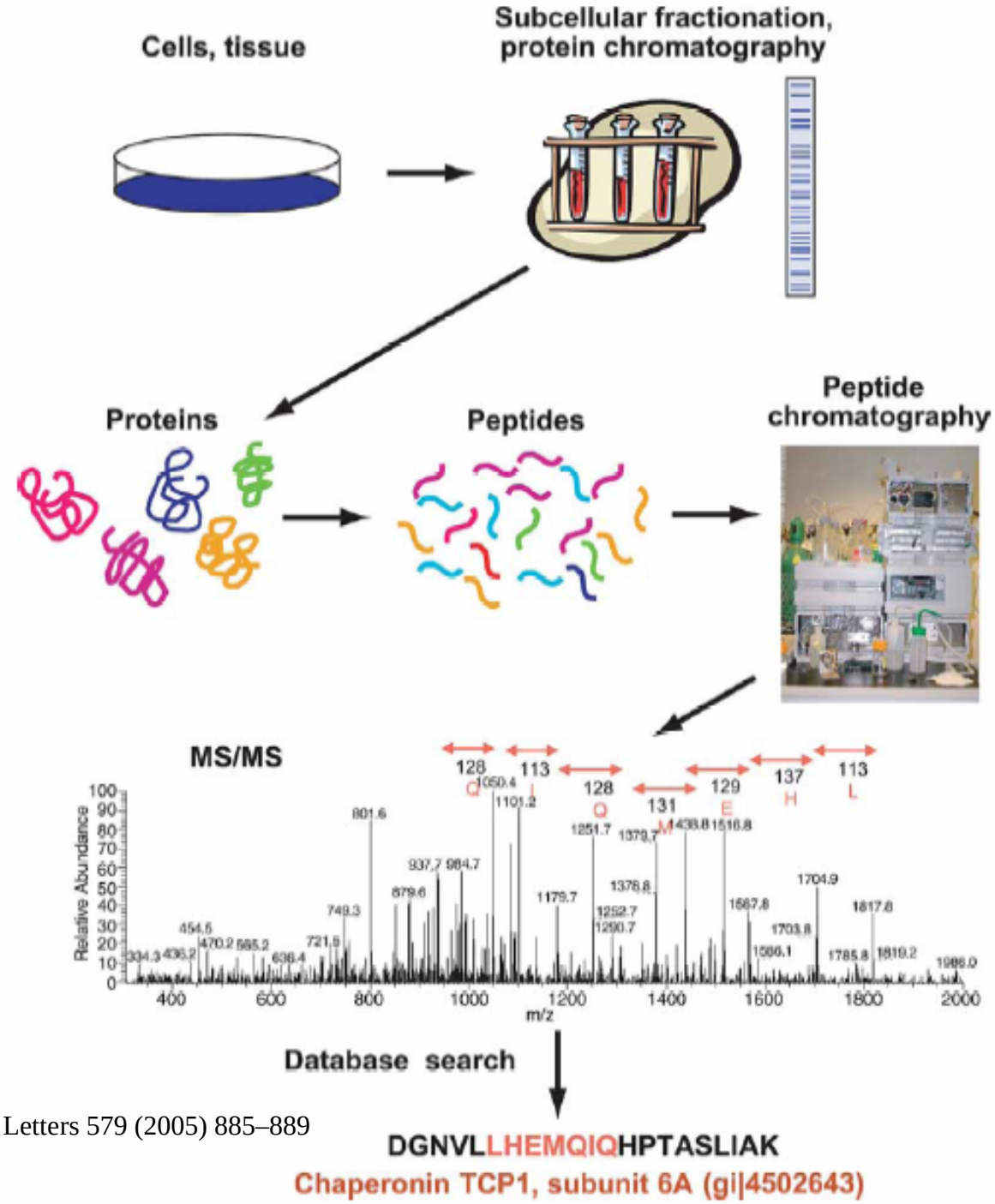

Linkage Analysis
Spaced Anonymous Markers
The Genome
x Many
"All" Genes
Microarray Analysis

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.

