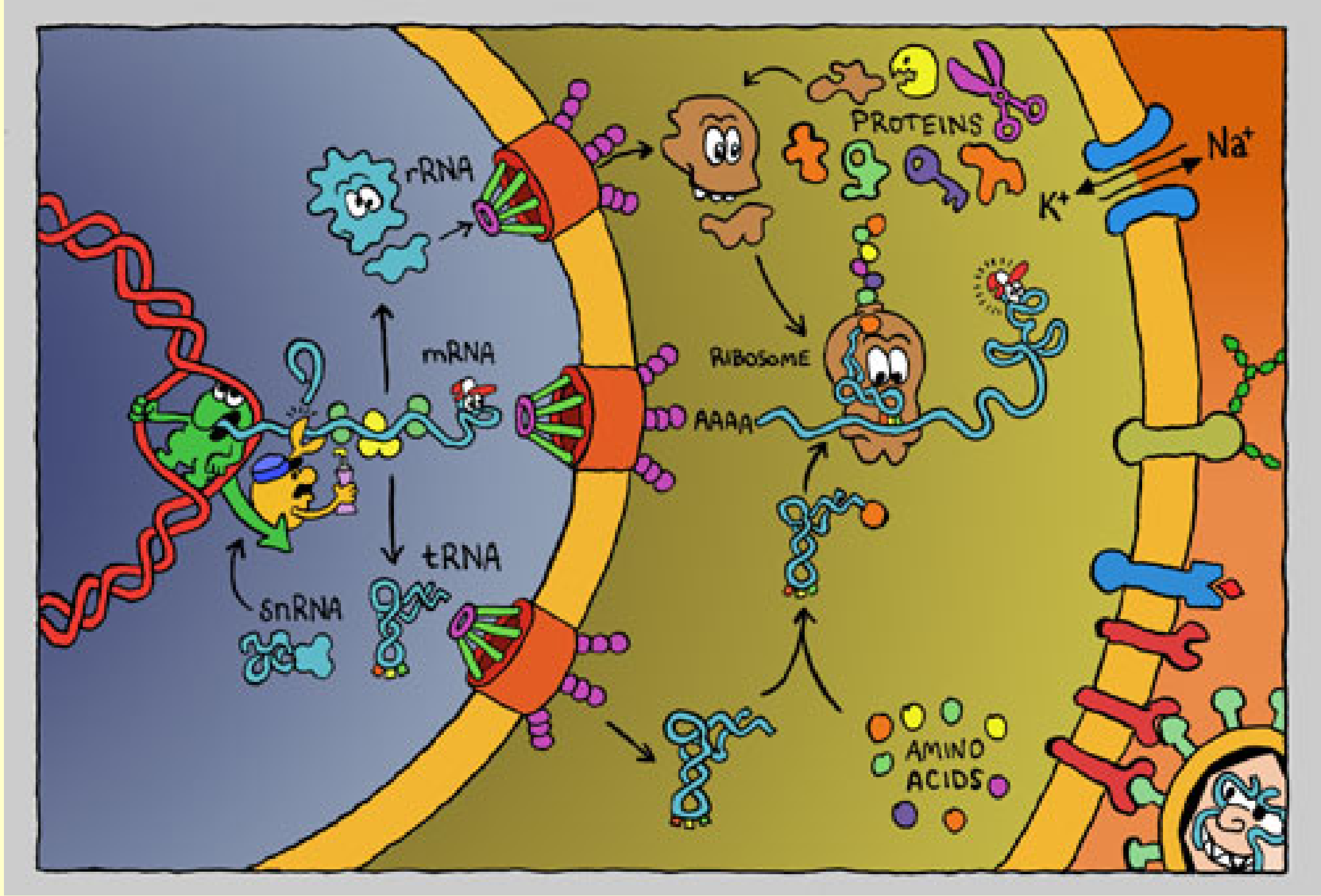
Improved Statistical Analysis of Data from Genetics, Microarrays, and Proteomics

Russ Wolfinger, PhD
Director of Scientific Discovery and Genomics
SAS Institute, Inc., Cary, NC

BASS XII
November 8, 2005

The Central Dogma of Molecular Biology Underlies Statistical Analysis



Drawn by Ebbe Sloth Andersen, http://130.225.13.7/dogma.html

Organize Data Types Along the Central Dogma

Genetic Marker Data – DNA, sequencing instruments, static

Transcript Abundance Data – RNA, microarrays, dynamic

Protein and Metabolite Abundance – peptides, gels and spectrometers, dynamic

Introductory Remarks

- Genomics data are mainstream now throughout discovery, pre-clin (e.g. tox) and Phases I-IV
- Driving major paradigm shifts towards personalized medicine, but major ROIs have been slow coming
- Analysis methods for genetics, transcriptomics, proteomics / metabalomics in different stages of maturity; have quite different scientific histories.
- Trend towards larger consortium-style projects with 500-1000+ subjects

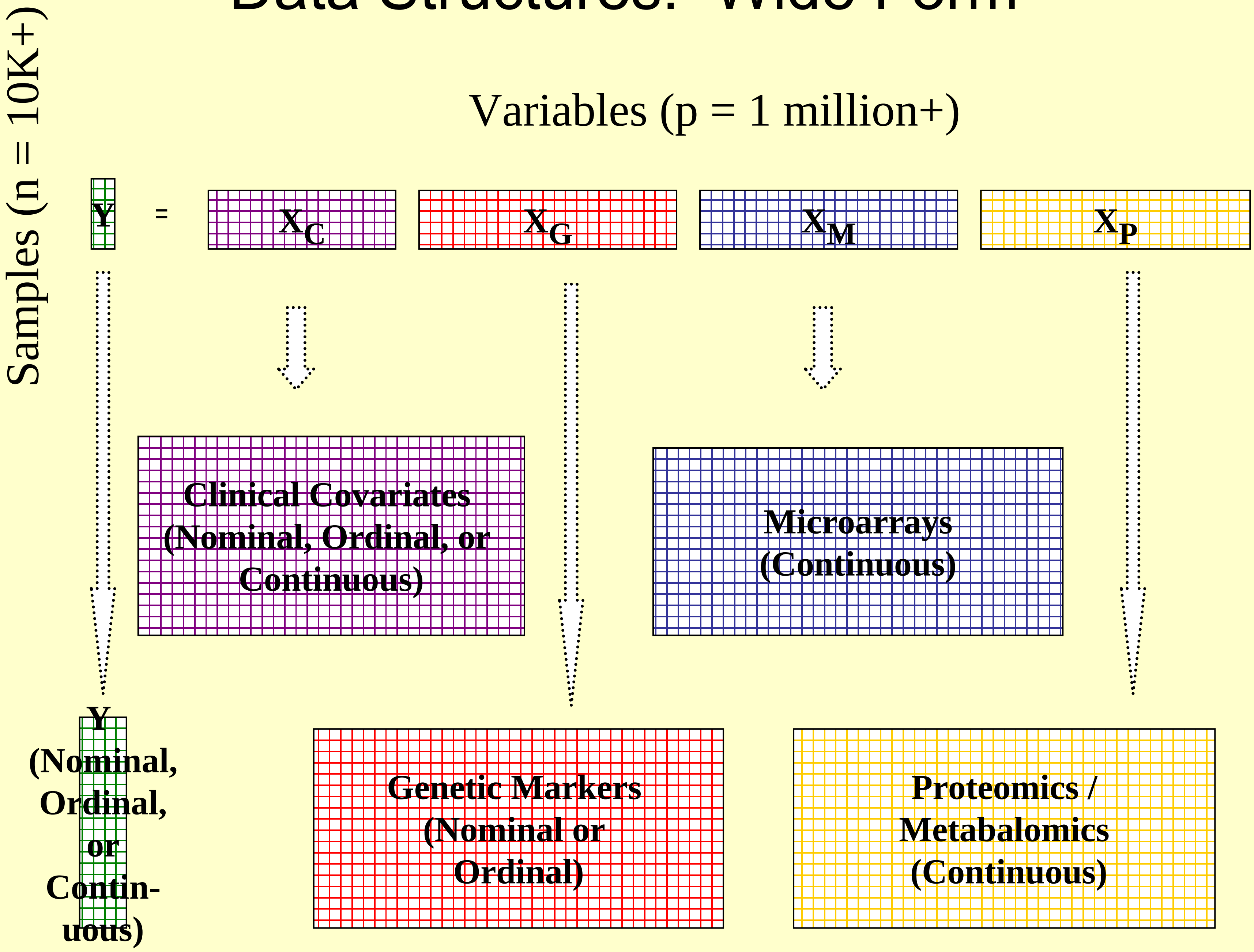
Introductory Remarks (continued)

- Tower of Babel: clinical biostatistics, discovery statistics, computer science, bioinformatics, biology, chemistry, medicine, PKPD, toxicology
- Huge number of new papers in stat and bioinformatics literature, try Google and Google Scholar
- Dire need for sound statistical reasoning, interpretation and communication

Ultimate Goals: Understanding and Prediction

- How do we get there? Numerous competing methods
- Definite risk of overfitting due to $n << p$
- Dimension reduction / variable selection is critical
- Honest cross-validation (both within and across studies) is essential for generalizability

Data Structures: Wide Form



Samples (n = 10K+)

Variables (p = 1 million+)

$Y = X_C \quad X_G \quad X_M \quad X_P$

Clinical Covariates (Nominal, Ordinal, or Continuous)

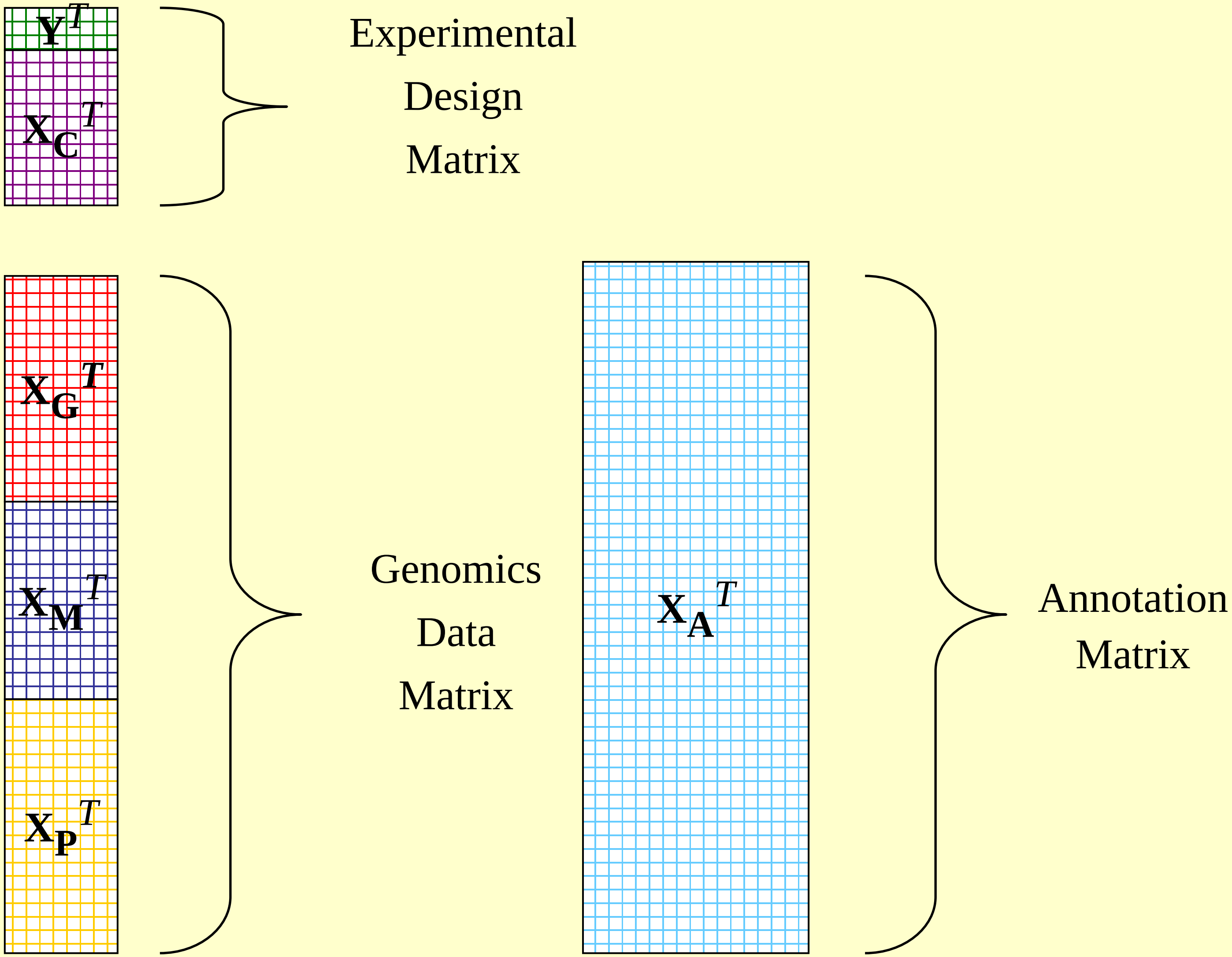
Microarrays (Continuous)

Y (Nominal, Ordinal, or Contin-uous)

Genetic Markers (Nominal or Ordinal)

Proteomics / Metabalomics (Continuous)

Data Structures: Tall Form



Experimental Design Matrix

Genomics Data Matrix

Annotation Matrix

Data Structures

- Tall form is well-suited for pre-processing, normalization, pattern discovery, and row-by-row modeling.
- Wide form is typical for data mining and keeps all variables in one file; allows data types
- Easy to transform from one to the other; both are useful
- Dimensions:
  - 10K+ samples
  - 1K+ clinical covariates
  - SNPs: Affy 500K SNP chip, theoretically 1M+ SNPs
  - RNA Expression: All-Exon Array with 1M+ features
  - Proteins: 100K+ peaks from LC-MS

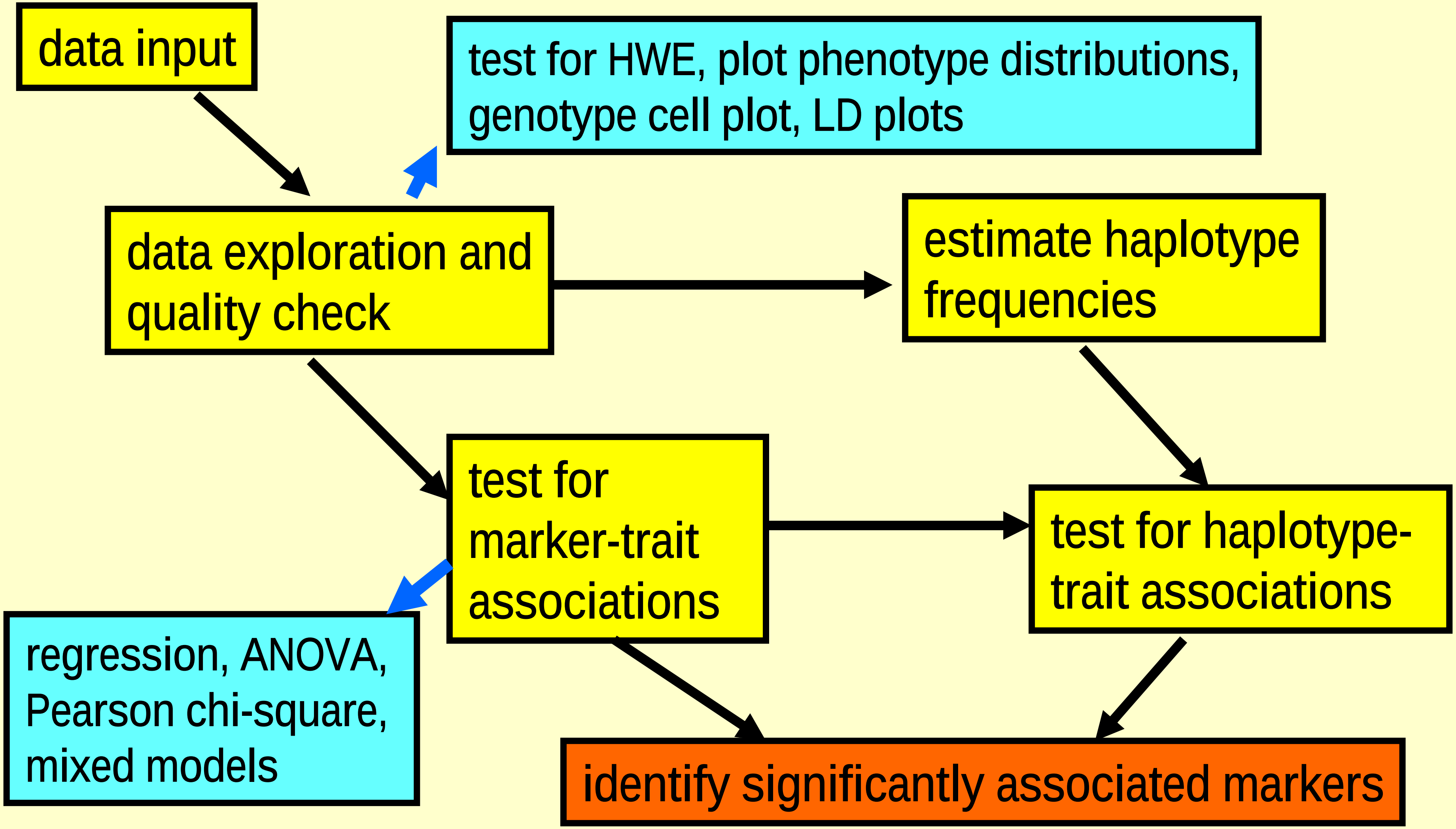
Genomics Data Analysis Challenges

- Understanding the Instrumentation
- Data Quality
- Missing Data
- Standardization / Normalization
- Transformations
- Confounding
- Association versus Causation
- n << p Prediction and Cross-Validation

Integrating Data from Clinical, Genetics, Microarrays, and Proteomics

- Pre-process and clean separately, then join by sample ID.
- Basic cross-correlations are a good place to start.
- Multivariate perspective is needed.
- Data mining methods are agnostic to combined predictor sets.

