Microarray Data Mining: Puce a ADN Recent Developments

Gregory Piatetsky-Shapiro

KDnuggets

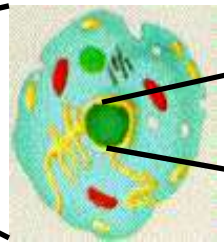
EGC 2005, Paris

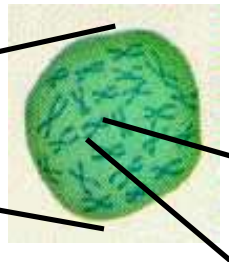
Role of Gene Expression

Cell

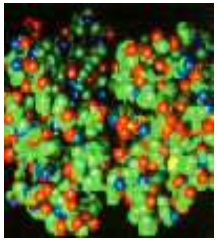
Nucleus

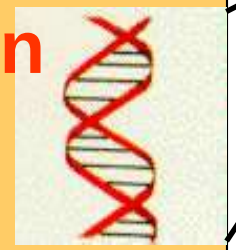
Chromosome





Gene expression





Protein

Gene (mRNA), single strand

Gene (DNA)

Graphics courtesy of the National Human Genome Research Institute

Gene Expression

- Cells are different because of **differential gene expression.**
- About 40% of human genes are expressed at one time.
- Gene is expressed by **transcribing** DNA exons into single-stranded mRNA
- mRNA is later **translated** into a protein
- Microarrays measure the level of mRNA expression

Gene Expression Measurement

- mRNA expression represents dynamic aspects of cell
- mRNA expression can be measured with latest technology
- mRNA is isolated and labeled with fluorescent protein
- mRNA is hybridized to the target; level of hybridization corresponds to light emission which is measured with a laser