End of Part 1

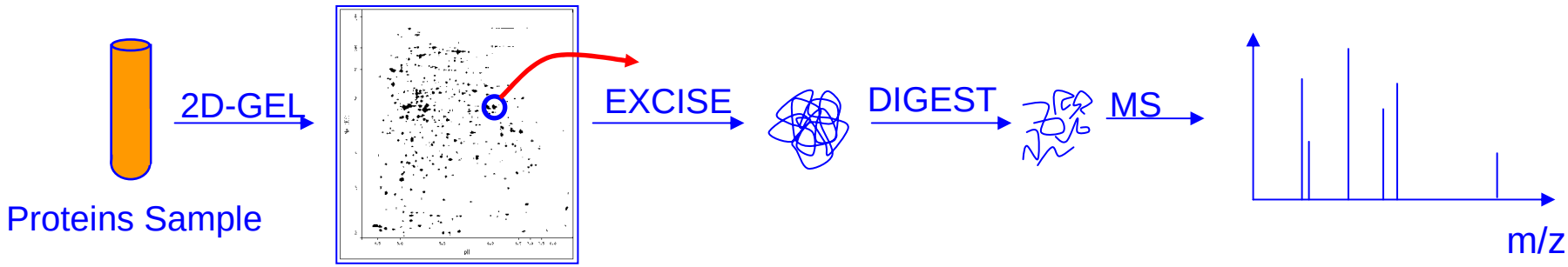
Statistical Analysis of Peptides



K.A. Resing, N.G. Ahn / FEBS Letters 579 (2005) 885–889

How to use MS for protein identification

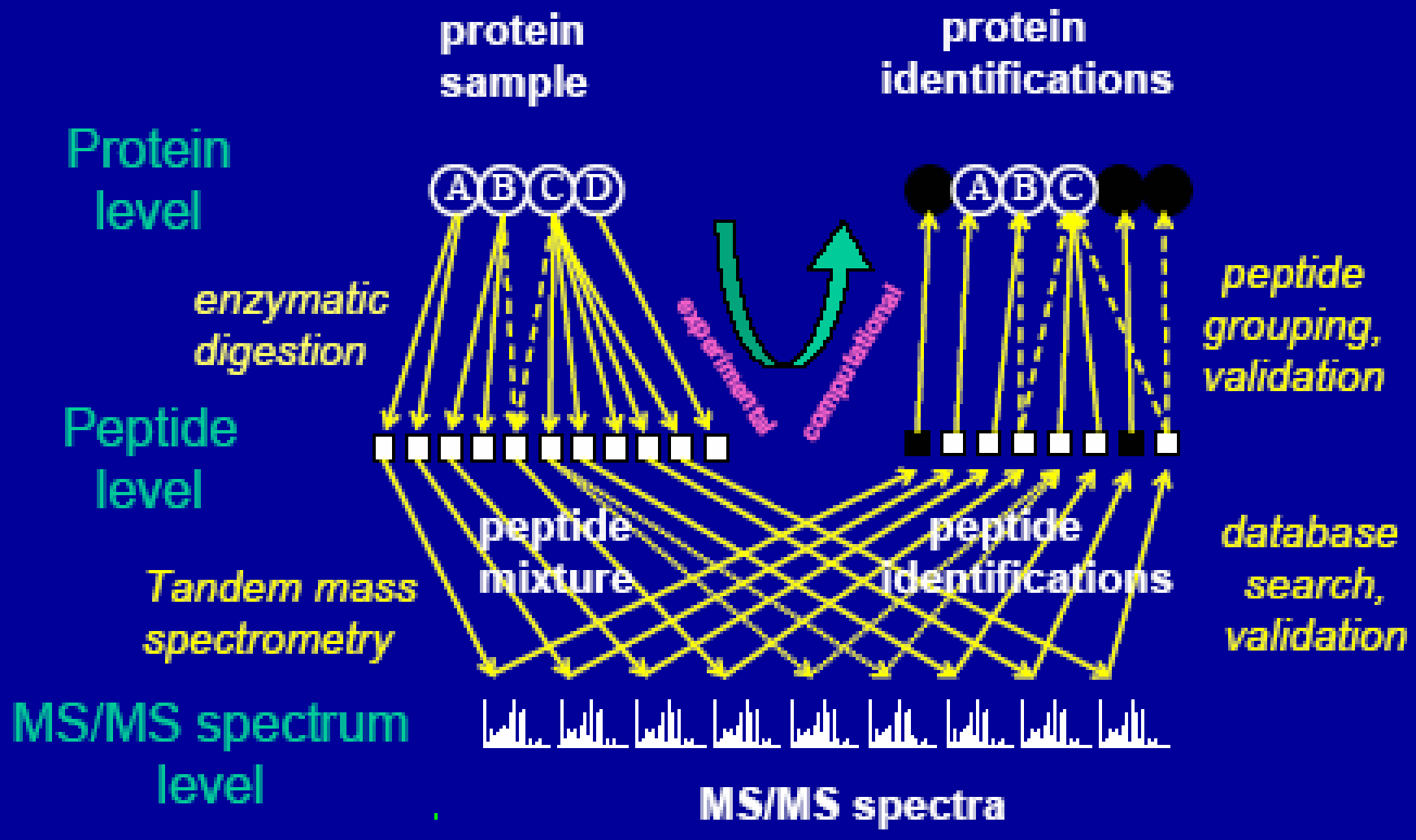
Peptide mass fingerprinting



Example: peaks at m/z 333, 336, 406, 448, 462, 889
The only protein in the database that would produce these peaks is
MALK|CGIR|GGSRPFLR|ATSK|ASR|SDD

- The exact protein needs to be in the database
- Works only with single protein fragmentations