Genetic Marker Statistical Work Flow



data input

test for HWE, plot phenotype distributions, genotype cell plot, LD plots

data exploration and quality check

estimate haplotype frequencies

test for marker-trait associations

test for haplotype-trait associations

regression, ANOVA, Pearson chi-square, mixed models

identify significantly associated markers

Genetic Marker Association Tests

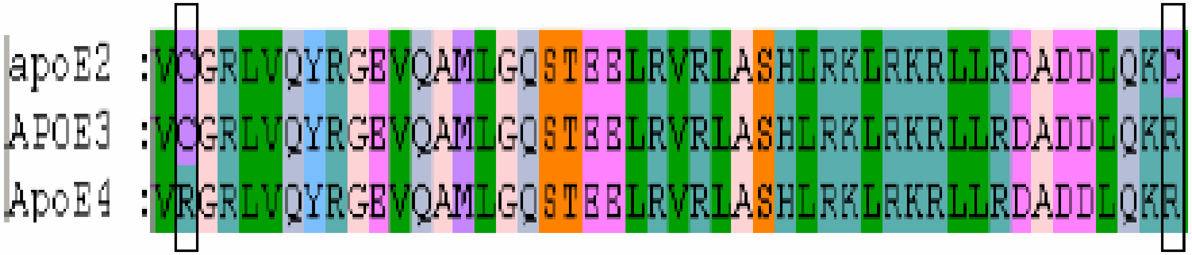
	Binary Trait	Quantitative Trait
Population	Case-Control Association Haplotype Estimation	Quantitative Trait Association Haplotype Trend Regression
Family	TDT	Quantitative TDT

Example: Alzheimer’s

Simulated mock genetic data based on 31 individuals – 9 control (Alz 0), 22 affected with Alzheimer's (Alz 1). Later we will subdivide the latter class into Incipient, Moderate, and Severe cases.

The data contain markers (SNPs) related to Alzheimer’s disease. One particular gene which has been found involved is APOE or apolipoprotein E .

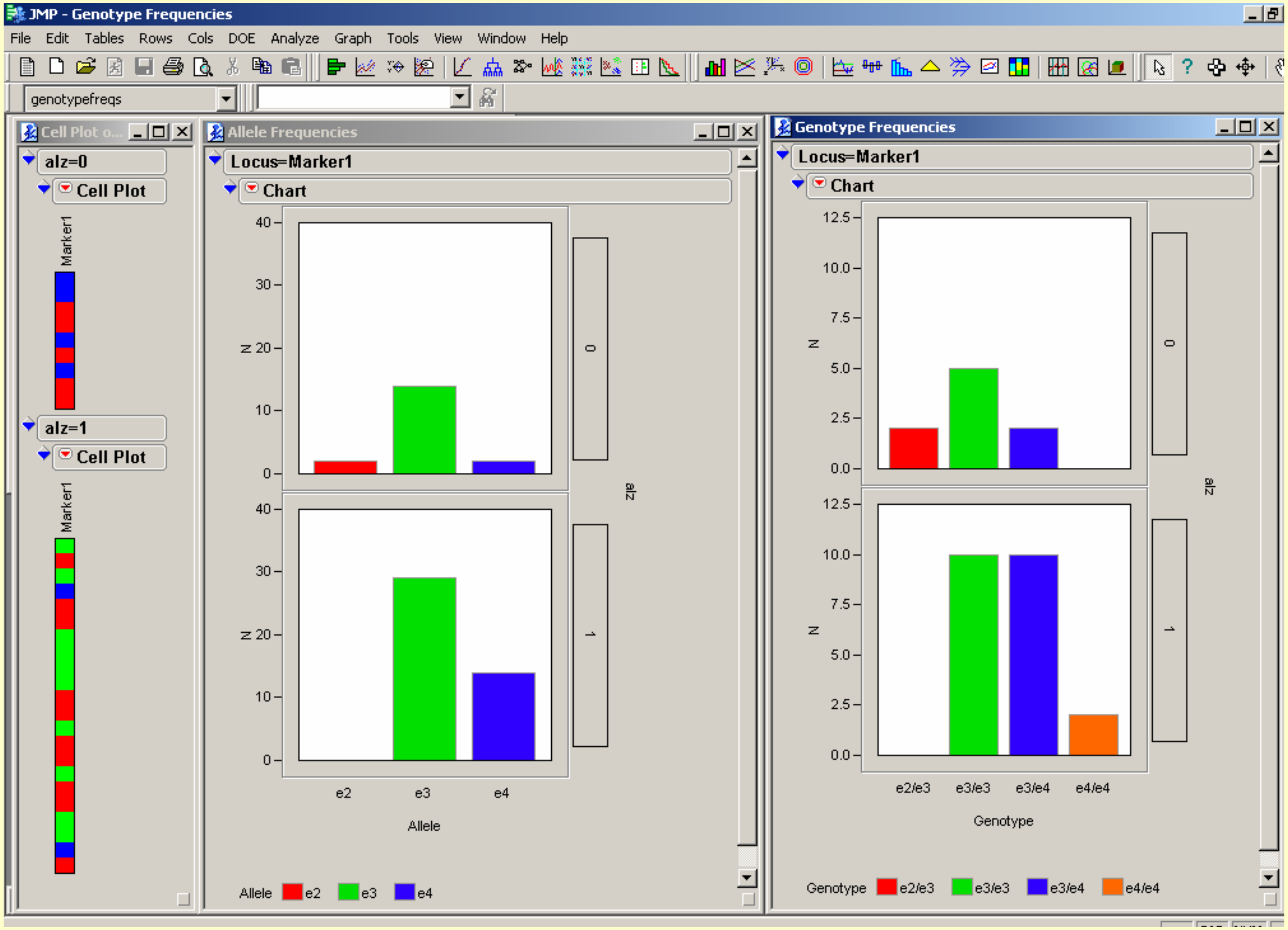
- Located on chromosome 19, ApoE is a gene that encodes for apolipoprotein E.
- The ApoE gene exists in 3 alleles, denoted e2, e3, and e4.
- ApoE4 is encoded by the e4 allele contains arginine (R) at positions 112 and 158.
- ApoE3 contains a cysteine (C) at position 112.
- ApoE2 contains cysteine (C) in both positions.



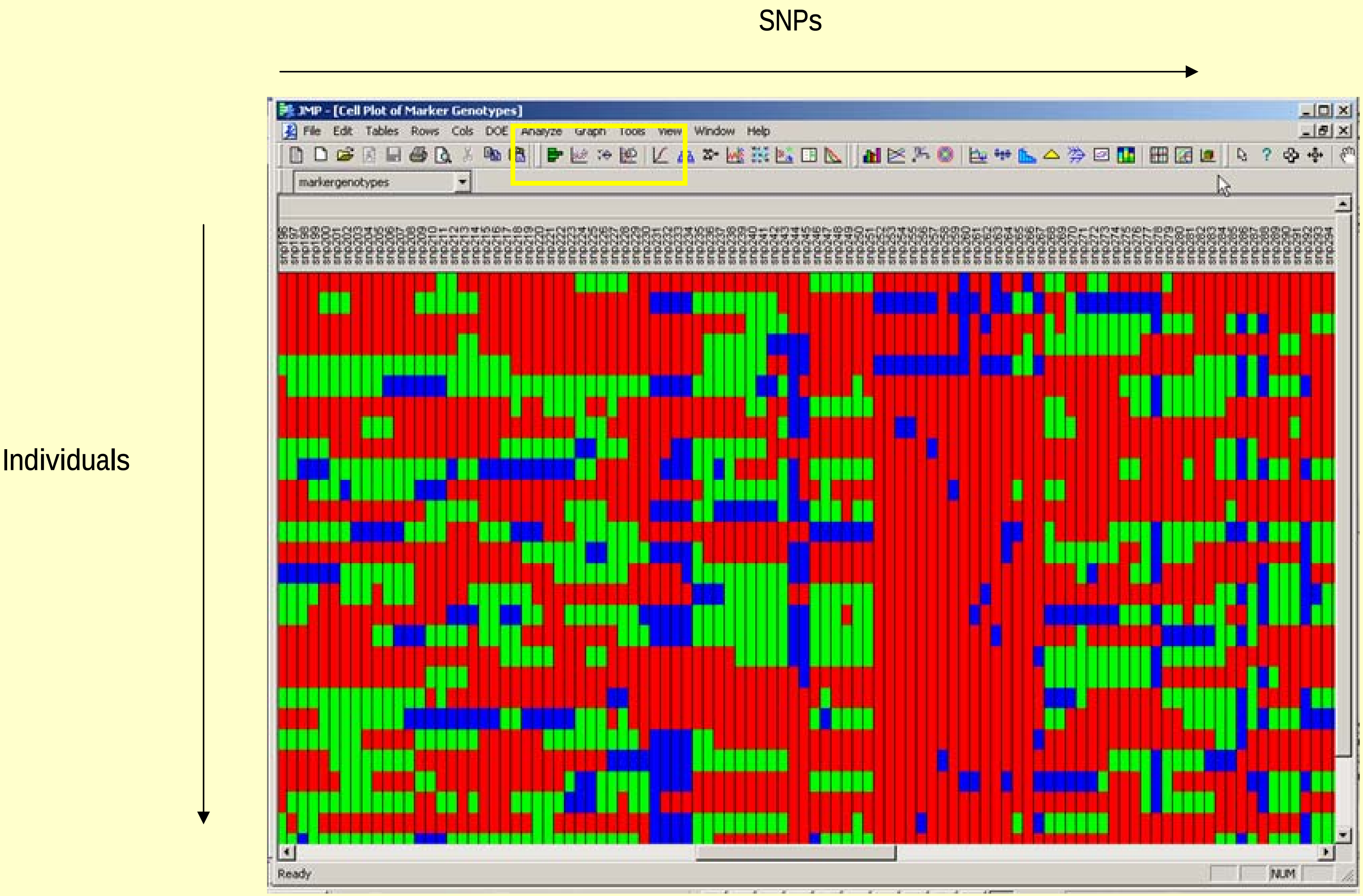
```
apoE2 :VCGRLVQYRGEVQAMLGQSTEELRVRLASHLRKLRKRLLRDADDLQKC
APOE3 :VCGRLVQYRGEVQAMLGQSTEELRVRLASHLRKLRKRLLRDADDLQKR
ApoE4 :VRGRLVQYRGEVQAMLGQSTEELRVRLASHLRKLRKRLLRDADDLQKR
```

- Six possible genotypes: e2/e2, e3/e3, e4/e4, e2/e3, e2/e4, e3/e4
- In Western populations, the e4/e4 genotype have the highest risk to develop the disease - 40 percent of all Alzheimer's patients have the e4 allele.
- The most common APOE genotype is e3/e3, which occurs in 40 - 90 percent of people. The e2 allele is rare, occurring in only 2 percent of the population.
- The e4/e4 genotype is found in only 1-3 percent of the Western population. However, the probability that the e4/e4 genotype will develop Alzheimer's disease is 60 percent