Affymetrix DNA Microarrays

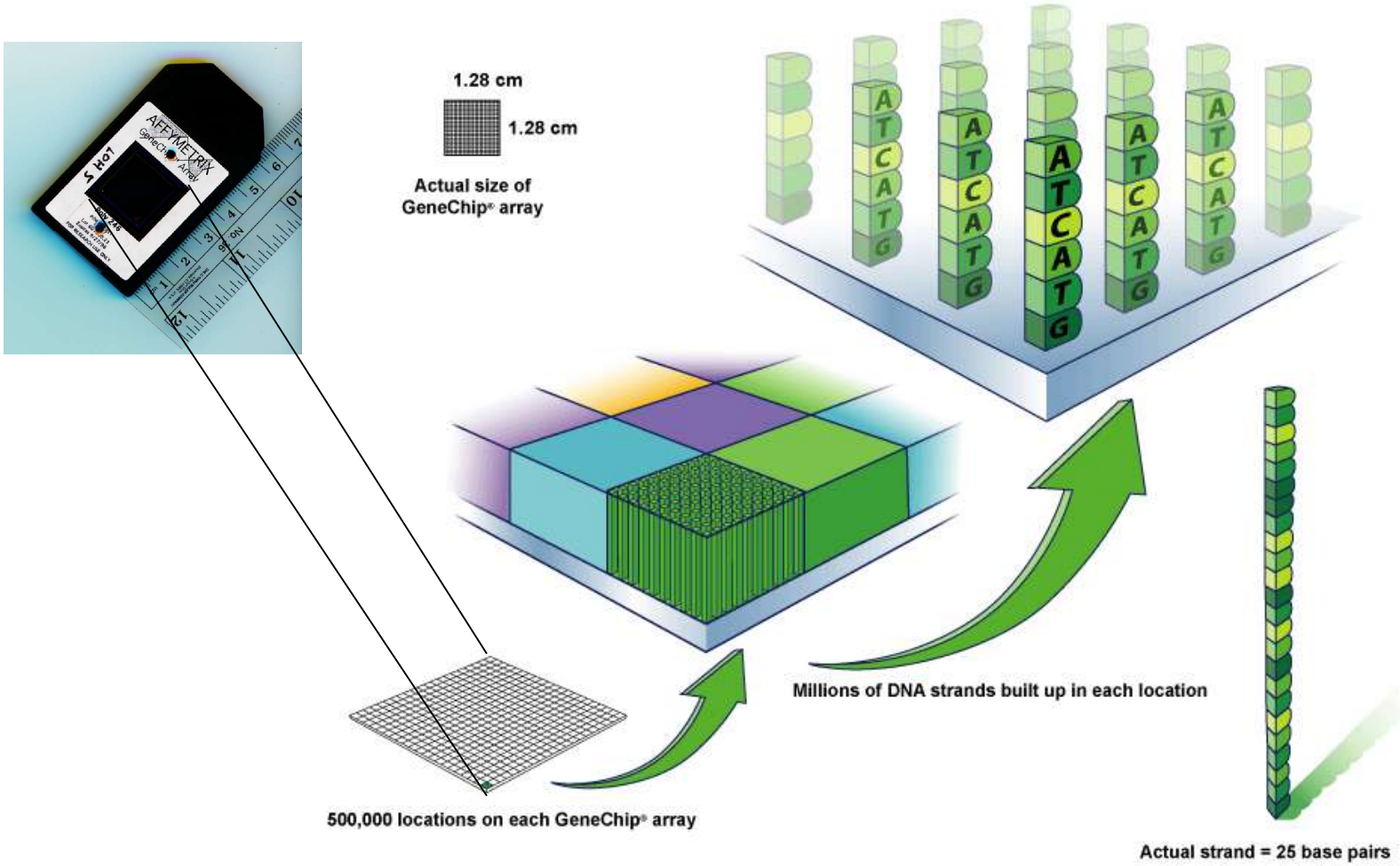