Shotgun Protein Identification

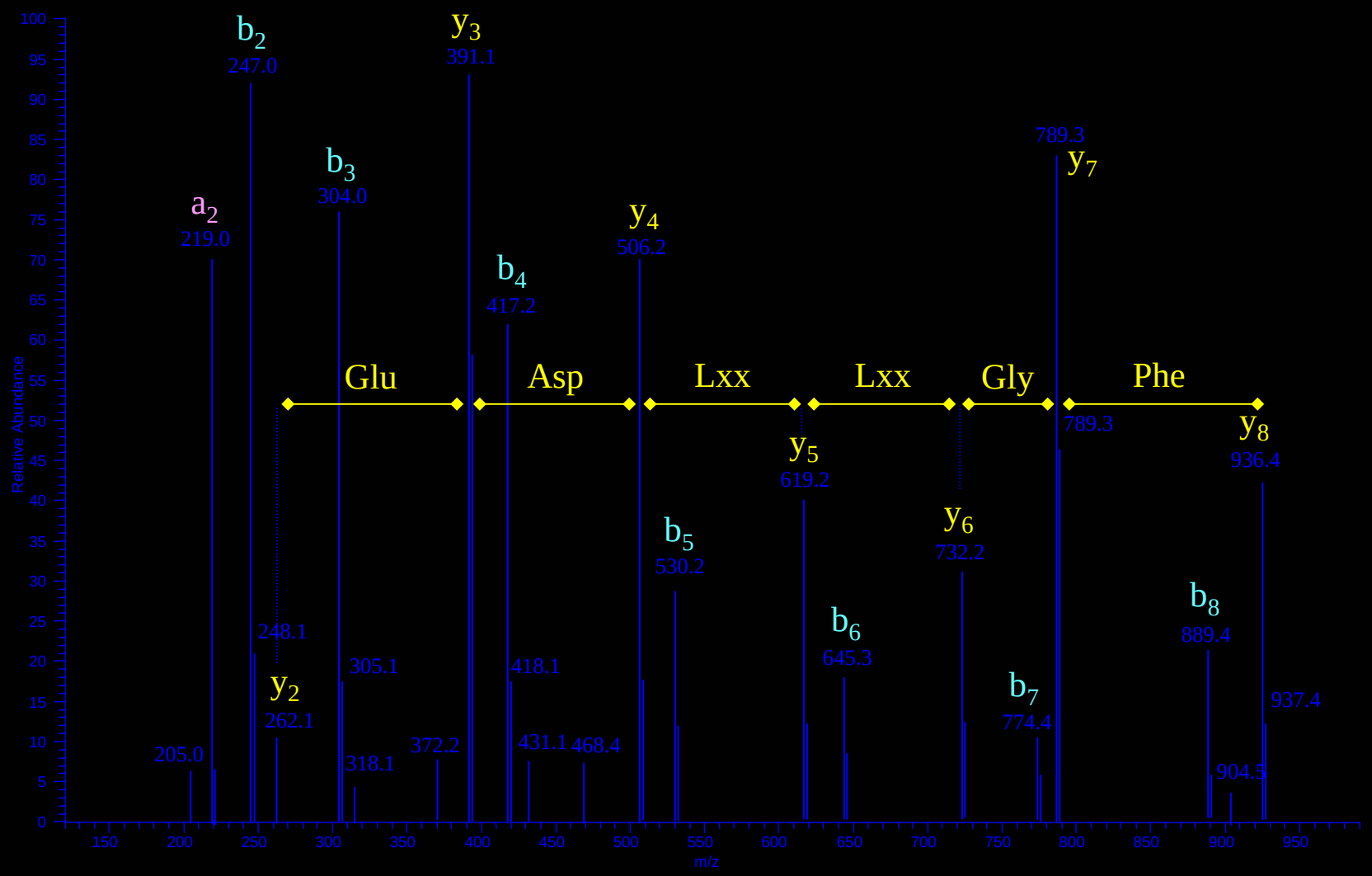


From Alexey Nesvizhskii

Tryptic fragment: Val Phe Gly Lxx Lxx Asp Glu Asp Lys

y_8 y_7 y_6 y_5 y_4 y_3 y_2

b_2 b_3 b_4 b_5 b_6 b_7 b_8

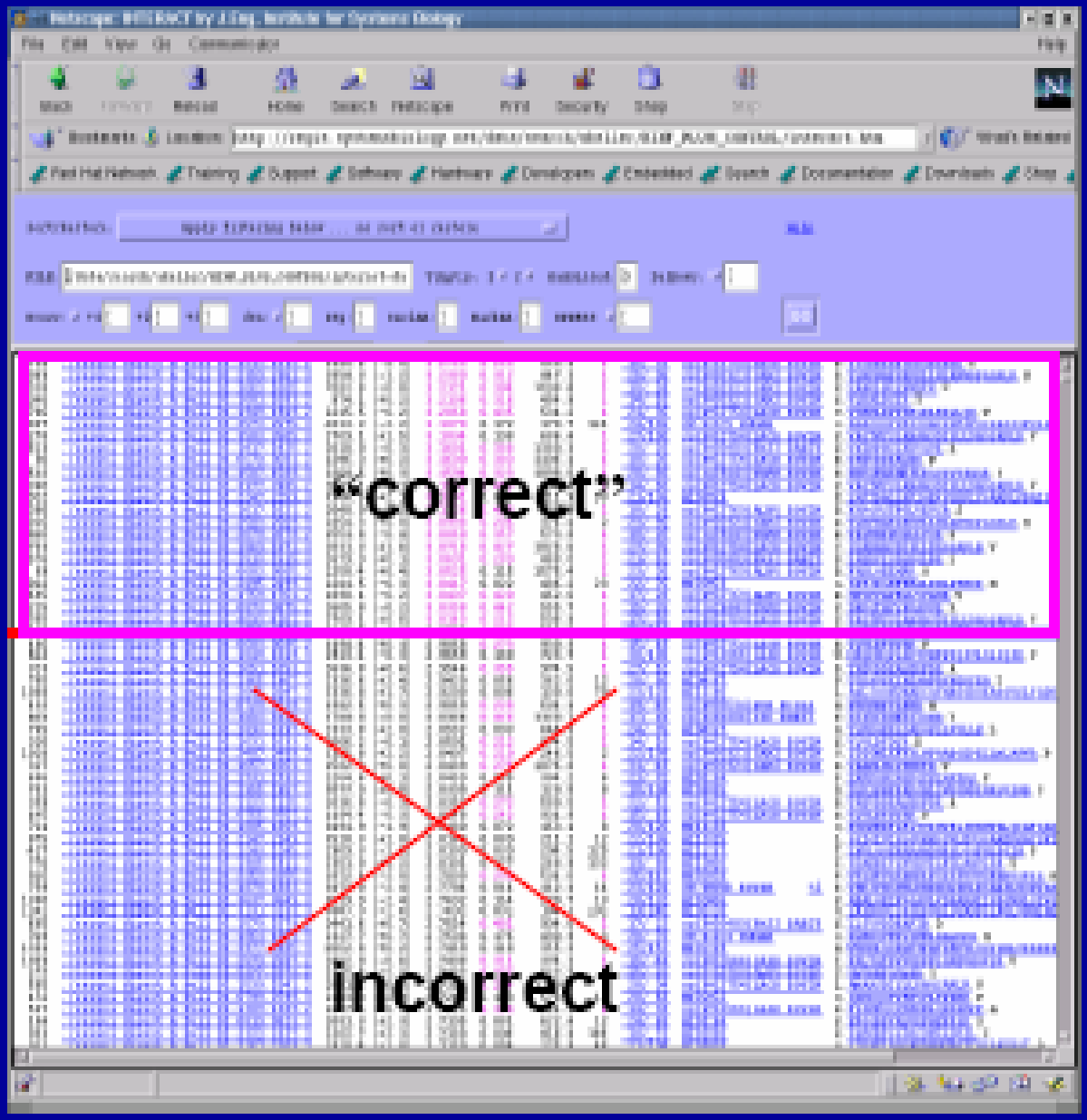


Example MS/MS spectrum

Interpretation of MS/MS data

- Direct interpretation ("de novo sequencing")
  - spectrum must be of good quality
  - the only identification method if the spectrum is not in the database
  - can give useful information (partial sequence) for database search
- General approach for database searching:
  - extract from the database all peptides that have the same mass as the precursor ion of the uninterpreted spectrum
  - compare each of them them to the uninterpreted spectrum
  - select the peptide that is most likely to have produced the observed data
- MASCOT:
  - simple probabilistic model