Marker Properties



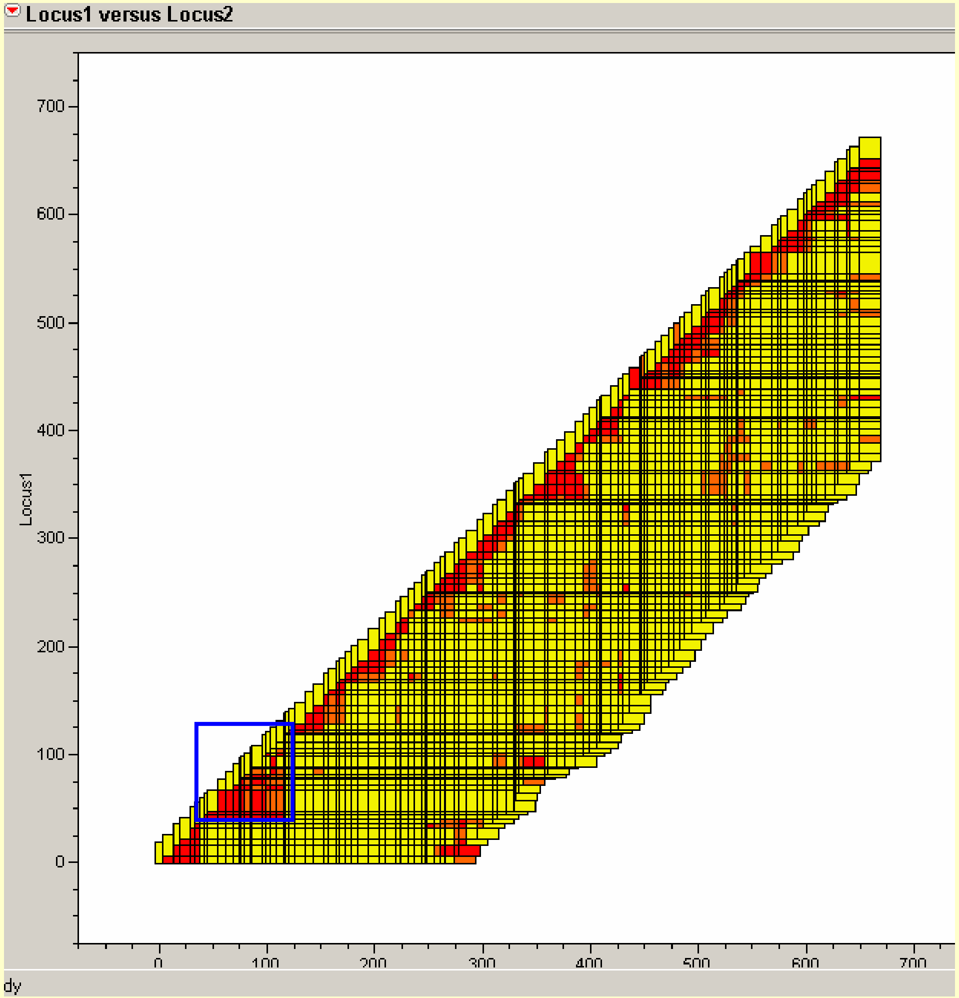
SNP Heat Map



SNPs

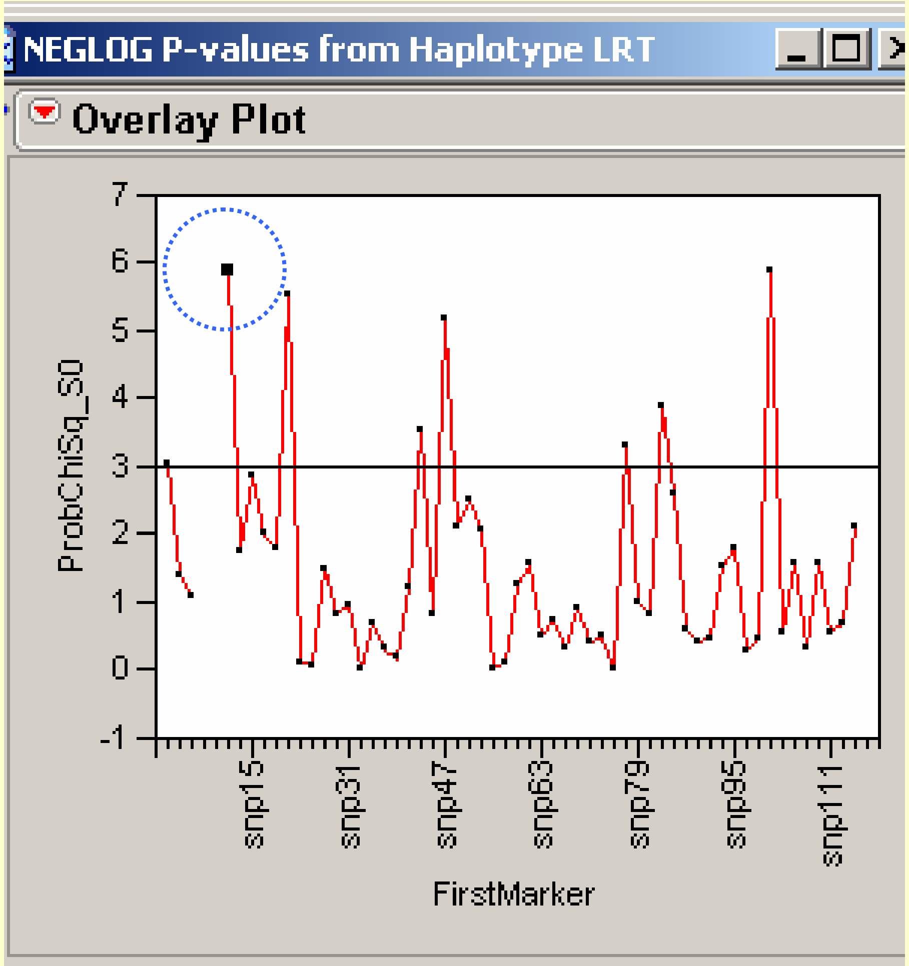
Individuals

Assocation Map



APOE Region in Blue Border

Haplotype Trait Assocation

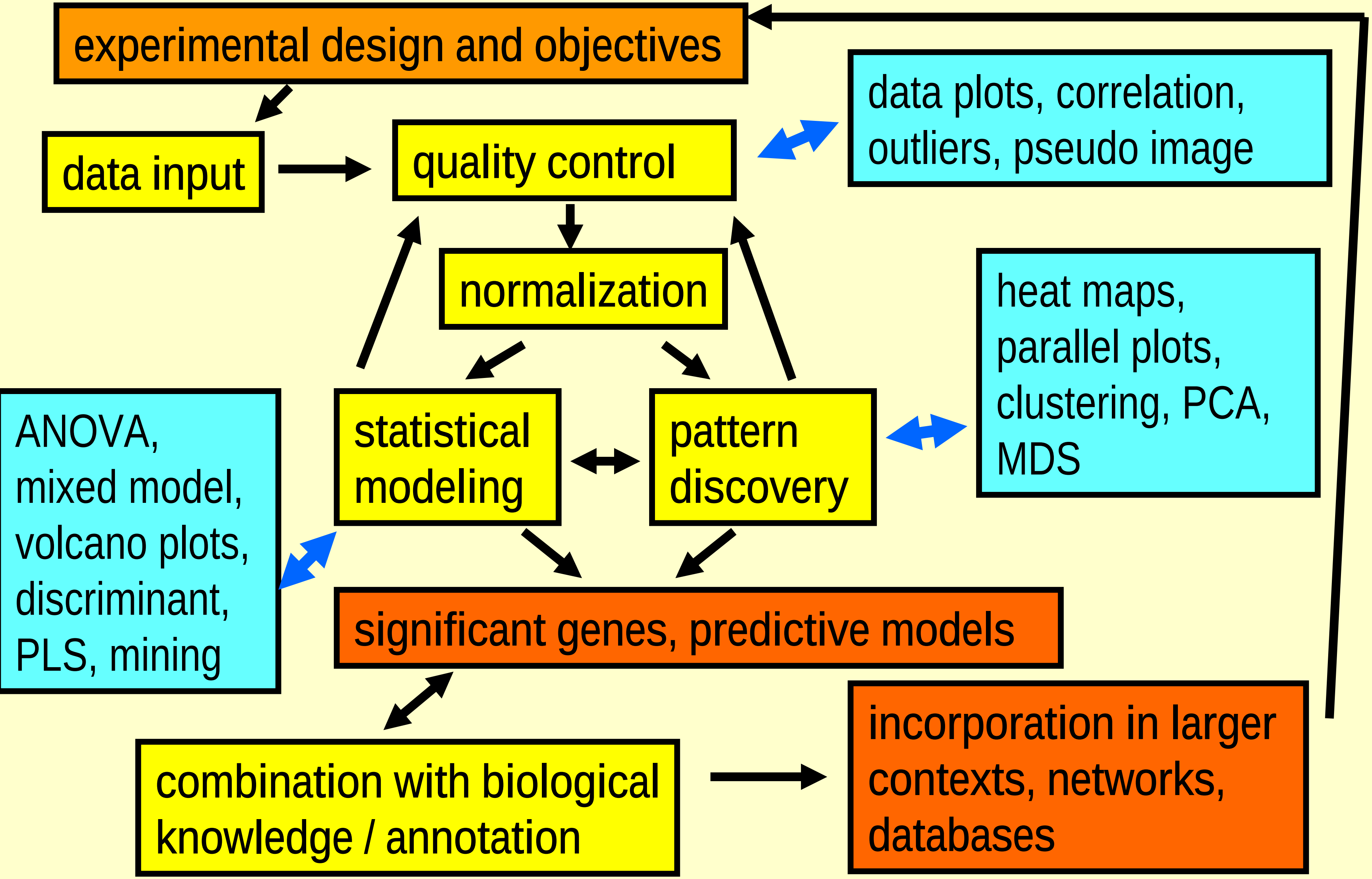


Many Different Technologies for Measuring RNA Abundance

- Polymerase Chain Reaction, Taqman
- Short Oligonucleotide Arrays (Affymetrix)
- Two-Color Arrays (long oligos: Agilent, full length CDNAs: homemade, various vendors)
- Radio Labeling (Clonetech)
- Beads (Illumina, Luminex)
- Serial Analysis of Gene Expression, SAGE
- Chemi-luminescence (Applied Biosystems)
- Nanotech

Competition over quality, content, and economics is hot!

Microarray Statistical Work Flow



experimental design and objectives

data input

quality control

data plots, correlation, outliers, pseudo image

normalization

heat maps, parallel plots, clustering, PCA, MDS

ANOVA, mixed model, volcano plots, discriminant, PLS, mining

statistical modeling

pattern discovery

significant genes, predictive models

combination with biological knowledge / annotation

incorporation in larger contexts, networks, databases

The Importance of Good Experimental Design

- Blazes best path from association to causality

- Especially for two-color arrays, reference designs are typically 2-4 times less efficient than incomplete blocks / loops.

- Be wary of various sources of variation and their nesting, including biological and technical replication

- Split-plot, incomplete block, fractional factorial designs all very relevant and under-utilized. See papers by Gary Churchill, Katie Kerr and colleagues.

- Sample size and power calculations are possible based on standard statistical modeling assumptions.

Microarray Basic Plots

Univariate

- Histograms
- Box Plots

Bivariate

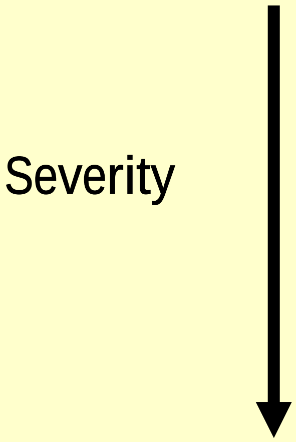
- Correlation Matrix Heat Map
- Array Group Correlation (plot replicates against one another)
- M-A Plots (“minus” versus “average”)

Multivariate

- PCA / Factor Analysis / Biplot
- MDS Plot
- Parallel Coordinate / Profile Plots
- Heat Maps, Dendrograms
- Pseudo-Images

Normalization

Univariate



Severity

Centering

Match a Few Moments or Quantiles

Nonlinear Regression Alignment / Loess

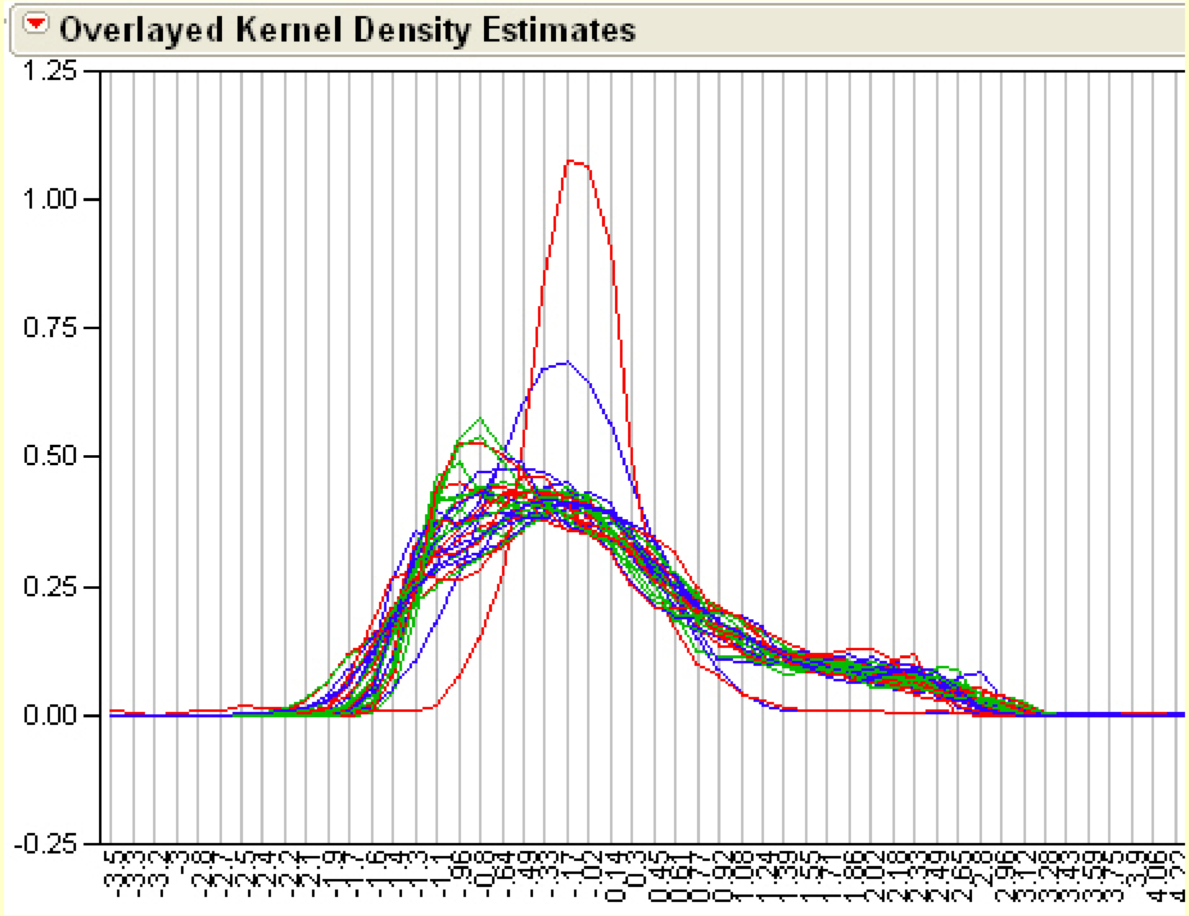
Fully Align All Quantiles or Ranks

Multivariate:

- SVD, Robust SVD (Hawkins and Young)

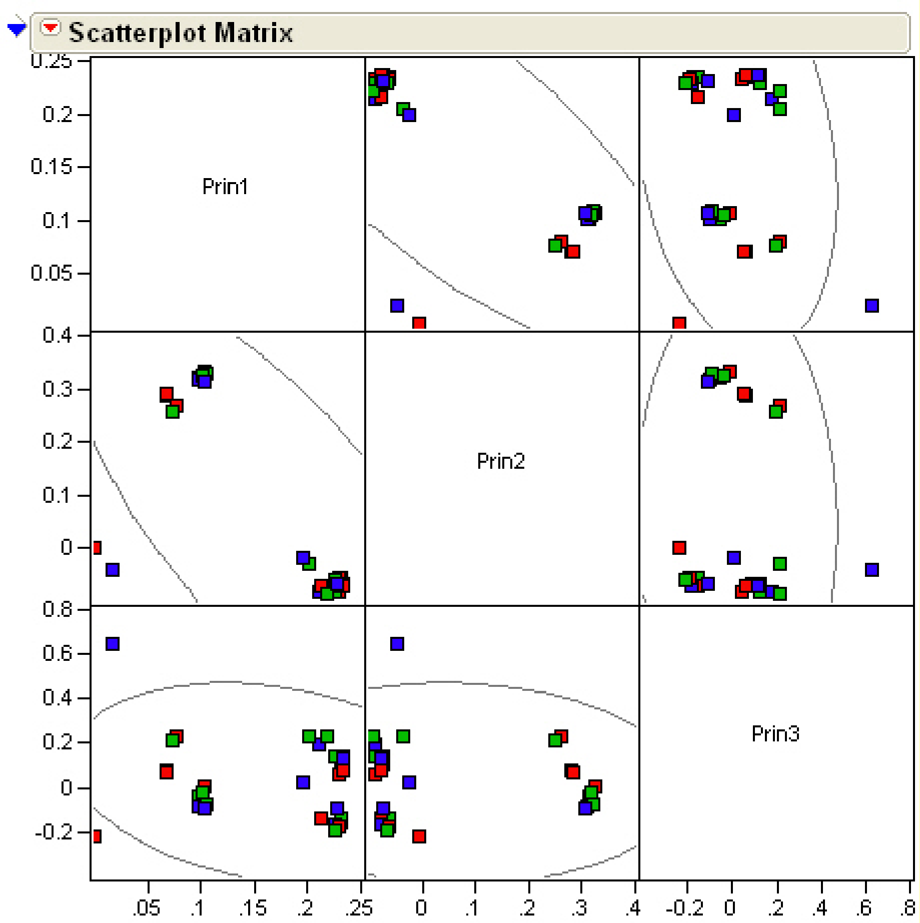
**- Data-mining style normalization for batch effects,
e.g. PLS and Distance Weighted
Discrimination (Marron and colleagues)**

Normalization: Univariate



Univariate standardization reveals two outlying arrays

Normalization: Multivariate



PCA also show outliers, but in addition reveals segmentation of data unrelated to treatment

Analysis Methods

- Scores of methods for assessing differentially expressed genes

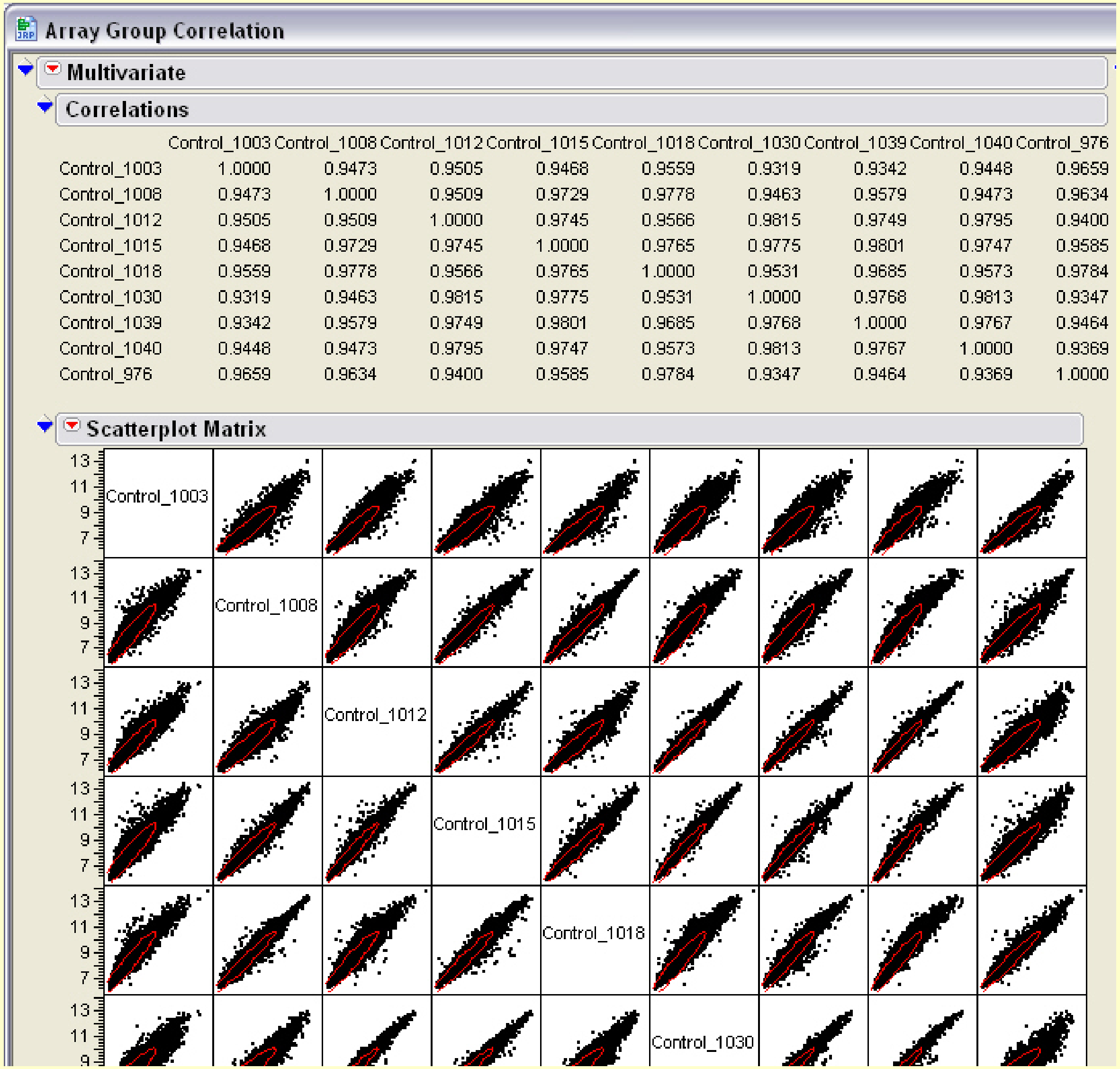
- Just as many for predictive modeling / data mining / classification