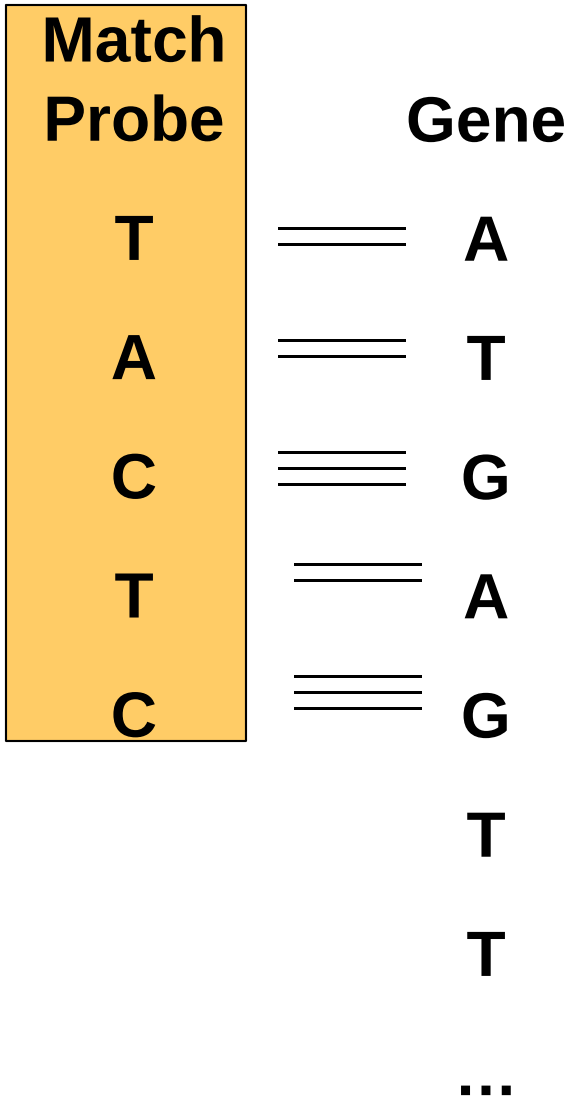
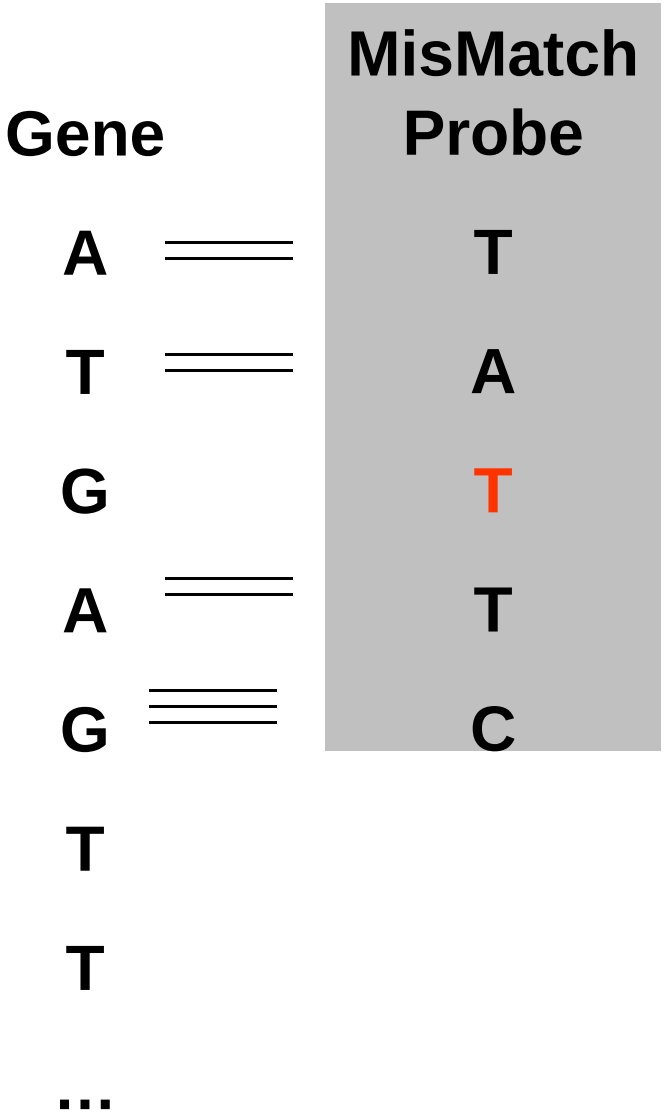