  - calculate the probability that a peptide could have produced the given spectrum by chance

Threshold Model



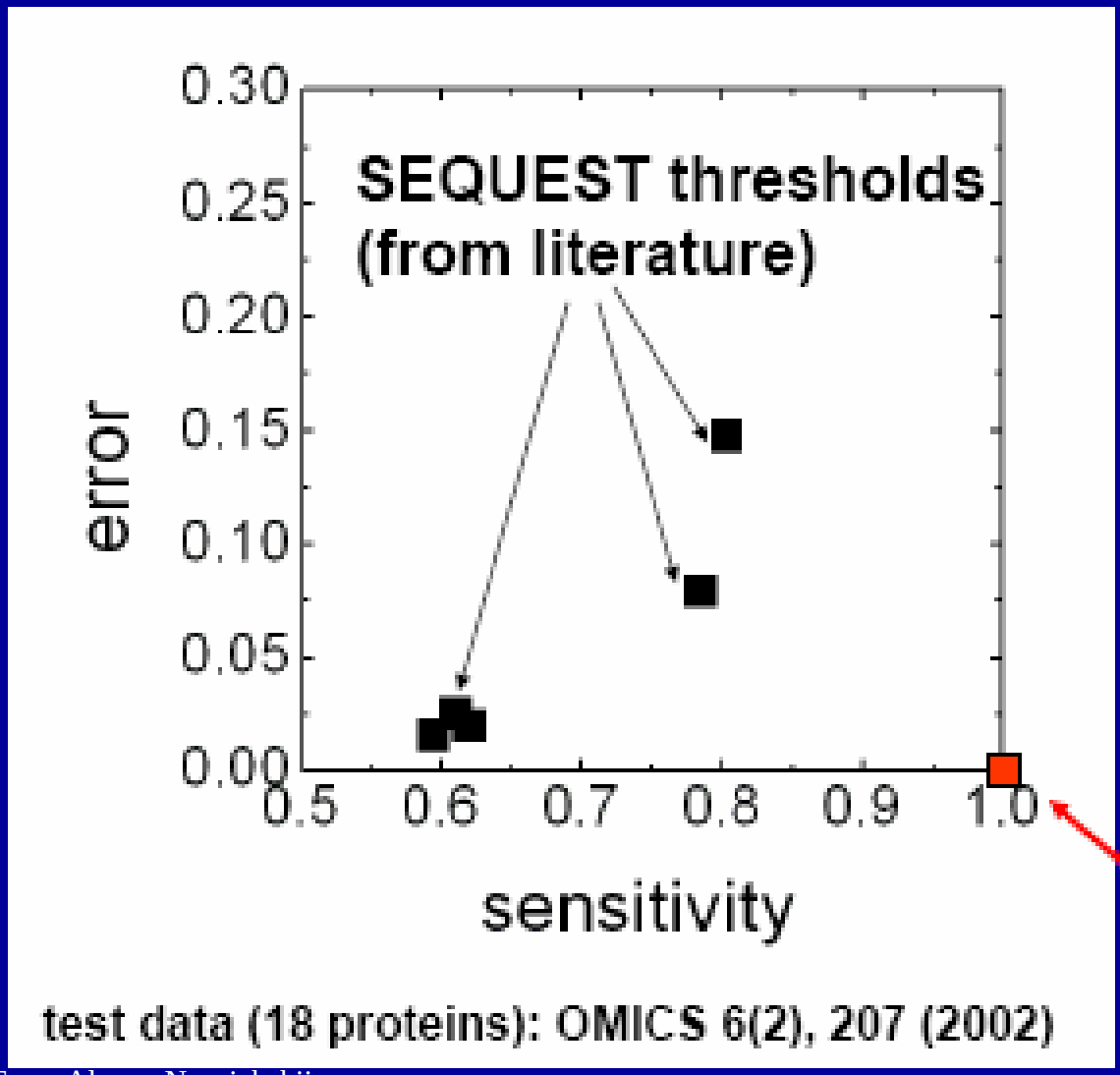
sort by search score

threshold

SEQUEST:
Xcorr > 2.0
ΔC_n > 0.1

MASCOT:
Ion Score > 30

Threshold Model: Bad Discrimination and Inconsistency



Sensitivity: fraction of all correct results passing filter

Error Rate: fraction of all results passing filter that are incorrect

Ideal Spot

From Alexey Nesvizhskii

Difficulties in Interpreting Peptide Identifications based on MS/MS

Applies to both SEQUEST and Mascot (as it is used in practice) and, to large degree, to more recent tools

- No 'useful' measures of confidence

(Mascot: 'identity threshold' guideline is not practical and rarely used)

- Different criteria used to filter data
- Unknown and variable false positive error rates