- And again as many for pattern discovery / clustering

- Detecting significant associations with molecular annotation, e.g. enrichment analysis on Gene Ontology categories (e.g GoStat by Speed and colleagues) or promoter regions

- Pathways and inferring genetic networks

Alzheimer’s Example from GEO



Array Group Correlation

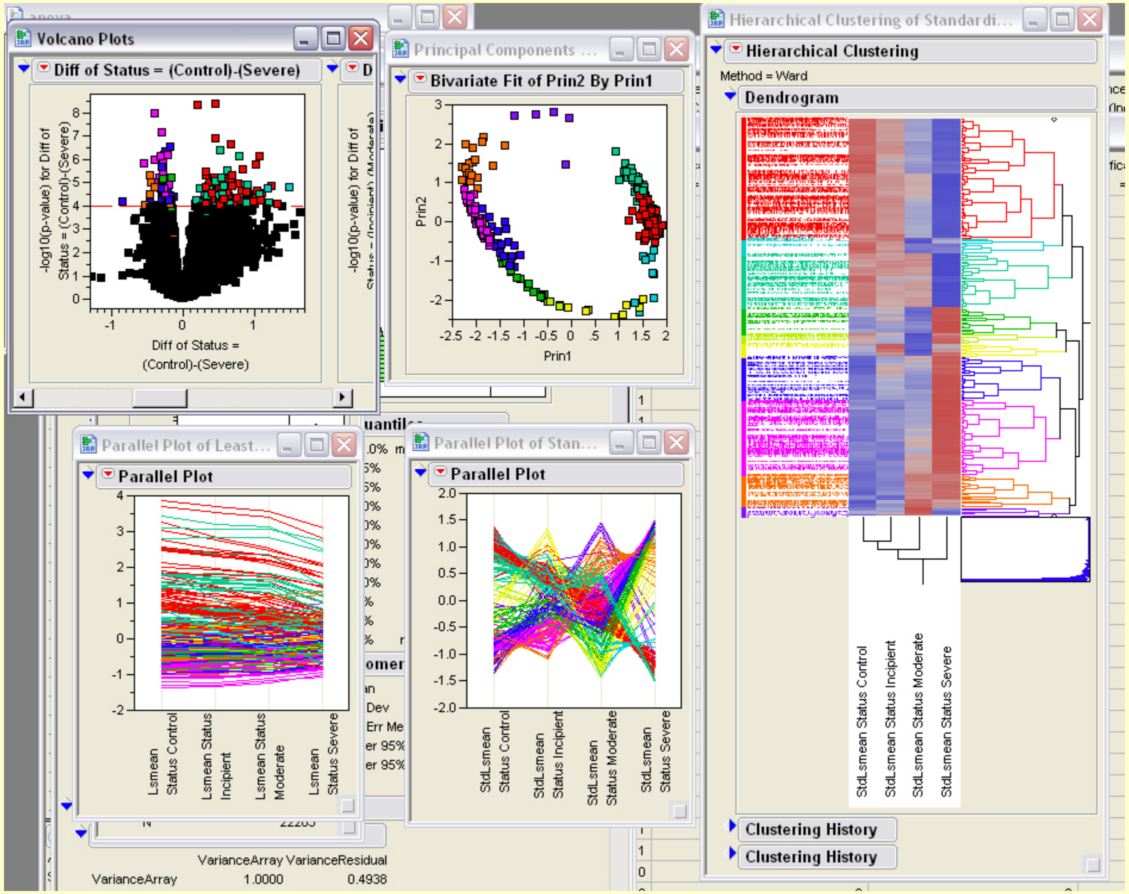
Multivariate

Correlations

	Control_1003	Control_1008	Control_1012	Control_1015	Control_1018	Control_1030	Control_1039	Control_1040	Control_976
Control_1003	1.0000	0.9473	0.9505	0.9468	0.9559	0.9319	0.9342	0.9448	0.9659
Control_1008	0.9473	1.0000	0.9509	0.9729	0.9778	0.9463	0.9579	0.9473	0.9634
Control_1012	0.9505	0.9509	1.0000	0.9745	0.9566	0.9815	0.9749	0.9795	0.9400
Control_1015	0.9468	0.9729	0.9745	1.0000	0.9765	0.9775	0.9801	0.9747	0.9585
Control_1018	0.9559	0.9778	0.9566	0.9765	1.0000	0.9531	0.9685	0.9573	0.9784
Control_1030	0.9319	0.9463	0.9815	0.9775	0.9531	1.0000	0.9768	0.9813	0.9347
Control_1039	0.9342	0.9579	0.9749	0.9801	0.9685	0.9768	1.0000	0.9767	0.9464
Control_1040	0.9448	0.9473	0.9795	0.9747	0.9573	0.9813	0.9767	1.0000	0.9369
Control_976	0.9659	0.9634	0.9400	0.9585	0.9784	0.9347	0.9464	0.9369	1.0000

Scatterplot Matrix

Linked Graphics


Volcano Plots
Diff of Status = (Control)-(Severe)
-log10(p-value) for Diff of Status = (Control)-(Severe)
Diff of Status = (Control)-(Severe)
Principal Components ...
Bivariate Fit of Prin2 By Prin1
Prin2
Prin1
Hierarchical Clustering of Standardi...
Hierarchical Clustering
Method = Ward
Dendrogram
StdLsmean Status Control
StdLsmean Status Incipient
StdLsmean Status Moderate
StdLsmean Status Severe
Clustering History
Clustering History
Parallel Plot of Least...
Parallel Plot
Lsmean Status Control
Lsmean Status Incipient
Lsmean Status Moderate
Lsmean Status Severe
Parallel Plot of Stan...
Parallel Plot
StdLsmean Status Control
StdLsmean Status Incipient
StdLsmean Status Moderate
StdLsmean Status Severe
VarianceArray VarianceResidual
VarianceArray 1.0000 0.4938

Note for Bayesian Interpretation of Volcano Plot:

p-value = 2 times probability sign is wrong

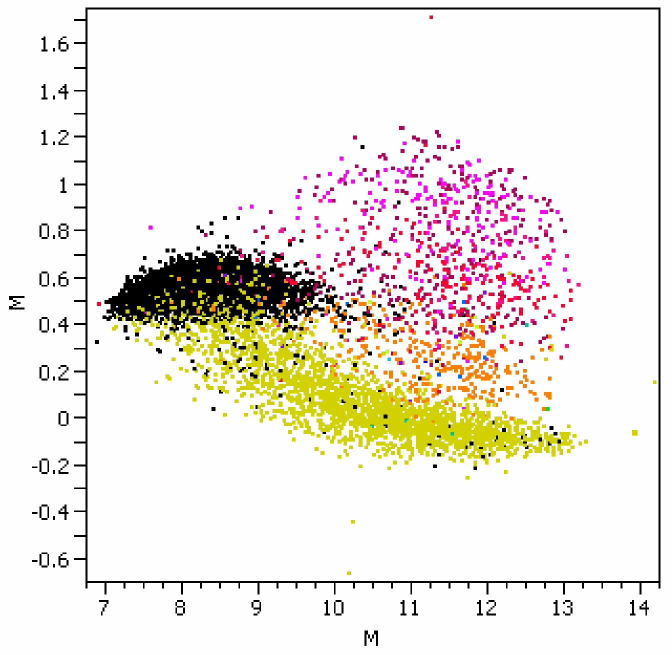
Spike-In Affy Experiment

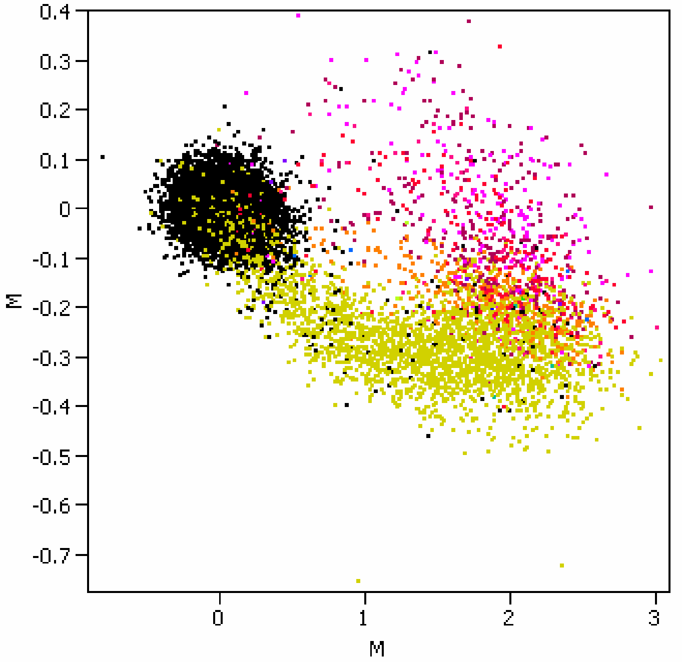
- Choe et al. (2005) *Genome Biology*
- 6 Chips (3 Control, 3 Treatment) with every RNA concentration known!
- Comparison of numerous algorithms in terms of ROC
- Subtracting MM helps remove cross-hyb effect at low end, but

  PM more consistent than PM-MM
- Three interesting clusters in M-A plot of log2(PM) corresponding to

  empty, 1-fold, and higher-fold spike ins

Interesting M-A Plots



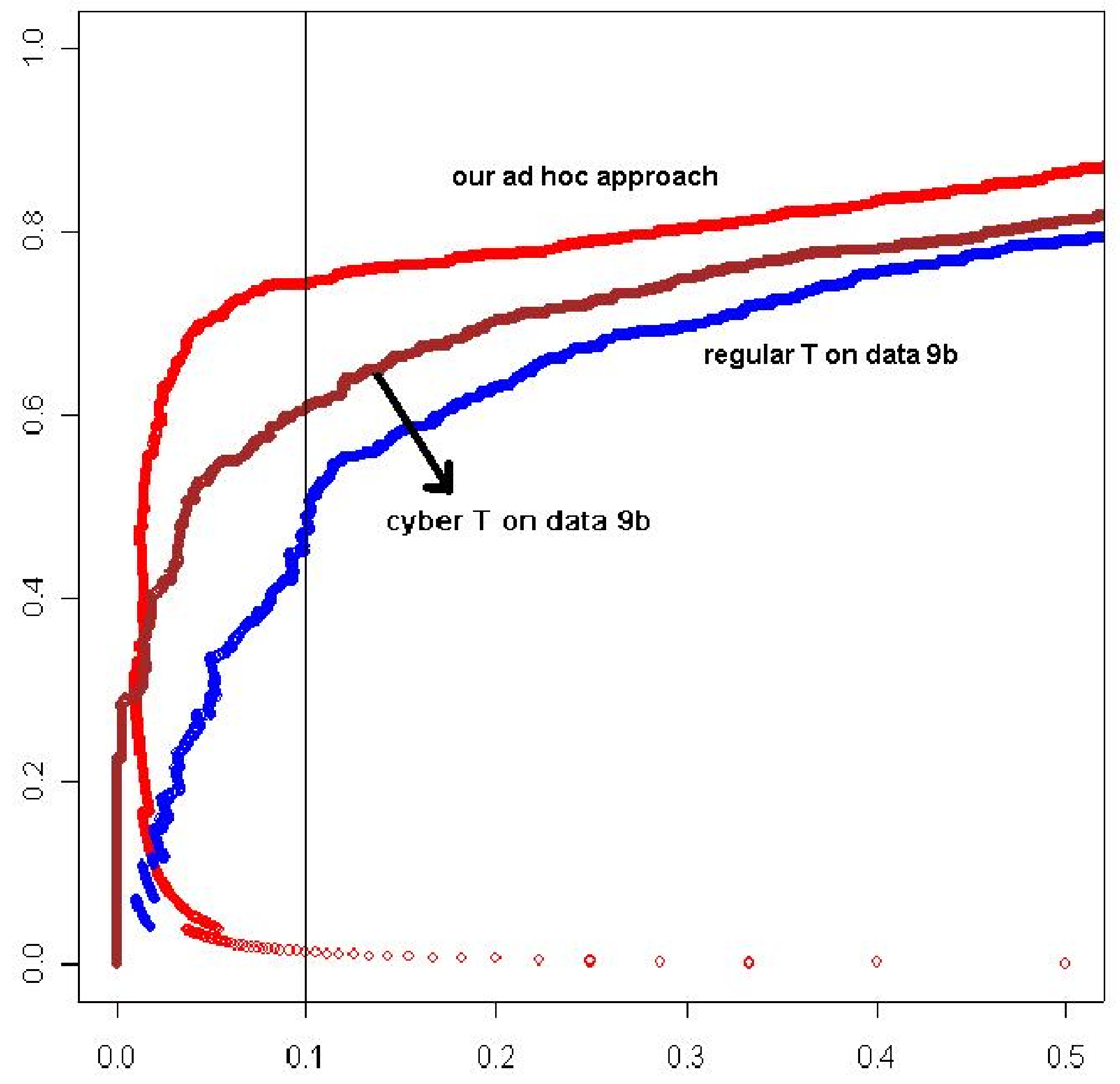


log2_PM

Log2_PM – log2_MM

Improving Spike-In ROC

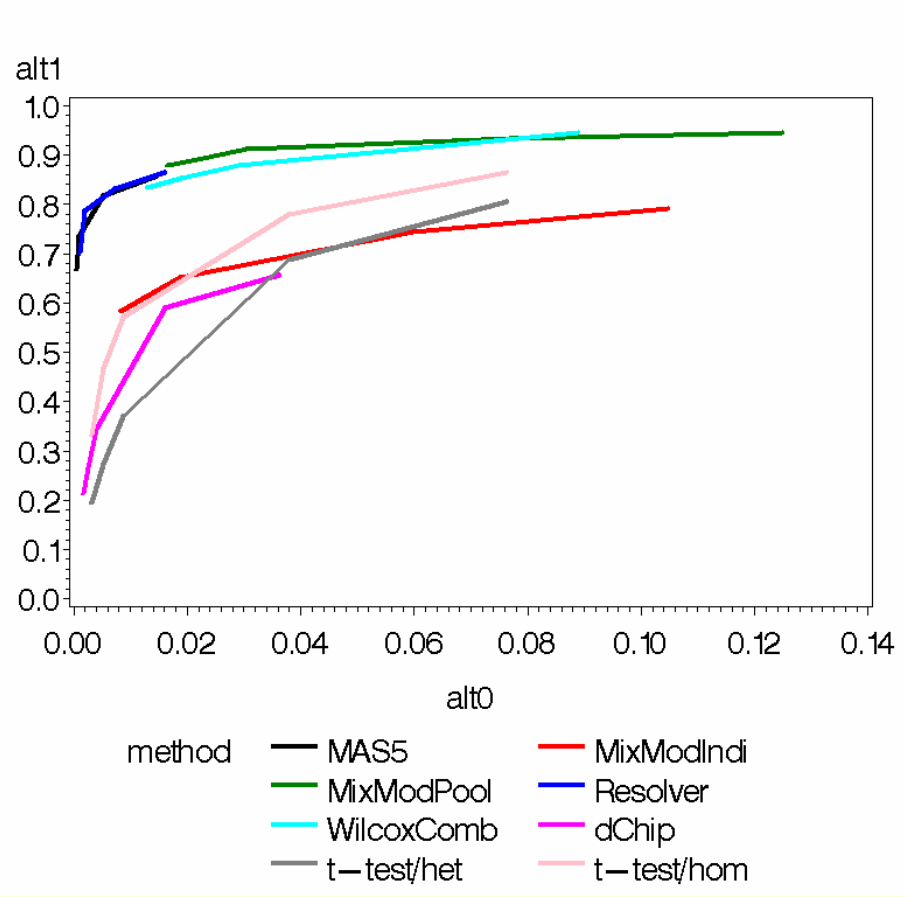
(preliminary unpublished results)



Many Choices for Optimizing ROC

- Response Variable: pm, diff, numerous transformations, including log, glog, linlog, symlog, sqrt, symsqrt
- Normalization/standardization: Global means, medians, loess, statistical and spatial models
- Data filtering: Outlying probes, outlying observations, before/after/during statistical testing
- Statistical decision rule: Various parametric and nonparametric methods, gene-specific versus pooling across replicates and genes

Affy Latin Square ROC



- See Chu et al (2004) Poster
- Some explanations on next slide

Previous Methods for Affy ROC

- MixModIndi: Previously described mixed model approach on log2(PM) and gene-by-gene model fits.
- MixModPool: Use the same variance component estimates for all genes, estimated as the medians from MixModIndi.
- WilcoxComb: Single Mann-Whitney-Wilcoxon rank sum test obtained by first subracting probe means and then doing gene-by-gene tests. Alternative to “5 out of 9” in MAS5 algorithm.