image courtesy of Affymetrix

Some Probes Match

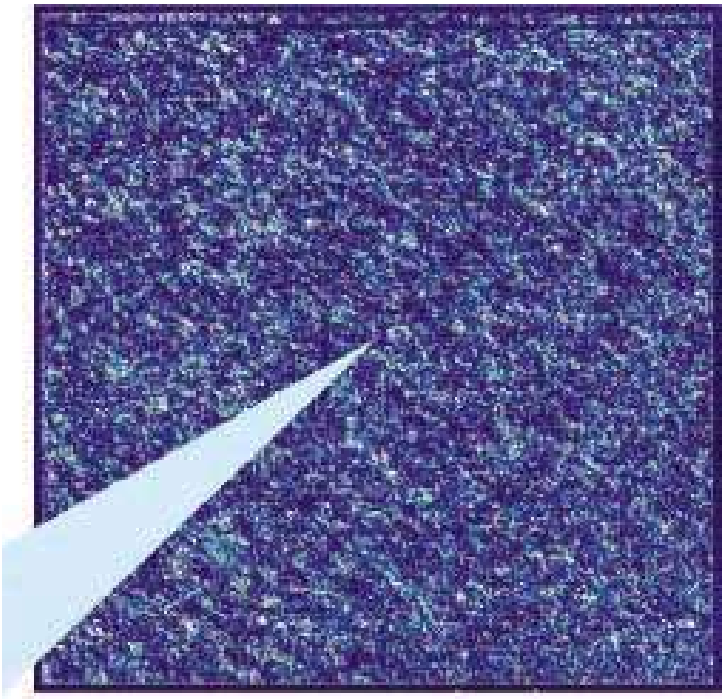


Other Probes MisMatch



Affymetrix Microarray Concept

1. mRNA segments tagged with fluorescent chemical
2. Matches with complementary probes
3. Fluorescence measured with laser



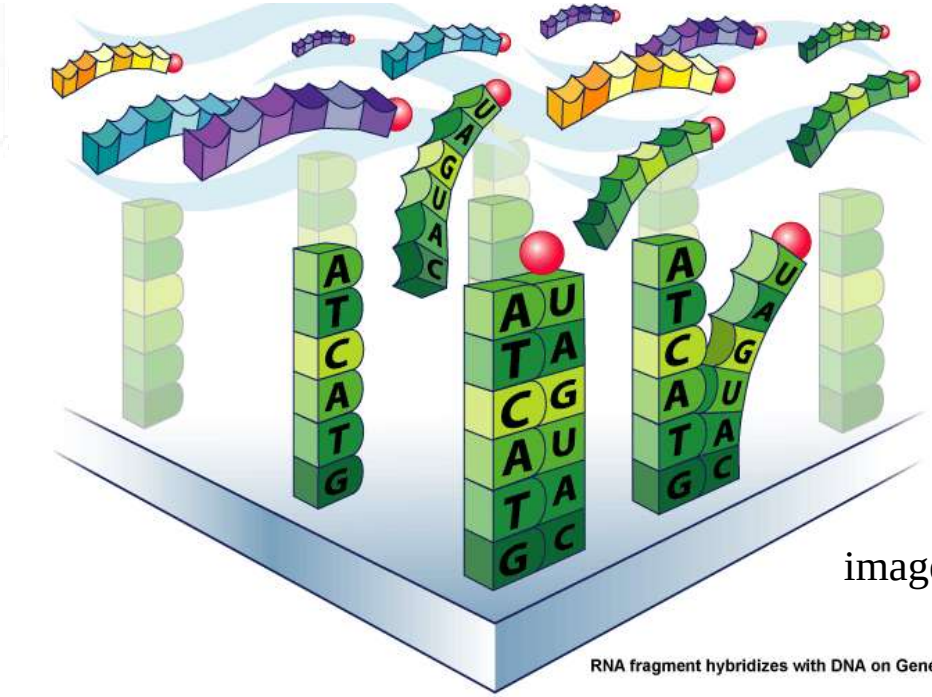
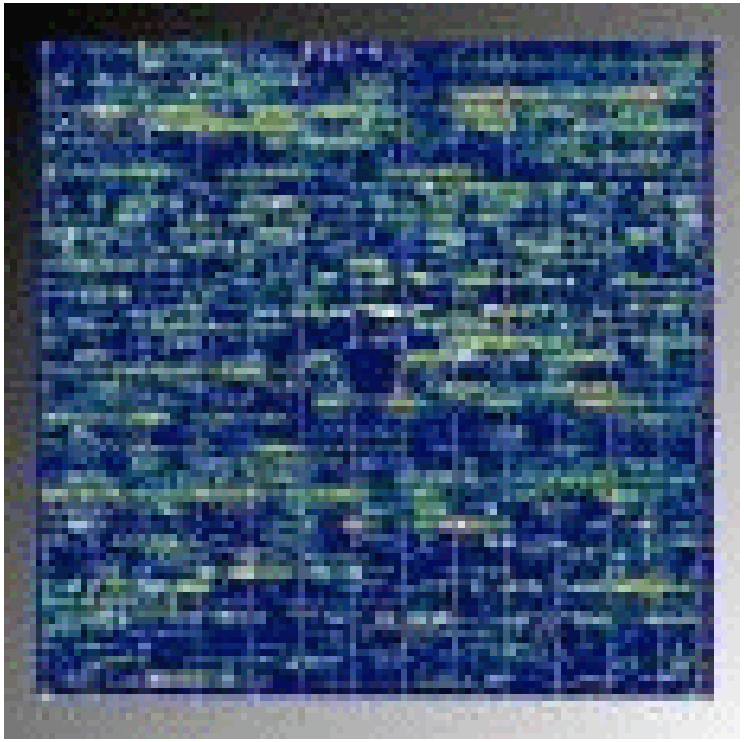