From Alexey Nesvizhskii

Just as assignment of quality scores to each base in DNA sequencing was essential for the genome sequencing programs, statistical models for estimating the accuracy of peptide and protein identifications are crucial for the success of high throughput proteomics

From Alexey Nesvizhskii

Statistical Validation

- p-values or expectation values

  used, e.g., in sequence similarity searching

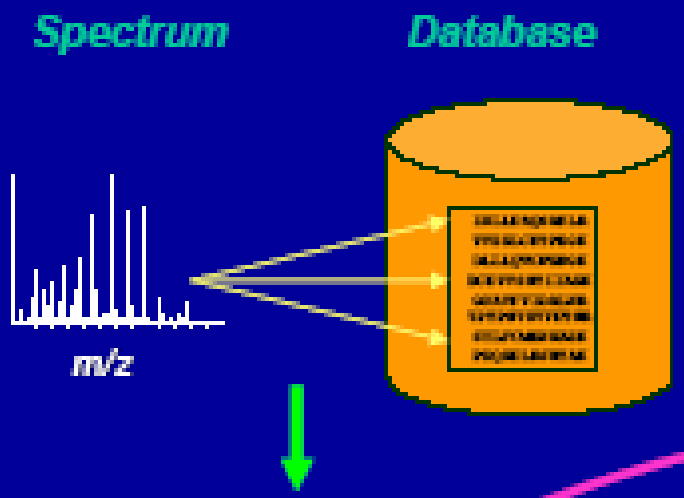
- Probabilities (Bayes)

  based on the ratio of two distributions (correct and incorrect) derived from the data (entire dataset)

  used, e.g., in information retrieval (relevant vs. non-relevant documents)

From Alexey Nesvizhskii

Expectation Values (empirical model)



$$p(s_m) = \sum_{s \geq s_m} P(s)$$

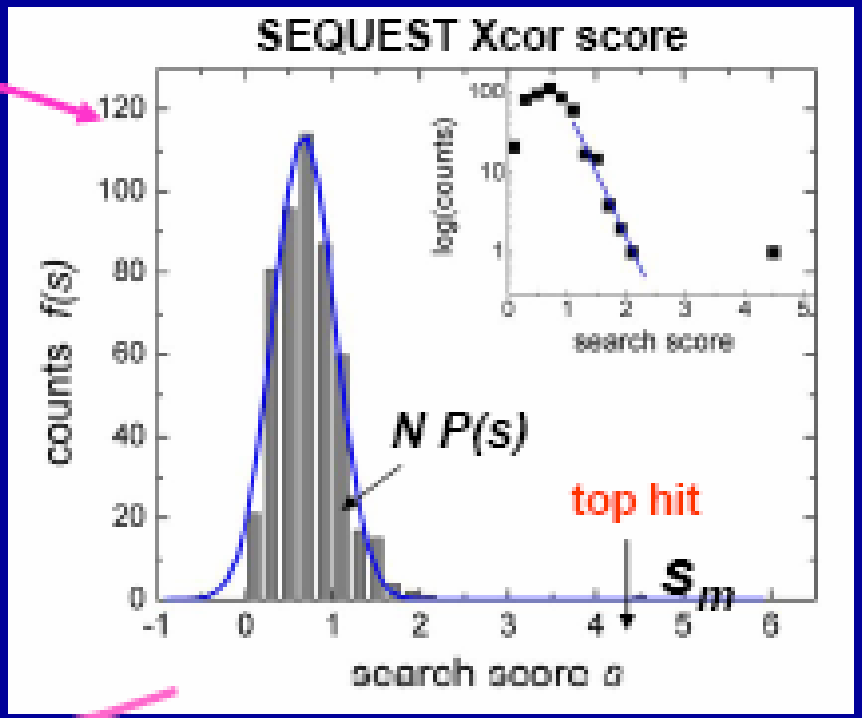
probability to get score $s \geq s_m$ by chance

$$E(s_m) = N \sum_{s \geq s_m} P(s)$$

expected number of random matches with $s \geq s_m$

Rank	Peptide	Score	p
1	ISLLDAQSAPLR	4.5	10^{-5}
2	VVEELCTPEGK	2.1	0.11
3	DLLLQWCWENGK	2.0	0.23
4	ECDVVSNTIIAEK	1.9	0.52
5	GDAVFVIDALNR	1.7	0.72
6	VPTPNVSVVTNR	1.6	0.86
7	SYLFCMEAEK	1.6	0.86
8	PEQSDLRSWTAK	1.5	0.94
	...		...

N peptides

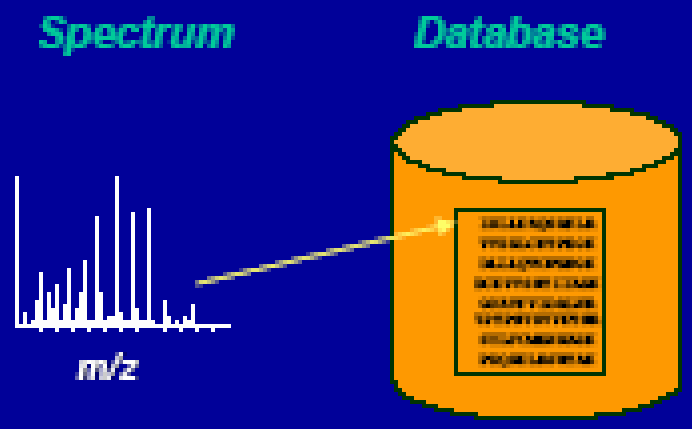


Fenyo & Beavis *Anal. Chem.* (2003)

From Alexey Nesvizhskii

Expectation Values (explicit model)