Future Directions for Affy ROC Improvement

- Try other similar data sets (U133 Latin Square from Affy and dilution experiment from GeneLogic)
- Calibration
- Incorporate prior information from probe sequence; deconvolute cross-hybridization
- Bivariate Mixed Model (Hsieh et al), models PM and MM as a bivariate pair

Future Directions for Affy ROC Improvement

- Empirical Bayes Mixed Model (Feng et al), shrinks variance components using priors obtained from all genes, exploits an orthogonalizing transformation based on classical ANOVA
- Redefine ROC in terms of false discovery rate (FDR).
- Papers by Storey and colleagues

Toxicogenomics Example from NCT

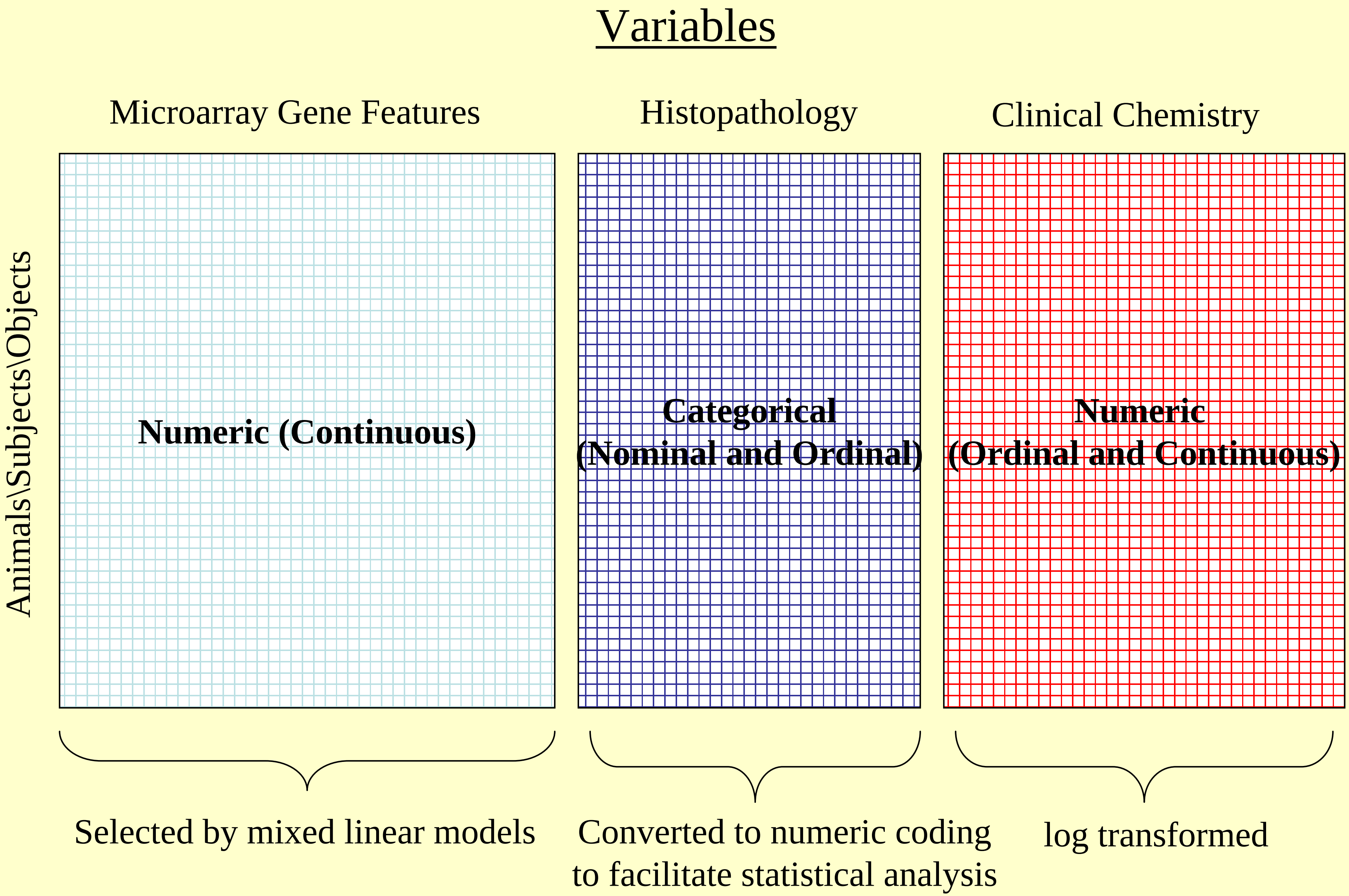
- Acetaminophen dose-response study on rat liver gene expression
- Four doses (50, 150, 1500, 2000 mg/kg) by three times (6, 24, 48 hours), plus dose-time-specific control pools = 24 unique conditions, three biological reps, two technical reps
- 72 total two-color microarrays on 6735 genes, including dye swaps
- Classical clinical chemistry and histopath variables, plus some ultrastructural data
- Heinloth et al. (2004) *Toxicological Sciences* 80: 193-202*.*

http://dir.niehs.nih.gov/microarray/datasets/home-pub.htm

Key Points from Heinloth et al.

- Array data provide a wealth of new information about acetaminophen-induced rat liver gene expression
- Expression changes at low doses can serve as precursors to toxicity.
- Down-regulated genes are involved in energy-consuming biochemical pathways.
- Up-regulated genes are involved in energy-producing biochemical pathways.

Representation of Mixed Data Types For Analysis



Variables

Microarray Gene Features

Histopathology

Clinical Chemistry

Animals\Subjects\Objects

Numeric (Continuous)

Categorical (Nominal and Ordinal)

Numeric (Ordinal and Continuous)

Selected by mixed linear models

Converted to numeric coding to facilitate statistical analysis

log transformed

Slide compliments of Pierre Bushel, NIEHS

Variables

Clinical Chemistry

- Serum enzymes of liver injury
  - Alanine Aminotransferase (ALT), Sorbitol dehydrogenase (SDH), Aspartate aminotransferase (AST)
- Serum glucose and cholesterol
- Indicator of renal injury
  - Urea Nitrogen (BUN)
- Evaluation of cholestasis (bile flow interruption)
  - Total bile acids, 5'-Nucleotidase, Alkaline Phosphatase (ALP)
- Total protein and albumin

Histopathology

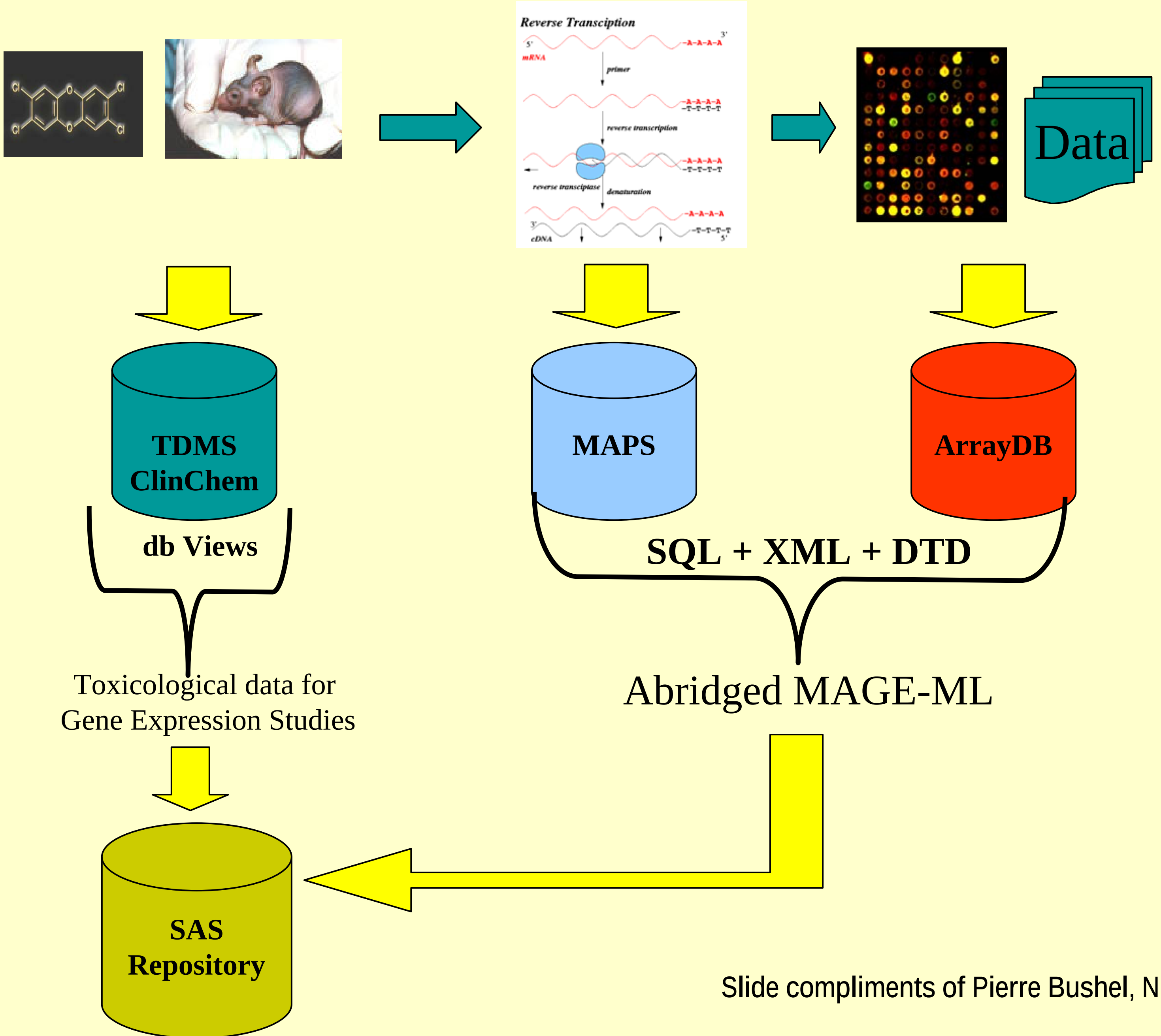
- Topography, site, system
- Morphology
- Severity code
- Chemical (amount, duration)
- Vehicle
- Route of exposure
- Study type (acute, chronic)

Microarray

- 2 channel arrays
- log2 ratios versus controls
- Mixed model results averaged within animals

Slide compliments of Pierre Bushel, NIEHS

Data Integration Process



TDMS ClinChem

MAPS

ArrayDB

db Views

SQL + XML + DTD

Toxicological data for Gene Expression Studies

Abridged MAGE-ML

SAS Repository

Slide compliments of Pierre Bushel, NIEHS

Statistical Workflow

1. Quality check, normalize and filter microarray data for statistically significant changes via mixed-model ANOVA, 6700 -> 444 cDNAs
2. Summarize microarray results to animal level and join with ancillary clinical chemistry and histopath data, remove largely incomplete records -> 30 animals
3. Use std Euclidean distance (<-> correlation) and multidimensional scaling to explore animal and molecular feature dimensions.
4. Fit more statistical models to perform rigorous inference.

Mixed-Model ANOVA

- Extension of analysis of variance to include fixed and random effects. The latter incorporate a Gaussian prior distribution and extend statistical inferences to the population of interest.
- Statistical theory exceptionally well-developed by Henderson (1950s), Harville (1970s), and others
- Strong connections to quantitative genetics methods, e.g. Cockerham and Weir variance component models
- Applied countless times with remarkable success in other areas, e.g. animal breeding, agriculture, clinical trials, epidemiology, pharmacokinetics, spatial statistics, and now tox & molecular biology

Advantages of Mixed-Model ANOVA

- Direct probabilistic modeling of all known sources of variability and correlation (experimental, biological and technical)
- Straightforward accommodation of complex and unbalanced experimental designs
- Provides formal means by which to conduct quality control, e.g. automatic filtering of outlying points
- Model produces rich output, including simultaneous estimates and standard errors for all scientific hypotheses of interest
- Empirical superiority to several other popular microarray analysis methods in terms of false positive / false negative rates

Side Note on Nonparametric Methods

Advantages: **Little to no distributional assumptions on the data, robust**

Disadvantages:

1. **Less power than parametric methods when the latter are well-chosen**
2. **Difficulty in handling more complex experimental designs with implied covariance dependence structures, e.g. time courses, incomplete blocks, intentionally or randomly missing data**
3. **For small experiments, the number of permutations or ranks are limited making sampling distributions too coarse. For large experiments, central limit theorem kicks in and normality assumption is reasonable.**
4. **The validity of the bootstrap is almost always assumed but typically not verified.**
5. **Most genomics data contain an accumulation of numerous small laboratory protocol and instrumentation errors -> central limit theorem and approximately normal errors**

Multidimensional Scaling

- Takes as input a distance matrix and produces low-dimensional coordinates (1, 2, or 3D) that optimally preserve inter-point distances.
- Contrast with principal components, which creates a projection with maximal variability / separation.
- Aside: A wide variety of distance metrics exist, and scientists should consider their appropriateness and devise new ones.

Toxicogenomics Parting Thoughts

Some essential components:

- Good experimental design
- Pre-processing, including quality control, normalization, and reduction
- Well-chosen statistical methods
- Dynamic visualization
- Careful annotation and connection to biochemical knowledge

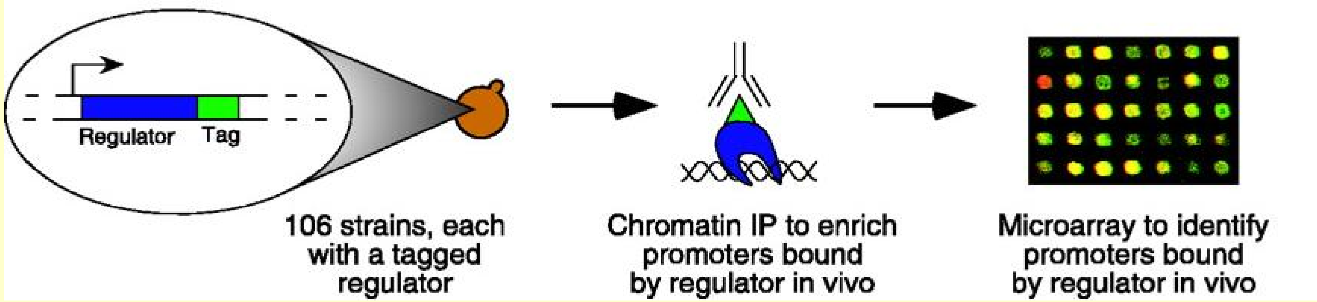
Joining array, histopath, and clinchem data is nontrivial but straightforward; analysis is tougher.

RNA profiles are clearly associated with tox phenotypes and have strong potential for inference and prediction.

Significant challenges remain for risk assessment!

ChIP-chip experiment setup: genome-wide location analysis method

Lee *et al*. 2002 *Science,* 298: 799-804

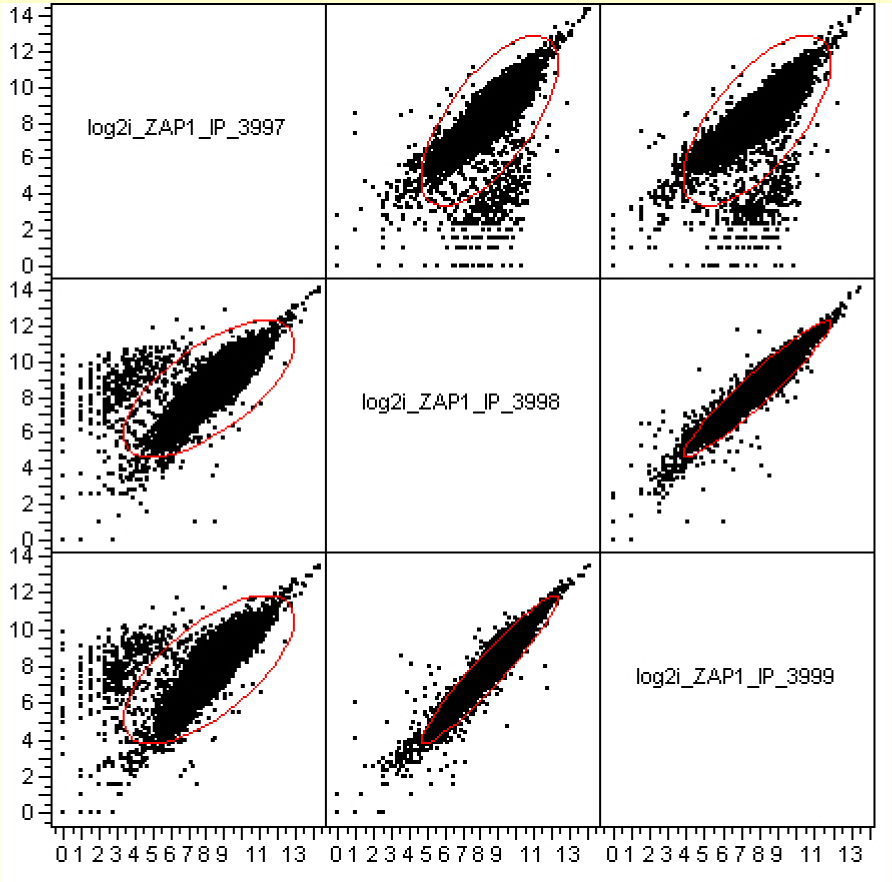

Regulator
Tag
106 strains, each with a tagged regulator
Chromatin IP to enrich promoters bound by regulator in vivo
Microarray to identify promoters bound by regulator in vivo

1. Construct yeast strains, each transcription factor (TF) has a *myc* tag.
2. Chromatin immunoprecipitation (ChIP) was used to separate promoters bound by tagged TF.
3. Two probes, ChIP and genome control, labeled with Cy3 or Cy5, were applied to a microarray chip containing 6279 gene promoters.