image courtesy of Affymetrix

Affymetrix Microarray Raw Image



enlarged section of raw image

Scanner

Gene	Value
D26528_at	193
D26561_cds1_at	-70
D26561_cds2_at	144
D26561_cds3_at	33
D26579_at	318
D26598_at	1764
D26599_at	1537
D26600_at	1204
D28114_at	707

raw data

Microarray Potential Applications

- New and better molecular diagnostics
  - Jan 11, 2005: FDA approved Roche Diagnostic AmpliChip, based on Affymetrix technology
- New molecular targets for therapy
  - few new drugs, large pipeline, …
- Improved treatment outcome
  - Partially depends on genetic signature
- Fundamental Biological Discovery
  - finding and refining biological pathways
- Personalized medicine ?!

Microarray Data Analysis Types

- Gene Selection
  - Find genes for therapeutic targets (new drugs)
- **Classification (Supervised)**
  - **Identify disease**
  - **Predict outcome / select best treatment**
- Clustering (Unsupervised)
  - Find new biological classes / refine existing ones
- Discovery of Associations and Pathways

Outline

- DNA and Biology
- **Microarray Classification - Best practices**
- Gene Set Analysis
- Synthetic microarray data sets

Microarray Data Analysis Challenges

- Biological, Process, and Other variation
- Model needs to be explainable to biologists
- Few records (samples), usually < 100
- Many columns (genes), usually > 1,000
  - This is very likely to result in **false positives, "discoveries" due to random noise**

Data Mining from Small Data: Example

- Les Américains boivent **beaucoup de vin** et ont **beaucoup** de maladie de coeur
- Les Francais boivent **beaucoup de vin** et ont **peu** de maladie de coeur
- Les Anglais boivent **beaucoup de biere** et ont **beaucoup** de maladie de coeur
- Les Allemands boivent **beaucoup de biere** et ont **peu** de maladie de coeur

Conclusion

- Vous pouvez boire tout que vous vouliez

- **C’est de parler en anglais que cause la maladie de coeur**
- ☺

Same Data, Different Results – Who is Right?

- Many examples (e.g. CAMDA conferences) where researchers analyzed the same data, and found different results and gene sets.
  - Who is right?
- Good methodology is needed
  - for avoiding errors
  - for best results

Capturing Best Practices for Microarray Data Analysis

- Worked with SPSS and S. Ramaswamy (MIT / Whitehead Institute) to capture best practices.
- Implemented as SPSS Microarray CATs (Clementine Application Template).
- Also implemented using Weka and open-source software

Capturing Best Practice for Microarray Gene Expression Data Analysis, G. Piatetsky-Shapiro, T. Khabaza, S. Ramaswamy, Proceedings of KDD-2003

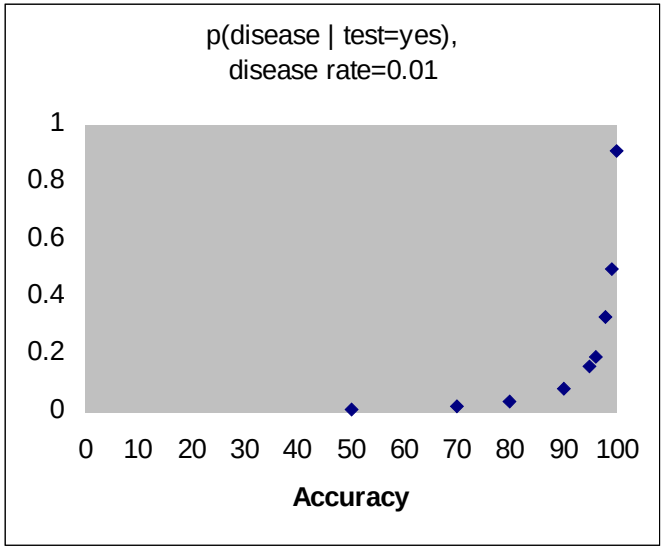
www.kdnuggets.com/gpspubs/

Best Practices

- Capture the complete process, from raw data to final results
- Gene (feature) selection inside cross-validation
- Randomization testing
- Robust classification algorithms
  - Simple methods give good results
  - Advanced methods can be better
- Wrapper approach for best gene subset selection
- Use bagging to improve accuracy
- Remove/relabel mislabeled or poorly differentiated samples