Sadygov & Yates *Anal. Chem.* (2003)
Geer et al. *J. Proteome Res.* (2004)



$$P(s) = \frac{\mu^s}{s!}\exp(-\mu)$$

Poisson distribution

s: number of matched peaks

μ: function of mass tolerance,
number experimental peaks
number of calculated ions
mass, charge

probability to get score $s \geq s_m$ by chance

$$p(s_m) = \sum_{s \geq s_m} P(s)$$

Geer et al. (upper bound)

expected number of random matches with $s \geq s_m$

$$E(s_m) = N(1-(1-\sum_{s \geq s_m} P(s))^N) \approx N^2 \sum_{s \geq s_m} P(s)$$

$$E(s_m) = N \sum_{s \geq s_m} P(s)$$

Sadygov et al. (lower bound)

From Alexey Nesvizhskii

Expectation Values (or p-values): Limitations

1. P-values or E-values are not well suited for the analysis of large-scale datasets (do not allow estimation of error rates as a function of filtering threshold)

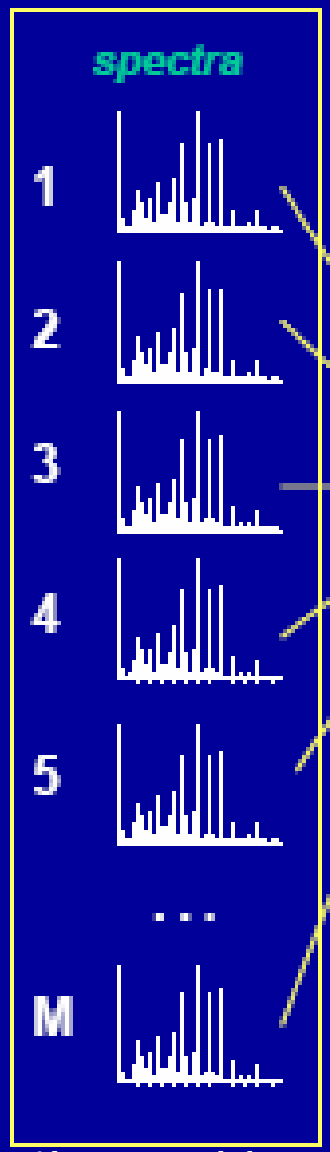
   see, e.g., recent papers by Tsibshirani and others on the subject of p -values vs. False Discovery Rate (FDR) approach

2. Difficult to take advantage of other useful information (e.g., number of missed cleavages, peptide retention time)

3. Need to compute protein probabilities by combining probabilities of peptides corresponding to the same protein. Whether peptide expectation values can be used for that purpose is not clear

From Alexey Nesvizhskii

Modeling Large-Scale Datasets



entire dataset, M spectra
(1 or more LC/MS/MS runs)

Spectrum	Peptide	Score
1	ISLLDAQSAPLR	4.5
2	VVEELCTPECK	3.9
3	DLLLQWCWENGK	1.2
4	ECDVVSNTIIAEK	0.9
5	GDAVFVIDALNR	3.6
	...	
M	SYLFCMEAEK	1.1

best match
to each spectrum

raw score
E value
p-value

From Alexey Nesvizhskii

Statistical Model for Computing Peptide Probabilities (PeptideProphet)

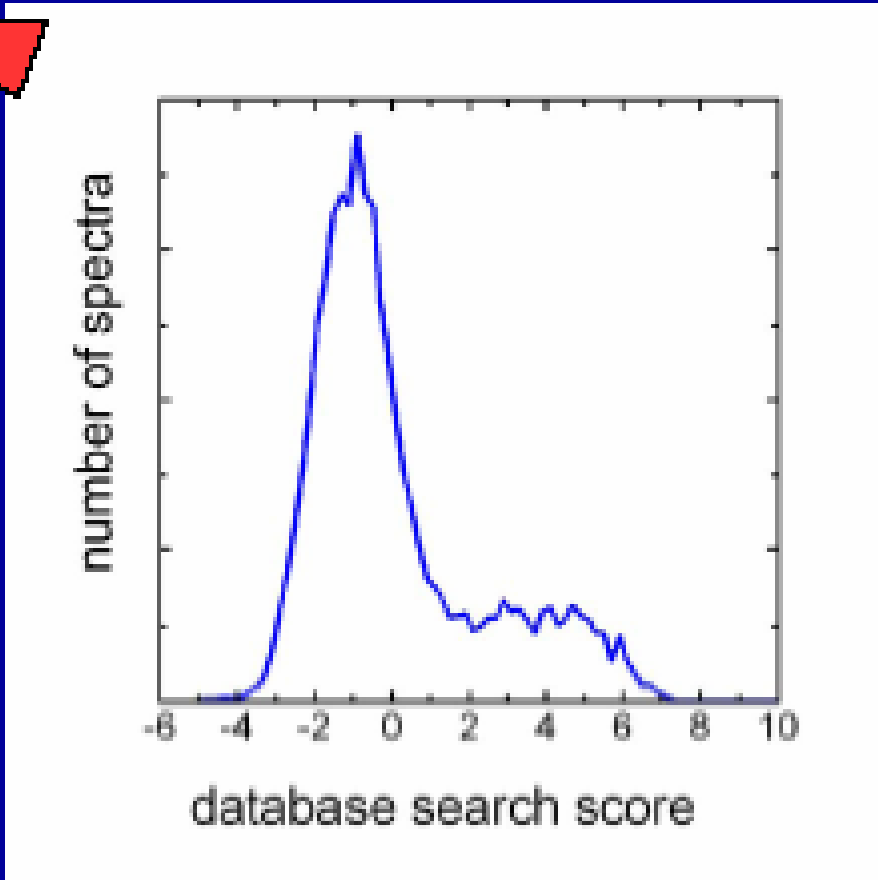
entire dataset:

Spectrum	Peptide	Score
1	ISLLDAQSAPLR	4.5
2	VVEELCTPEGK	3.9
3	DLLLQWCWENGK	1.2
4	ECDVVSNTIIAEK	0.9
5	GDAVFVIDALNR	3.6
	...	
M	SYLFCMEAEK	1.1

spectrum

best match

score



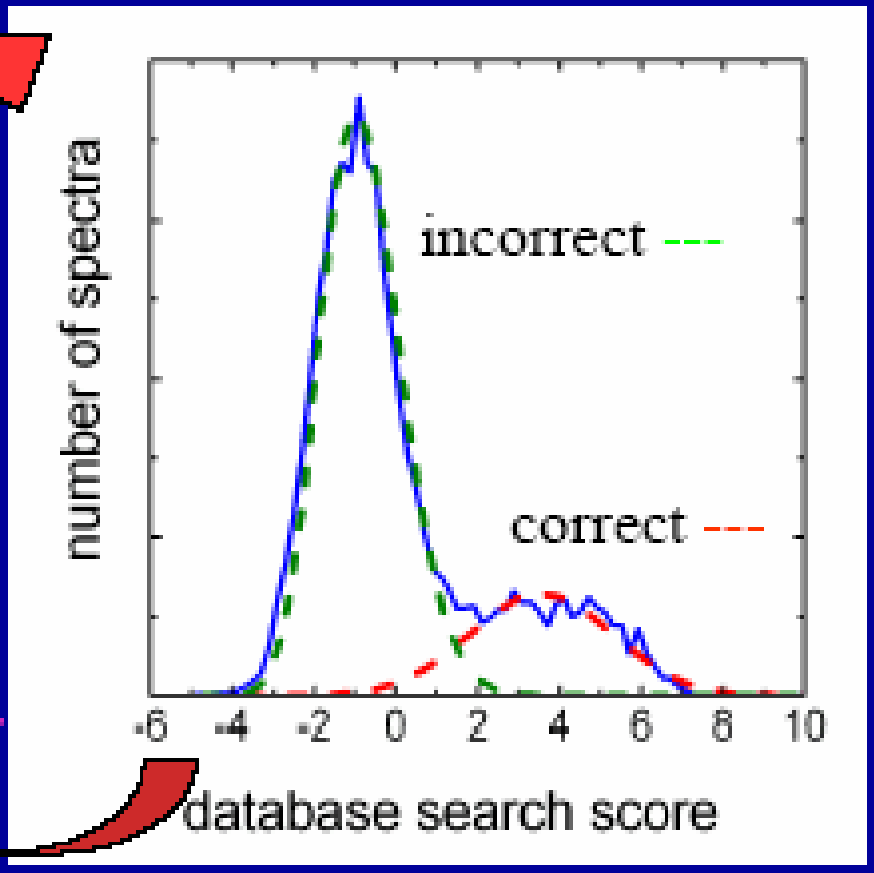
A. Keller, A.I. Nesvizhskii, E. Kolker, R. Aebersold *Anal. Chem.* 74, 5383 (2002)

From Alexey Nesvizhskii

Statistical Model for Computing Peptide Probabilities (PeptideProphet)

entire dataset:

Spectrum	Peptide	Score	probability
1	ISLLDAQSAPLR	4.5	1.00
2	VVEELCTPEGK	3.9	0.99
3	DLLLQWCWENGK	1.2	0.11
4	ECDVVSNTIIAEK	0.9	0.00
5	GDAVFVIDALNR	3.6	0.87
	...		...
M	SYLFCMEAEK	1.1	0.02