ChIP-Chip Yeast Example

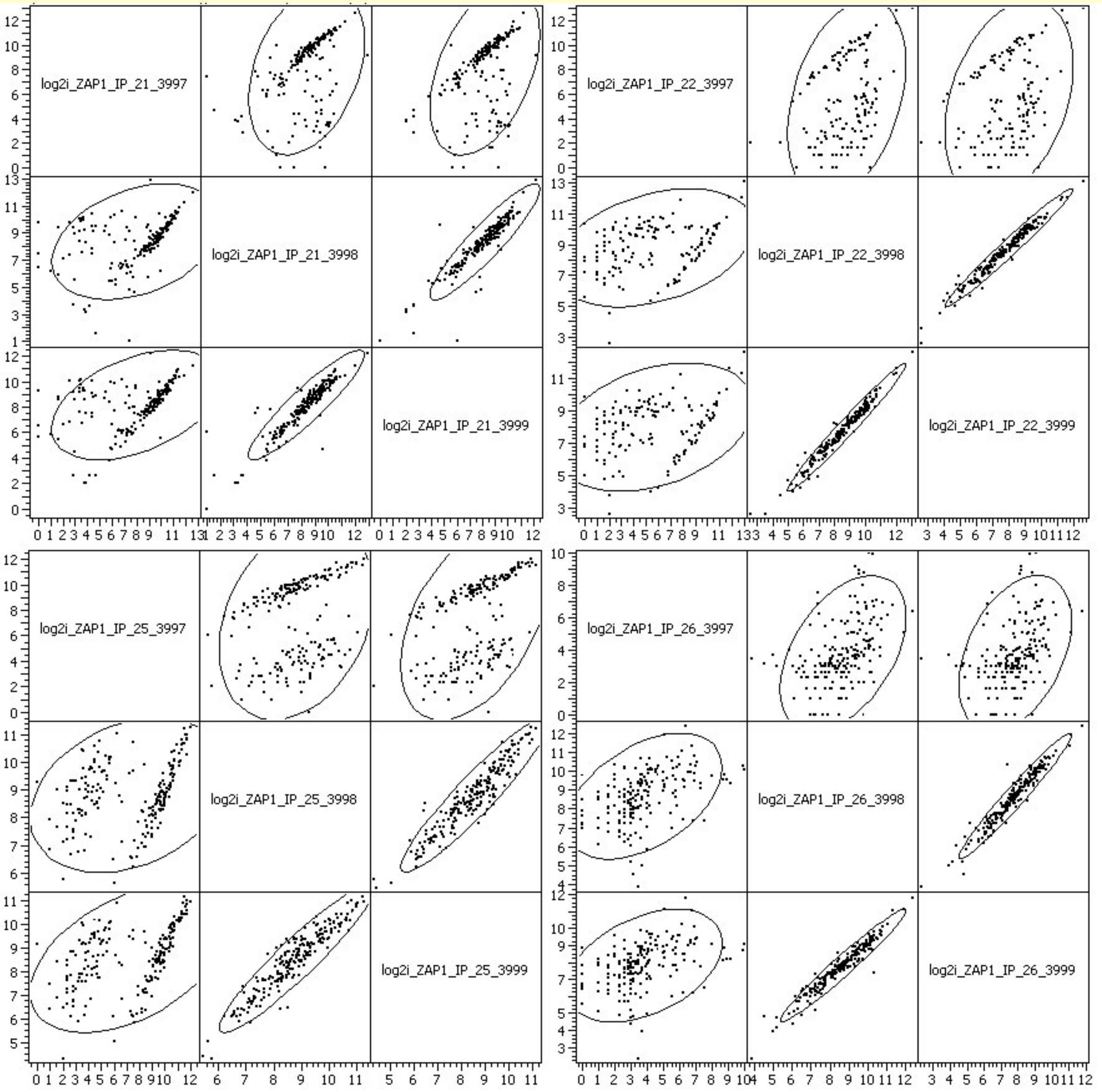
- 106 transcription factors
- 300 arrays, with 2~4 replicates (mostly 3) for each TF
- 7,200 spots on each array, representing promoters of ~6,300 genes
- Two channels, Cy3 and Cy5 dye for each spot
- Two probes, IP and WCE, corresponding to IP and genome control DNA, completely confounded with dye
- Study design and analysis results using Rosetta Error Model described in Lee *et al*. (2002)
- Re-analysis in Yu, Chu, Gibson, Wolfinger (2004) *SAMG,* 3:1, Article 22

Quality inspection: 3 replicates from Zap1



Scatter plot of 3 replicates for Zap1

Scatter plot by block for ZAP1



Block 21

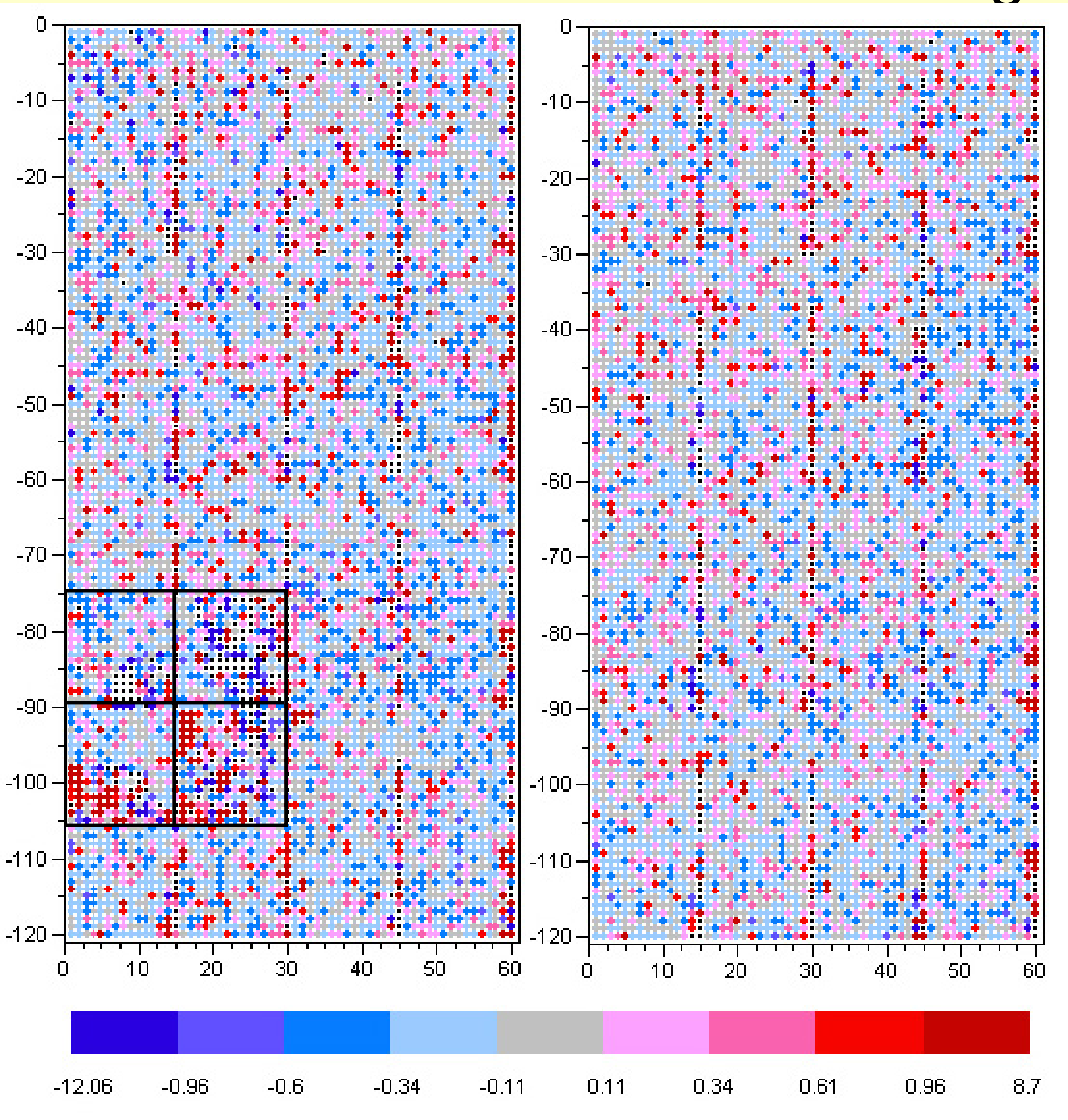
Block 22

Block 25

Block 26

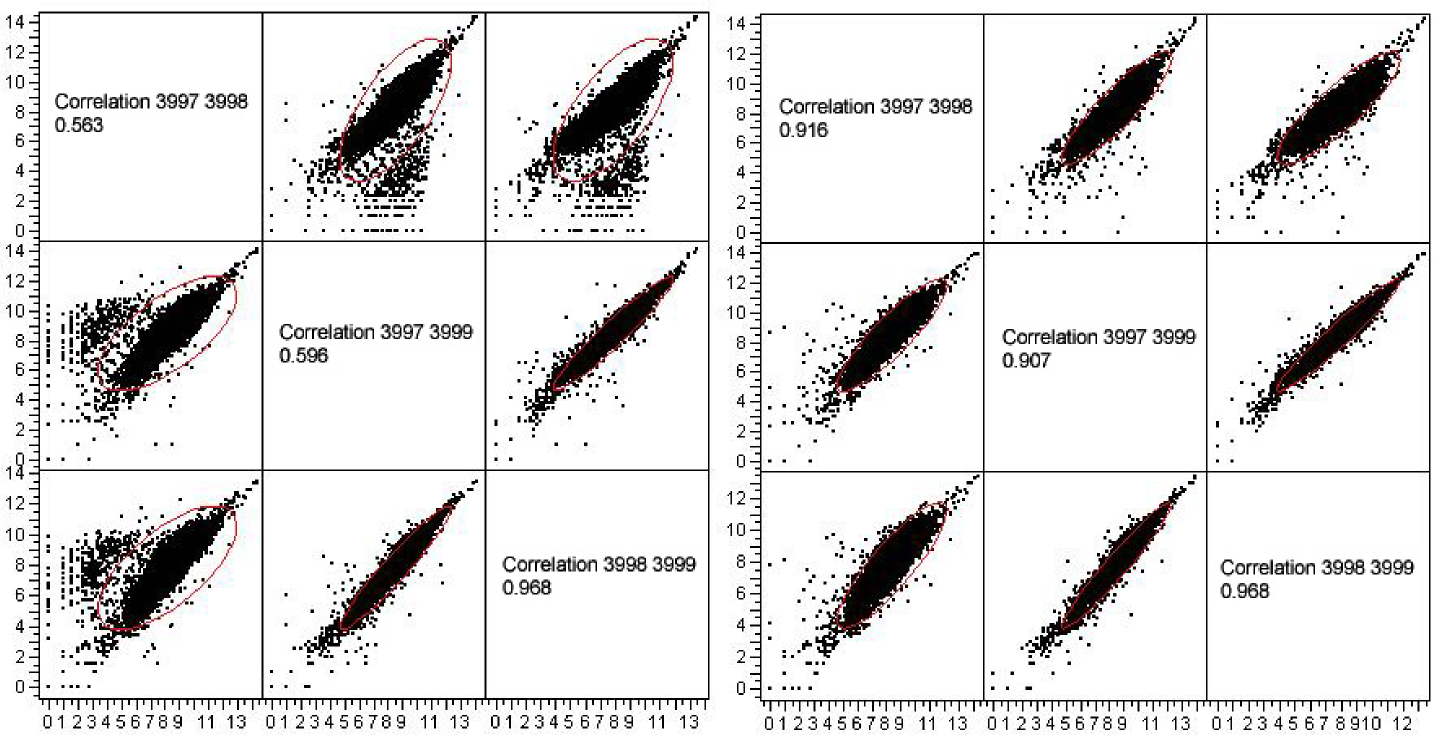
Further visualization: Pseudo image

Array 3997



Array 3998

Removing bad blocks - example of Zap1



Mixed Model Analysis

- Residuals from the normalization model were taken as input
- Fit a linear mixed model for each gene

$$r_{tpa} = \mu + T_t + P_p + A_a + TP_{tp} + \varepsilon_{tpa}$$

$T,$ main effect of TF

$P,$ main effect of probe, $p = 1$ for IP and $p = 2$ for control

$TP,$ main effect of TF-probe interaction

$A,$ array random effect, $A \sim N(0, \sigma_A^2)$

Mixed Model Assumptions Check

Verify normal error assumption by looking at standardized or studentized residuals from the gene-by-gene model fits.

- histograms
- Q-Q plots
- first 4 moments

Can also look at standardized or studentized empirical BLUPs of chip and channel effects.

Hypothesis Testing

- Goal: for each gene, find if a transcription factor T_t significantly binds the promoter of that gene.
- As from the experimental design, promoter bound by TF is enriched in the IP probe.
- Form the contrast of the probe-TF interaction and do a one-sided test:

$$H_0 : \ TP_{t1} - TP_{t2} <= 0$$

$$H_1 : \ TP_{t1} - TP_{t2} > 0$$

Affymetrix All-Exon Array

Collaboration with Eric Hoffman and Marina Bakay,
Children's National Medical Center, PEPR Database

Chip Design – Rough Outline

1. "Exon" prediction information compiled from a variety of publicly available sources.
2. "Exons" are divided into Probe Selection Regions (PSRs)
3. 4 probe pairs are designed for each PSR (when possible)
   - PSR must be >17 nt
4. PSRs are joined into *Exon Clusters*
5. Exon Clusters are joined into *Transcript Clusters*

Slide compliments of Eric Hoffman, CNMC

Exon prediction categories used for initial prototype design

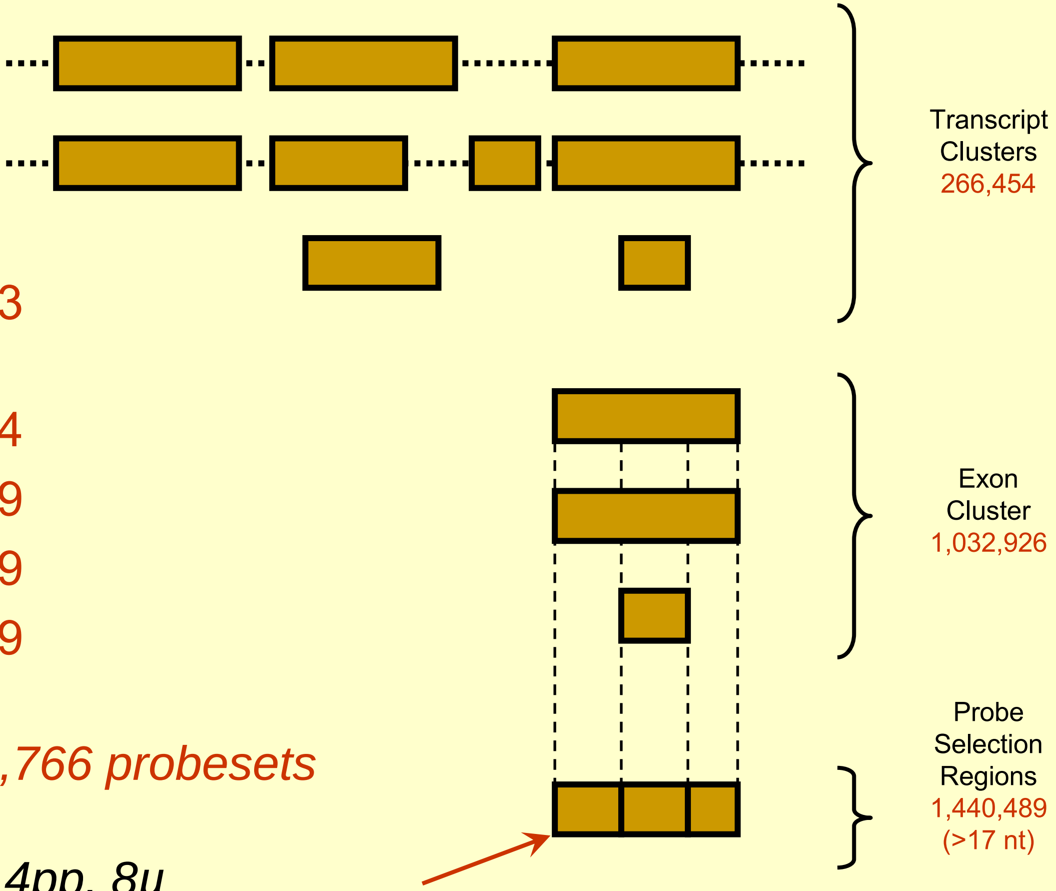
- EnsGene
  - Heterogenous prediction sets that uses protein homology, cDNA alignments, and ab inition predictions. This prediction set was generated by the Ensembl group
- Genscan
  - HMM-based ab initio gene prediction set generated by Chris Burge
- GenscanSubopt
  - High-scoring exons which are not in the optimal parse of the genscan HMM. This set was also generated by Chris Burge
- Twinscan
  - HMM-based gene predictor, similar to genscan, which uses synteny to improve predictions. The twinscan predictions were generated by Michael Brent
- SLAM
  - Paired HMM-based ab initio gene prediction that generates orthologous predictions in both human and mouse. Affymetrix internal
- Full-length cDNA, mRNA, and ESTs
  - Transcript-derived alignments to the human genome using BLAT