Observations

- Simplest approaches are most robust
- Advanced approaches can be more accurate
- "Small" increase in diagnostic accuracy (e.g. 90% to 98%) can greatly reduce rate of false positives (5-fold)



Microarray Classification Desired Features

- Robust in presence of false positives
- Stable under cross-validation
- Results understandable by biologists
- Return confidence/probability
- Fast enough

Popular Classification Methods

- Decision Trees/Rules
  - Find smallest gene sets, but not robust – poor performance
- Neural Nets - work well for reduced number of genes
- K-nearest neighbor – good results for small number of genes, but no model
- Naïve Bayes – simple, robust, but ignores gene interactions
- Support Vector Machines (SVM)
  - Good accuracy, does own gene selection,
    but hard to understand
- Specialized methods, D/S/A (Dudoit), …

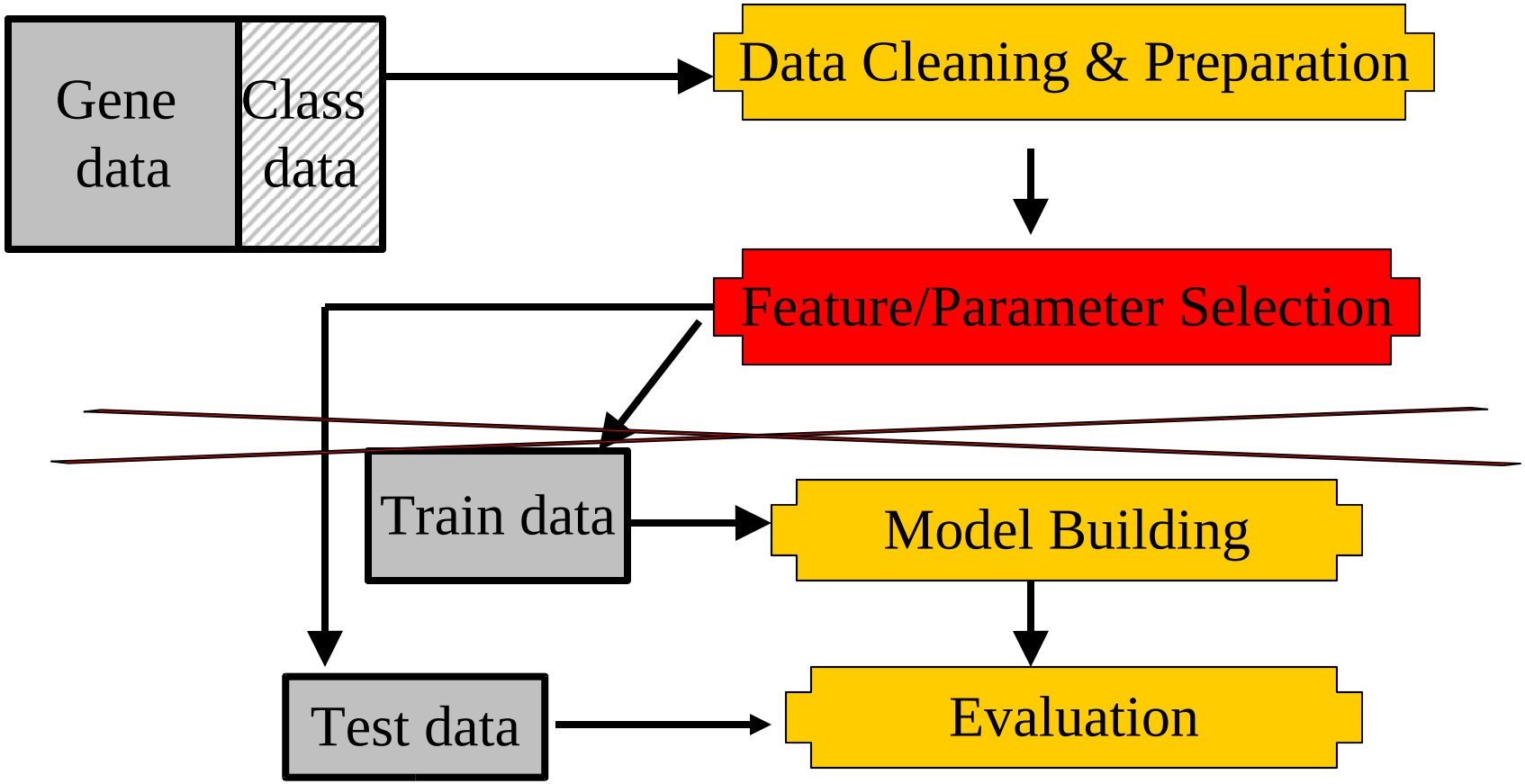
Gene Reduction Improves Classification Accuracy - Why?