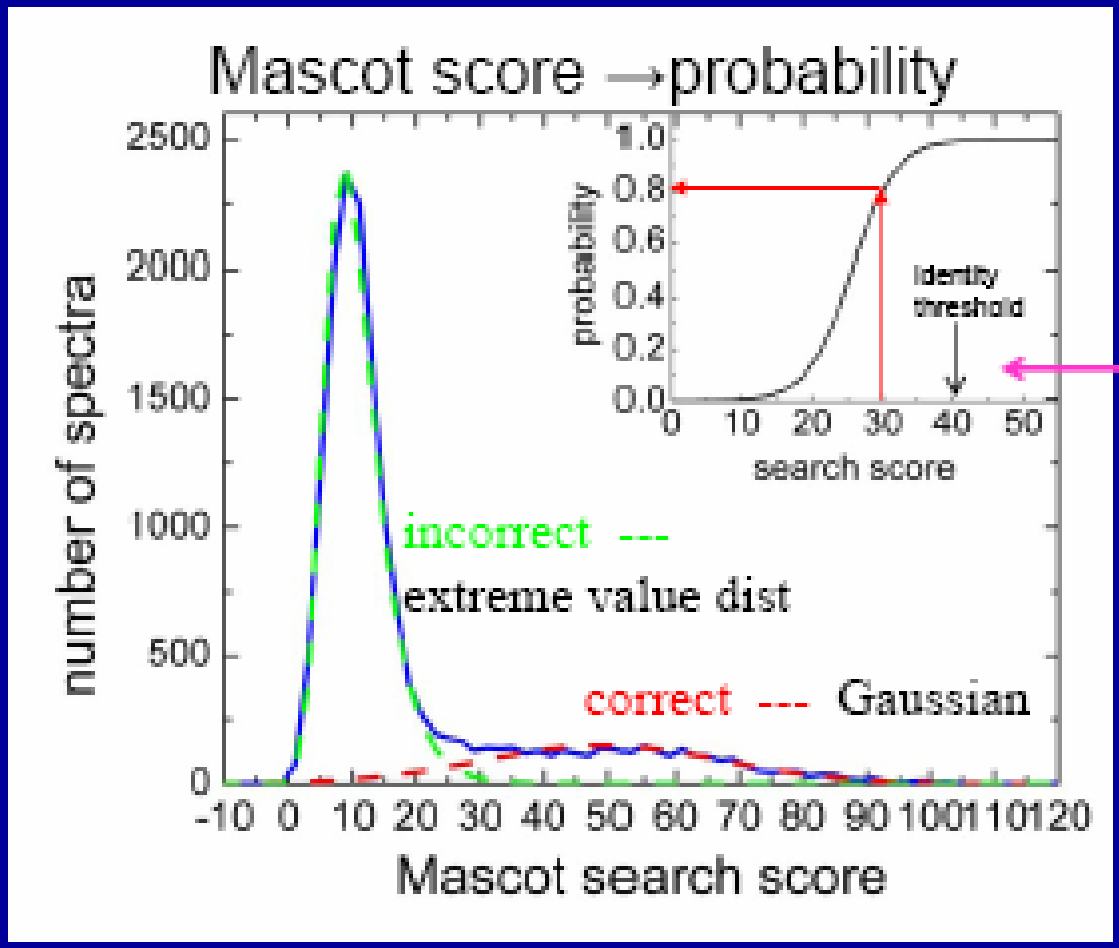


'unsupervised clustering'

EM mixture model algorithm learns the most likely distributions among correct and incorrect peptide assignments given the observed data

From Alexey Nesvizhskii

Illustration: Assigning Probabilities to Mascot Search Results



distributions are learned from the data

conversion of Mascot search scores into probabilities

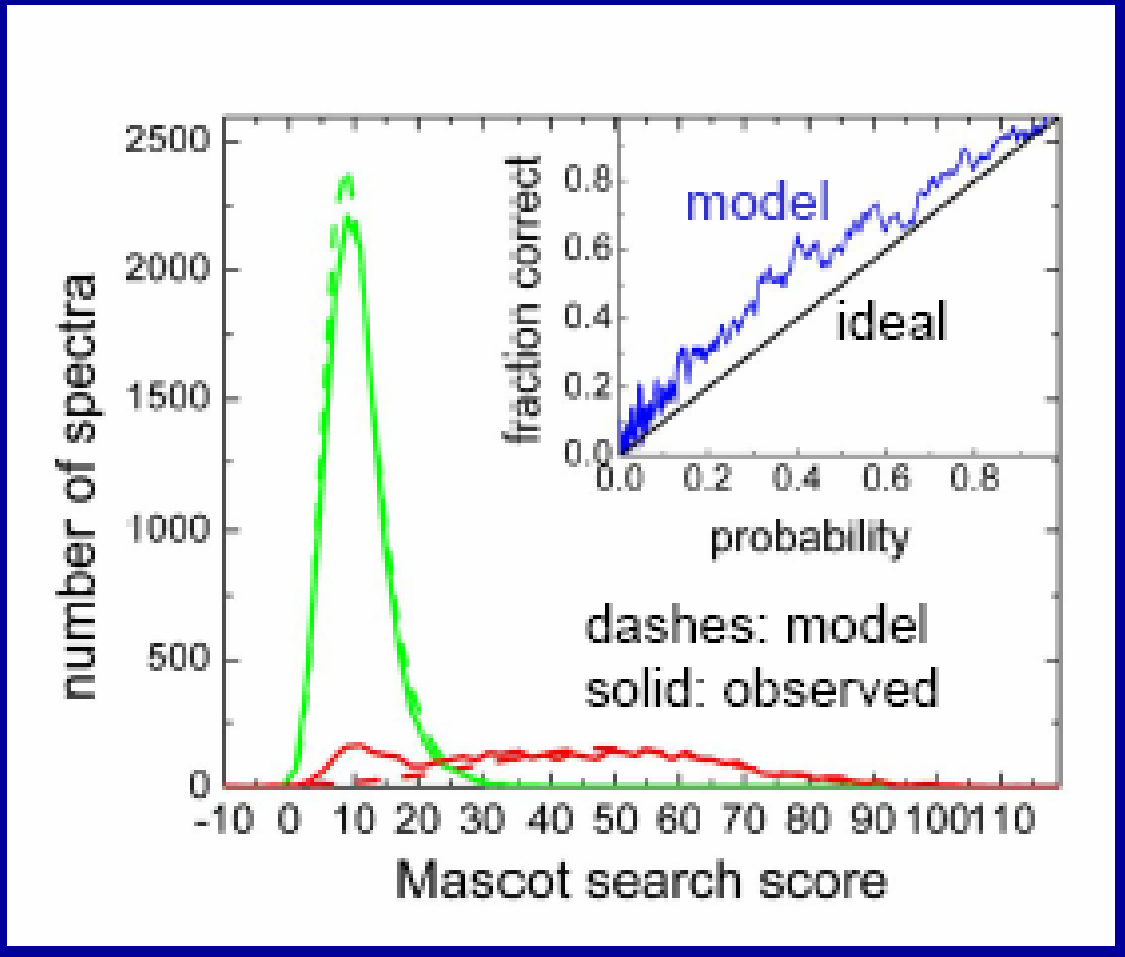
<u>To address a common misunderstanding</u>:

distribution parameters ARE NOT determined using a control dataset of 18 proteins, or any other training dataset for that matter. They are learned from each analyzed dataset anew using the EM mixture model algorithm

H. Influenzae, membrane fraction, 15 LC/MS/MS runs (~30,000 spectra)

From Alexey Nesvizhskii

Accuracy of Learned Distributions and Computed Probabilities



database searched:

size ratio: ~ 20:1

For those familiar with "reverse database search" approach:
This is an equivalent of appending 20 randomized databases of equal size.

Method is accurate

H. Influenzae, membrane fraction, 15 LC/MS/MS runs
~30,000 spectra

From Alexey Nesvizhskii

Question?