Slide compliments of Eric Hoffman, CNMC

Human exon arrays: An inclusive design to discover alternative splicing

Predictors

EnsGene	342,843
GenScan(SubOpt)	326,514
SLAM	176,759
TwinScan	191,589
cDNA/EST	502,019



Transcript Clusters 266,454

Exon Cluster 1,032,926

Probe Selection Regions 1,440,489 (>17 nt)

Final design includes 1,495,766 probesets

10 million features

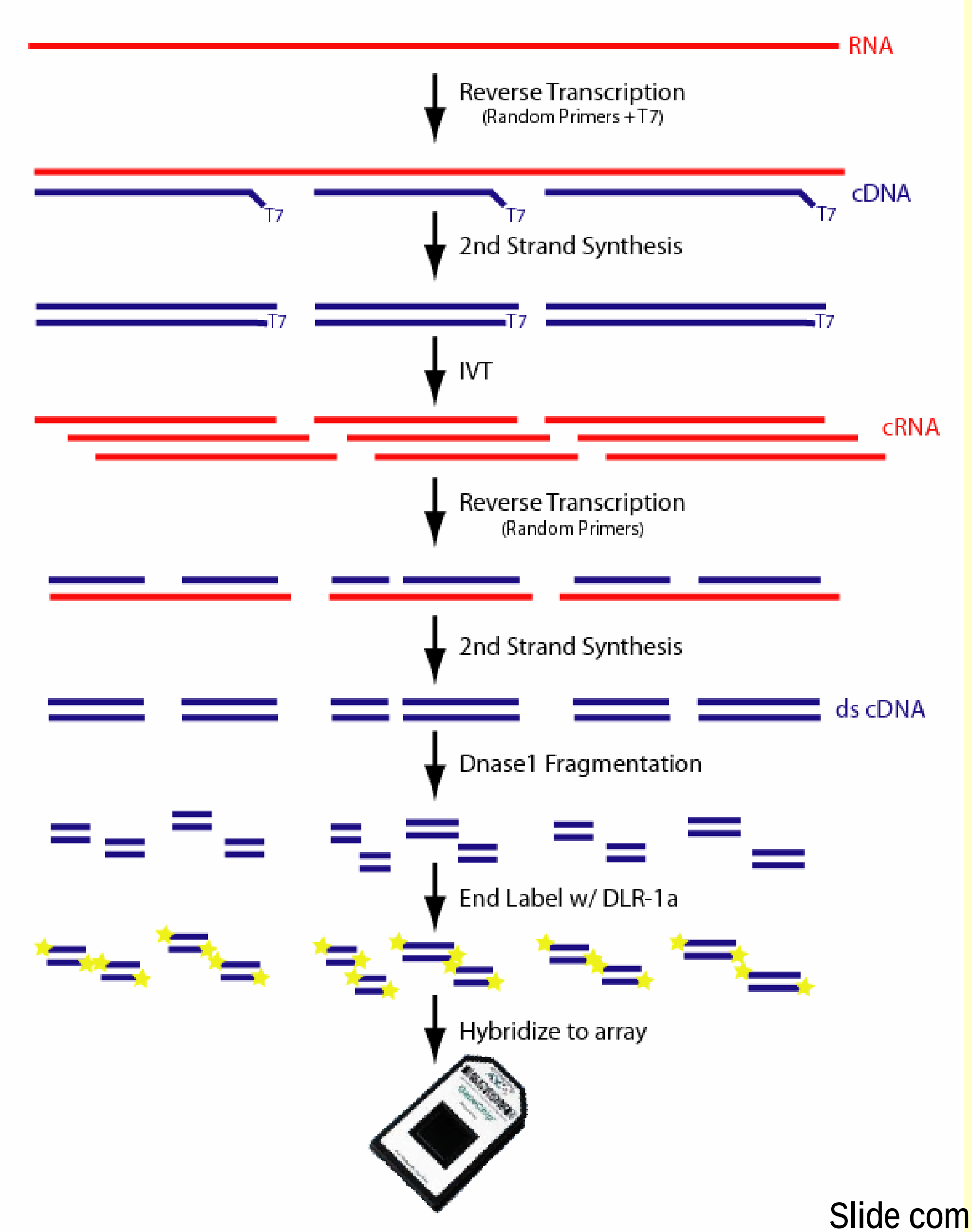
Four 49-format arrays, 4pp, 8µ

Divided by chromosome

Median length 119 nt

Slide compliments of Eric Hoffman, CNMC

Version 2.0 Target Prep Protocol (sWTA)



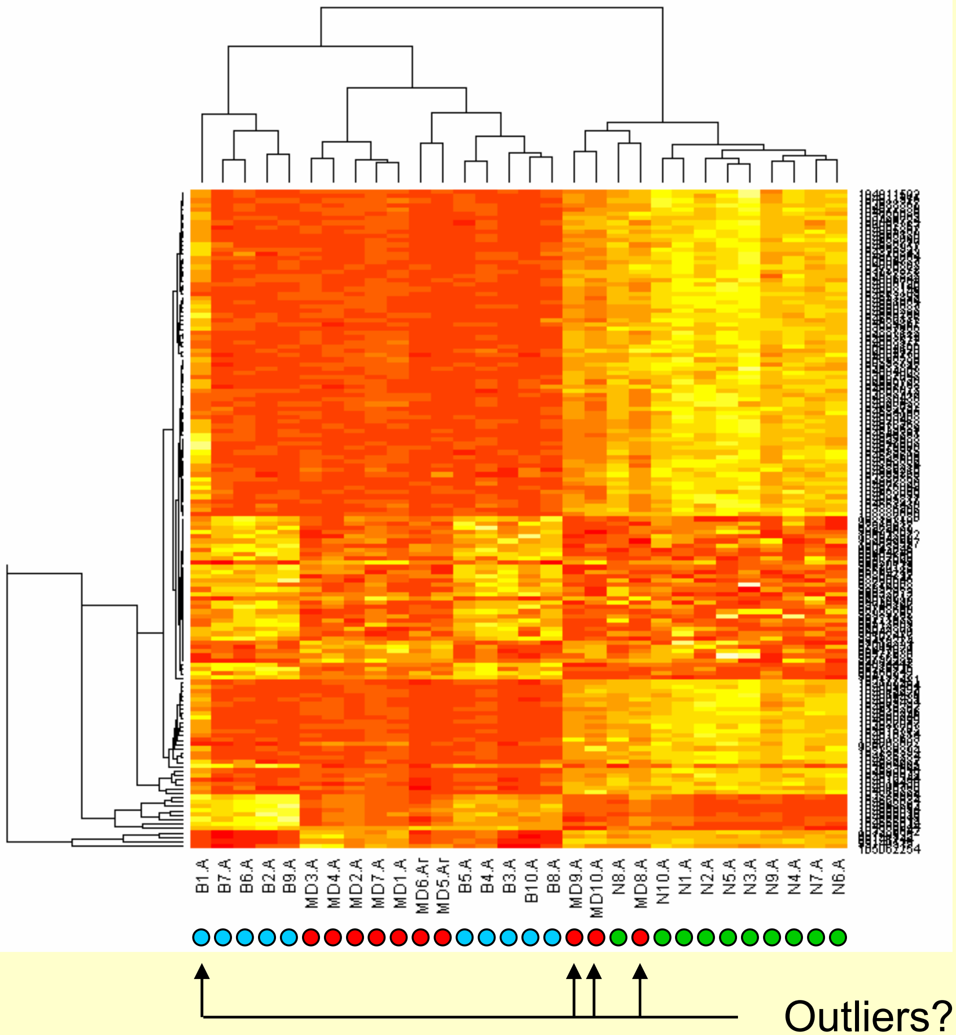
Slide compliments of Eric Hoffman, CNMC

All-Exon Example from CNMC

- Myotonic dystrophy
  - Dominant disorder
  - Trinucleotide repeat expansion (CTG)
  - 3’ UTR of kinase gene
    - **Not** a poly-glutamine disorder like all other dominant expansion disorders
  - Dominant “RNA toxicity” disorder (Wang et al. 1994)
  - Alters splicing “in trans” by sequestering splicing machinery
- Goal:: Define abnormal splicing from patient tissue on genome-wide scale
- Experimental Design:
  - 10 Normal volunteers
  - 10 Myotonic dystrophy

Slide compliments of Eric Hoffman, CNMC

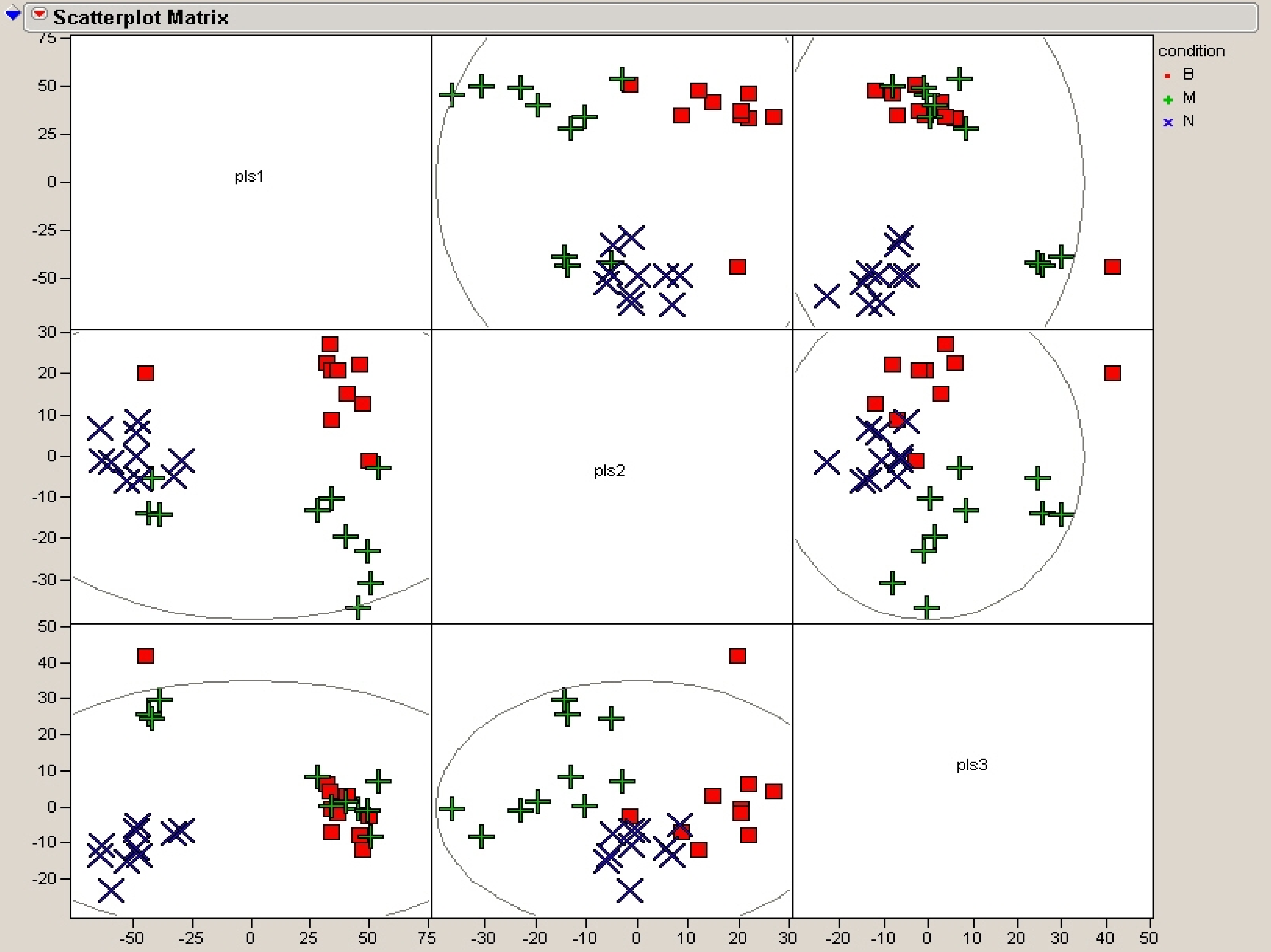
Clustering using the top 149 (top 50 from each pair-wise comparison) for Normal, Myo Dys, and Becker's (DMD)



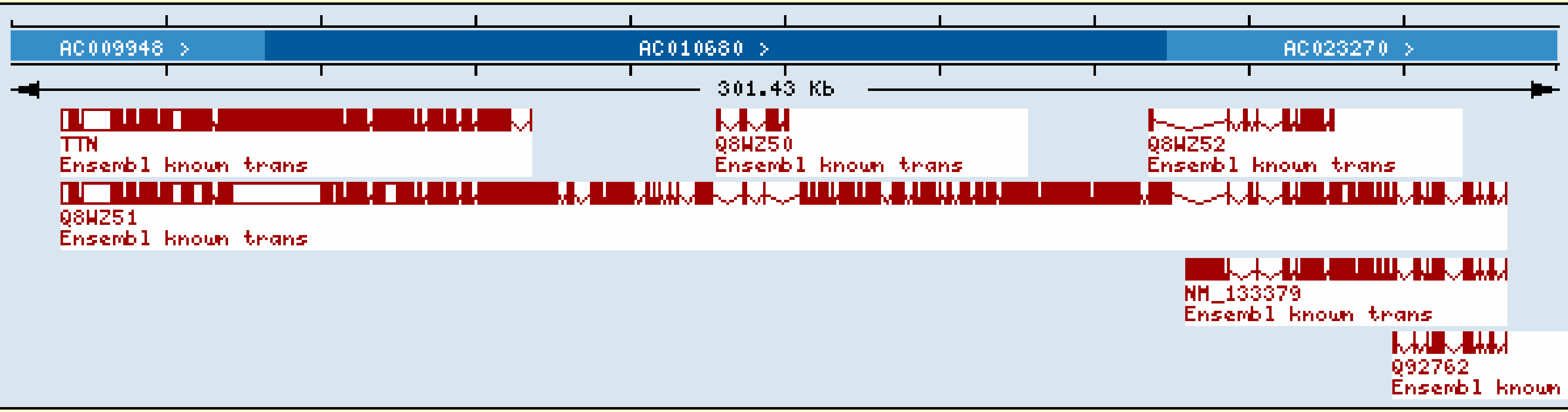
Outliers?