- Most learning algorithms look for non-linear combinations of features
  - Can easily find ***spurious*** combinations given few records and many genes – "false positives problem"
- Classification accuracy improves if we first reduce number of genes by a linear method
  - e.g. T-values of mean difference

Feature selection approach

- Rank genes by measure & select top 100-200
- T-test for Mean Difference= $\frac{(Avg_1 - Avg_2)}{\sqrt{(\quad {}_1^2/N_1 + \quad {}_2^2/N_2)}}$
- Signal to Noise (S2N) = $\frac{(Avg_1 - Avg_2)}{(\quad {}_1 + \quad {}_2)}$
- Other methods …

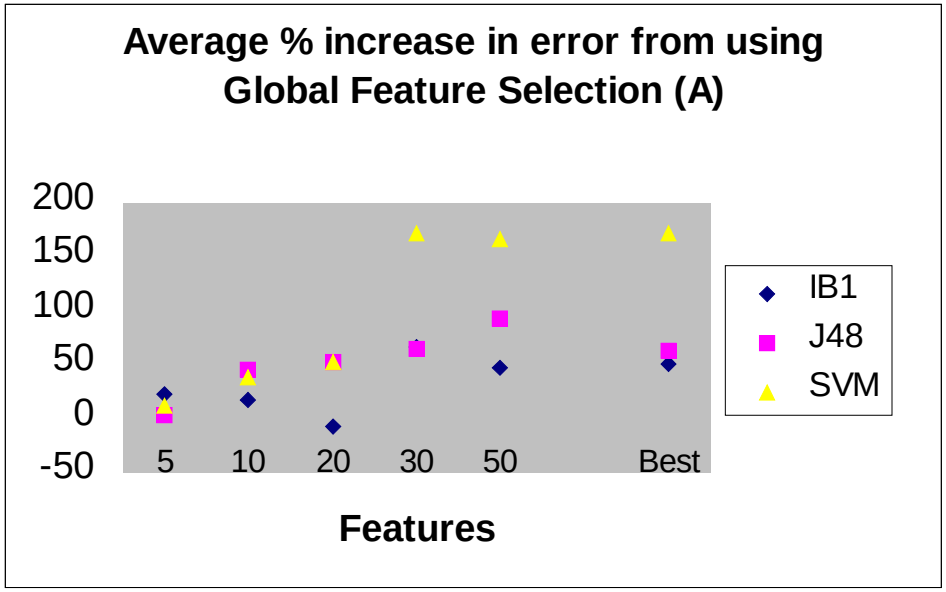
Global Feature / Parameter Selection is wrong



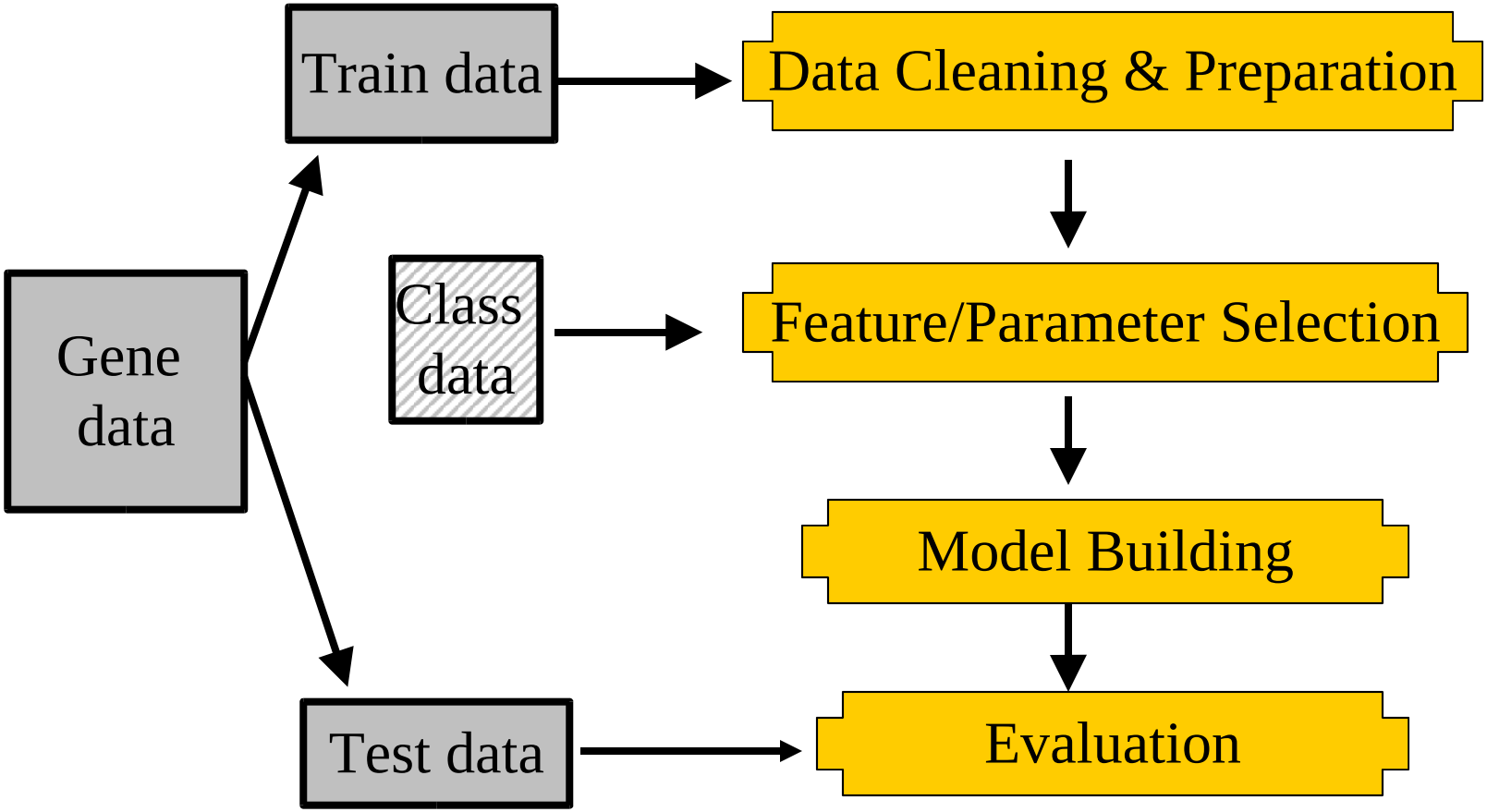
Because the information is “leaked” via gene selection.
Leads to overly “optimistic” classification results.