- 🟢 Normal
- 🔴 Myo Dys
- 🔵 Becker's

Slide compliments of Eric Hoffman, CNMC


Scatterplot Matrix
condition
B
M
N
pls1
pls2
pls3
75
50
25
0
-25
-50
30
20
10
0
-10
-20
-30
50
40
30
20
10
0
-10
-20
-50
-25
0
25
50
75
-30
-20
-10
0
10
20
30
-20
-10
0
10
20
30
40
50

Titin isoform novex-3

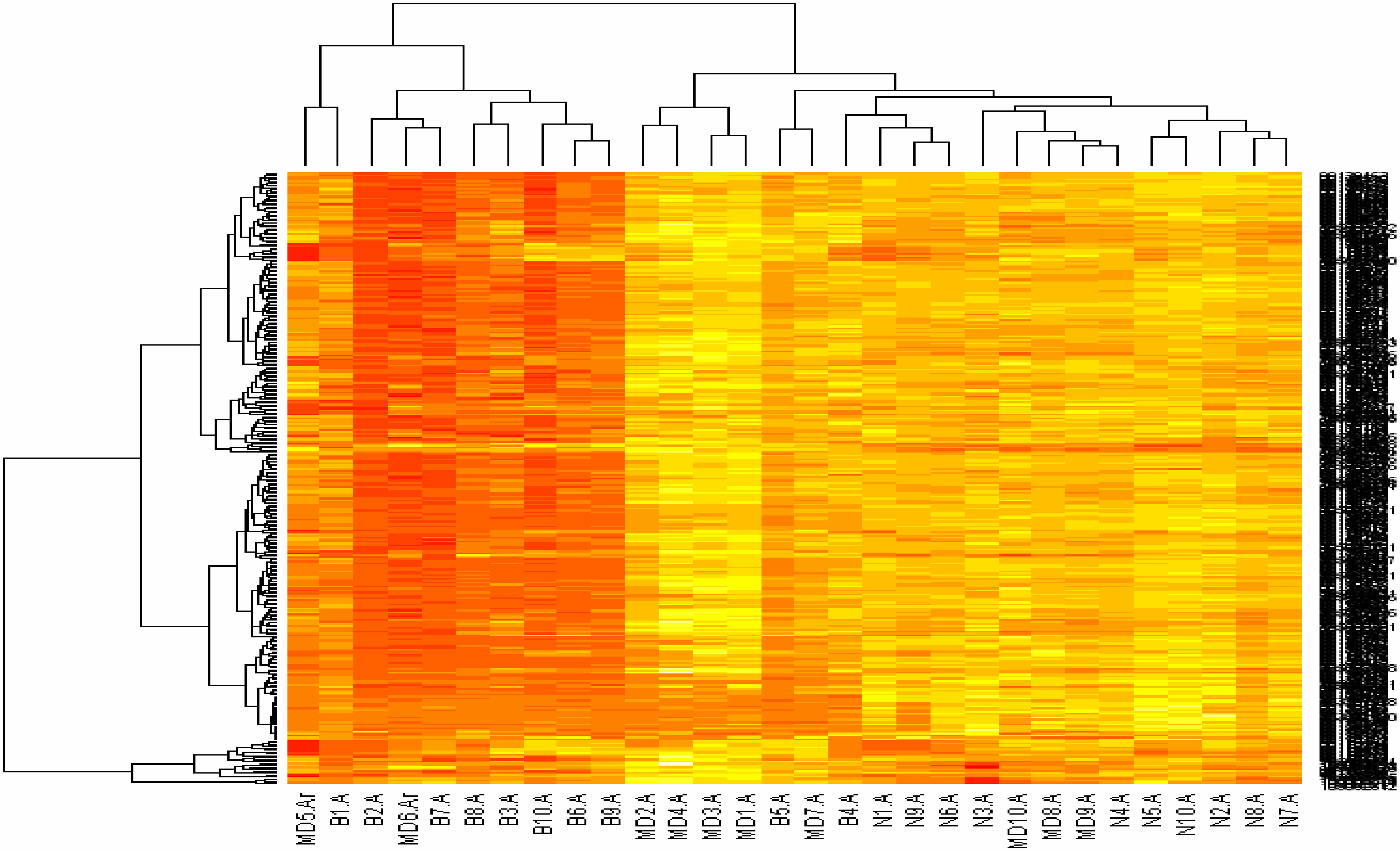


Transcript Variant: This variant (novex-3) is the shortest transcript and encodes the shortest protein. The last exon in the novex-3 variant is nearly 7 kb and is not found in the N2-A transcript. The novex-3 isoform, found in all striated muscle, lacks the PEVK region and is a C-terminal truncation.

589 probe sets on Chip A for ENSG00000155657

Slide compliments of Eric Hoffman, CNMC

Clustering for 312 of the 589 probe sets for Titin isoform novex-3



Challenges for All-Exon Array Analyses

- What underlies chip/sample outliers?
  - Trinucleotide repeat length
  - Diagnosis
  - Confounding variables (sex, age, ethnicity)
- Normalization of all three groups?
  - Find alternative splicing of MyoDys vs Becker
  - Then compare splicing patterns to normal
- Validation

Slide compliments of Eric Hoffman, CNMC

Proteomics and Metabalomics

- Waves of new data
- Real chemical action in peptides and metabolites, hope and potential for biomarker discovery is high
- Appropriate data pre-processing is absolutely critical, not to mention good experimental design (e.g. controversy over results from Petricoin and Liotta).

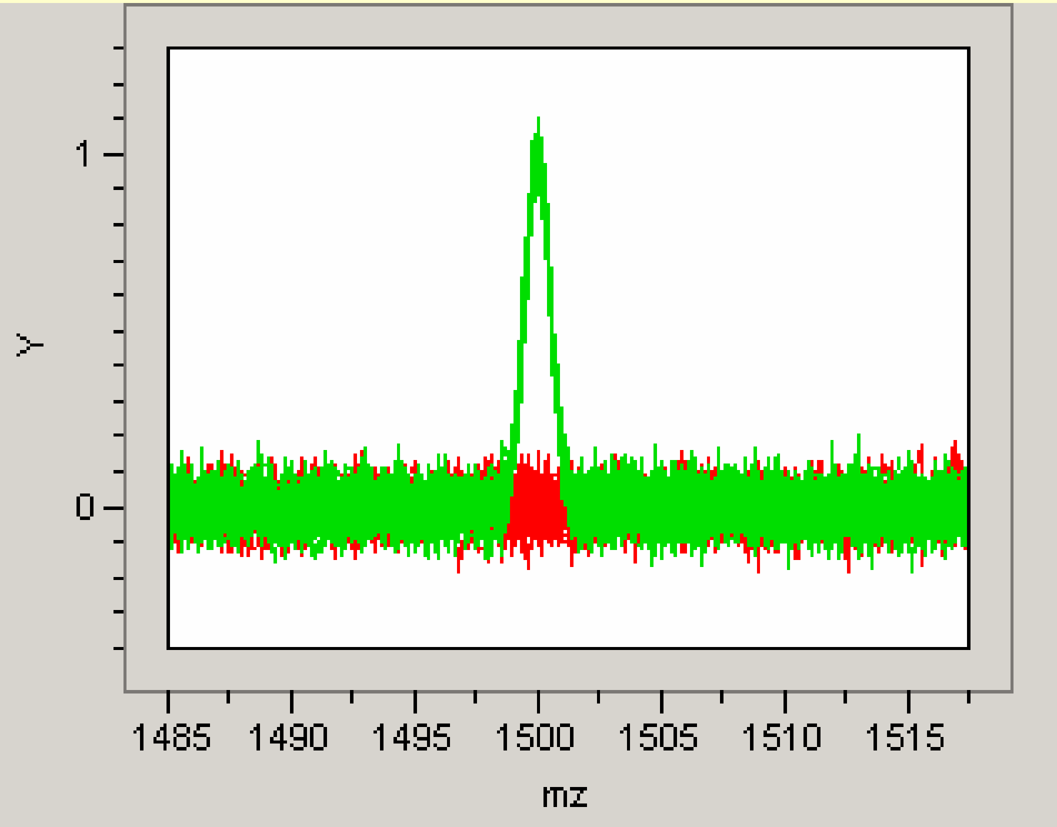
Common Technologies

- 2-D gels
- Mass spectrometry (2D and 3D, many flavors and acronyms, e.g. MALDI TOF, SELDI, QTOF, Electrospray, LC/MS)
- Using the preceding with dual labeling techniques such as ICAT
- Nuclear magnetic resonance (NMR)
- Antibody chips

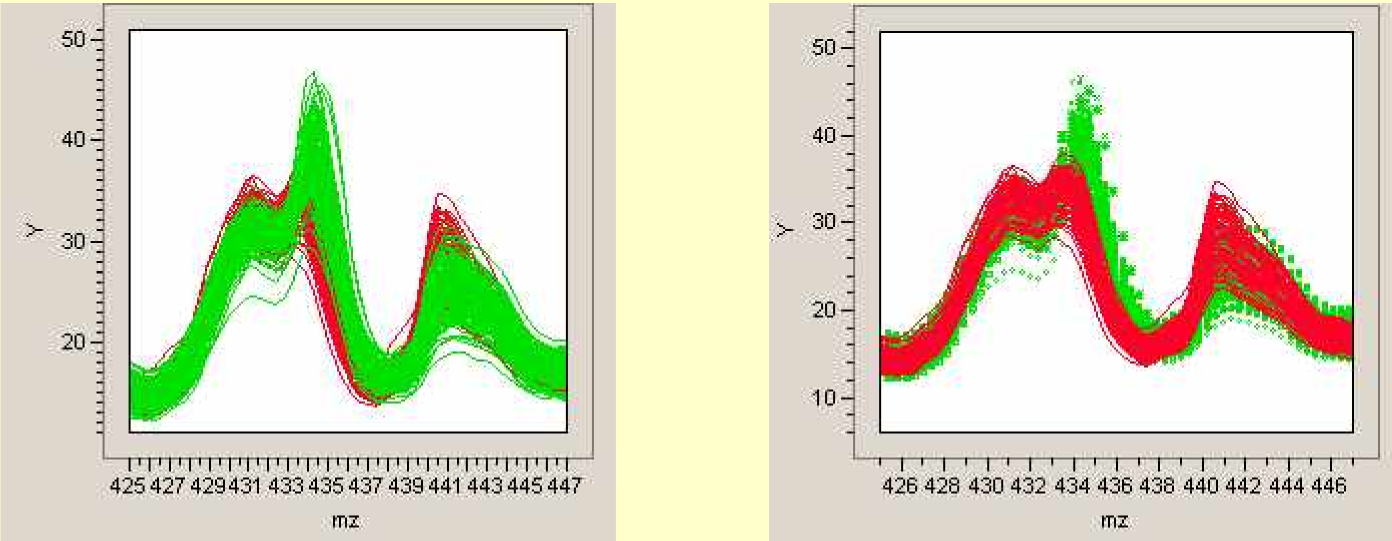
Analysis Ideas

- Utilize best methods from:
  - statistics
  - econometrics / time series
  - chemometrics
  - image analysis, e.g. image alignment / registration
- Once metabolites or peptides are appropriately aligned and quantitated, apply similar workflow as with microarray data.
- Chemist ≠ Biologist

What We’d Love to See



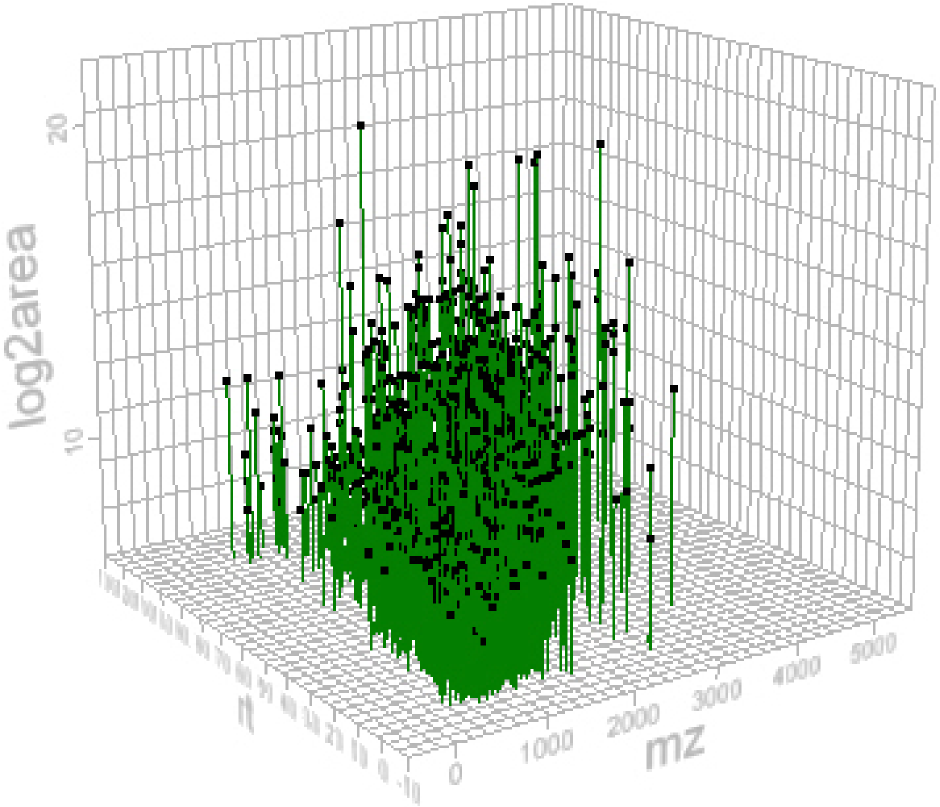
What We Often See



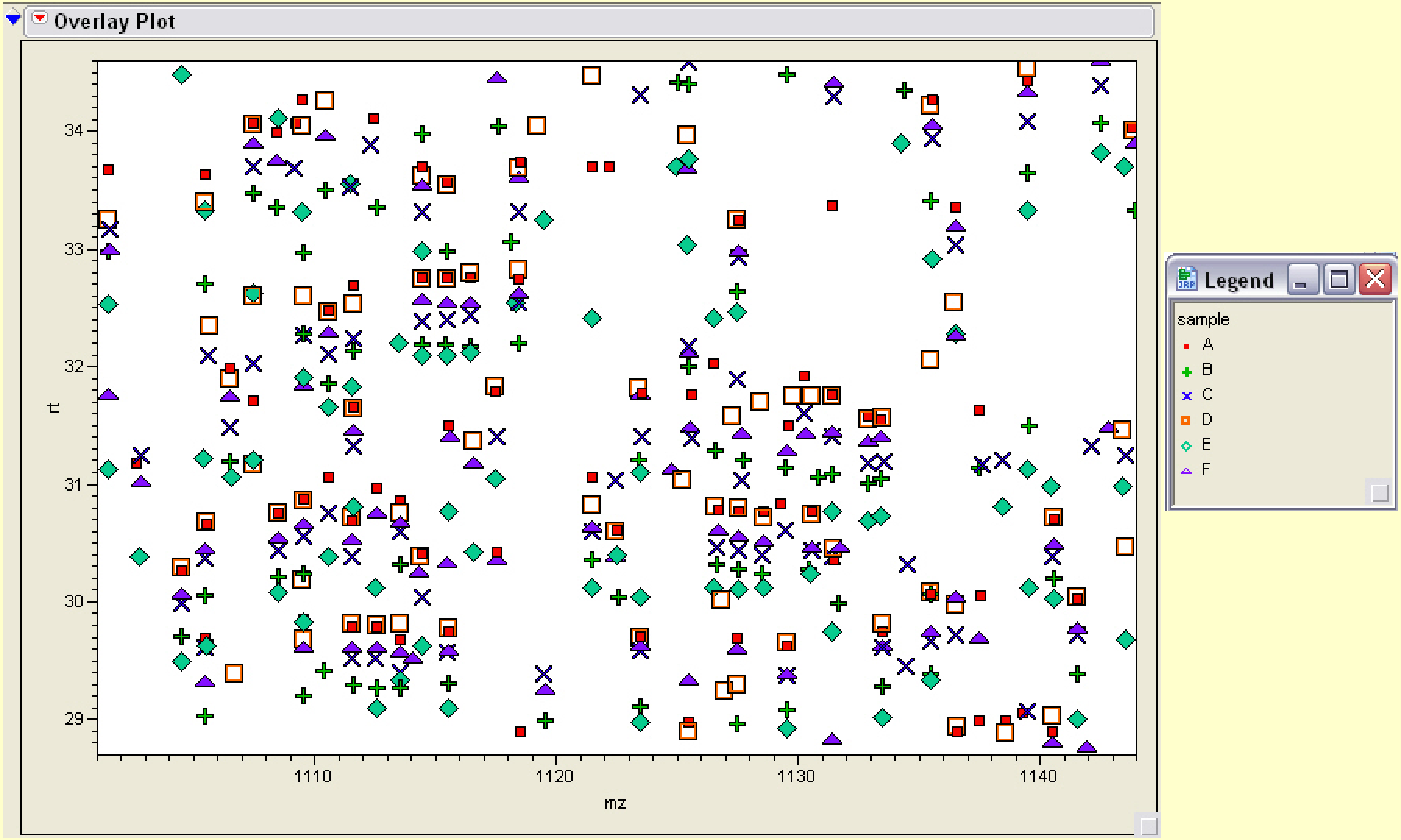
Green: Cancer, Red: Normal

Left: Green in Front, Right: Red in Front

1/10 of Data from 1 LC/MS Sample



LC/MS Alignment/Clustering Problem



Recent Proteomics Review Paper

Listgarten, J. and Emili, A. (2005) Statistical and Computational Methods for Comparative Proteomic Profiling Using Liquid Chromatography-Tandem Mass Spectrometry, *Molecular & Cellular Proteomics 4.4,* 419-434.

eQTL

- Expression Quantitative Trait Loci analysis
- View gene or protein expression as a quantitative trait, and map it onto the genome.
- Computational complexity increases. e.g. 100K+ expression measurements by 500K SNPs
- Basic idea: Use K-means clustering for dimension reduction