Global Feature Selection Bias

- Example: Data w 6000 features, ~ 100 samples
- Used 3 Weka algorithms: SVM, IB1, J48
- Error with Global feature selection was 50% to 150% lower than using X-validation



Microarray Classification Process

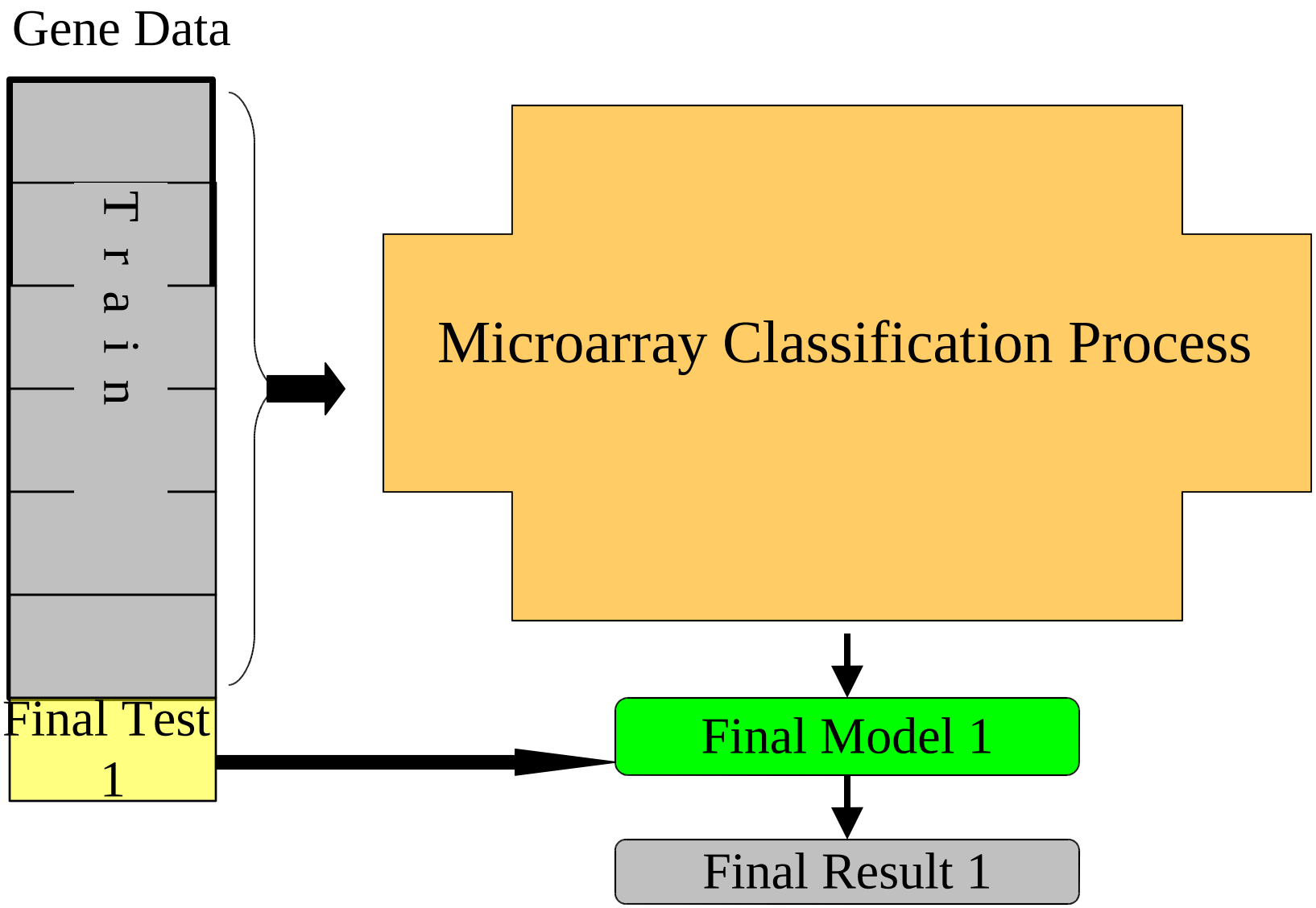


Evaluation

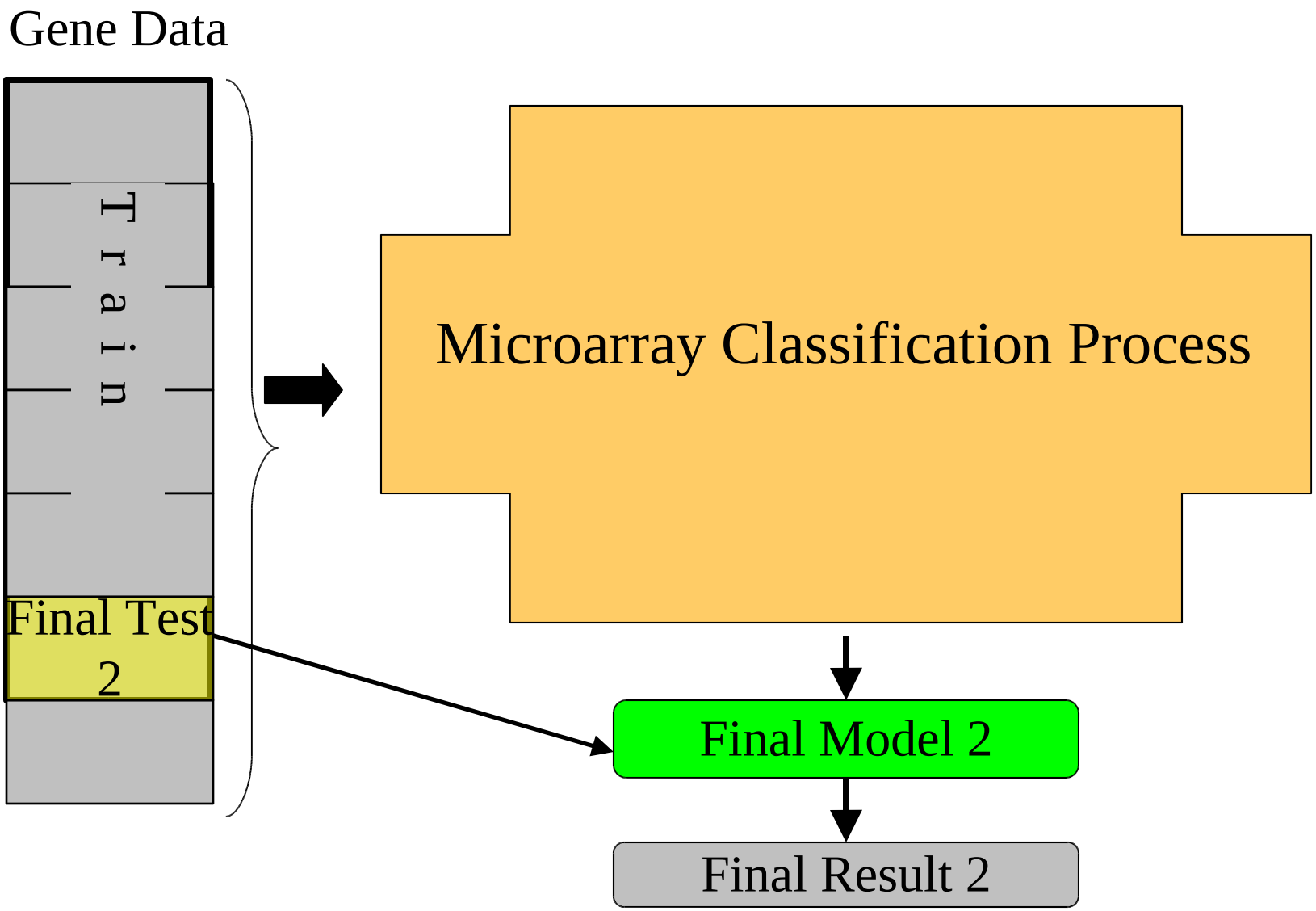
- Evaluation on one test dataset is not accurate (for small datasets)

- Evaluation, including feature/parameter selection needs to be inside cross-validation (**Validation Croisee)**

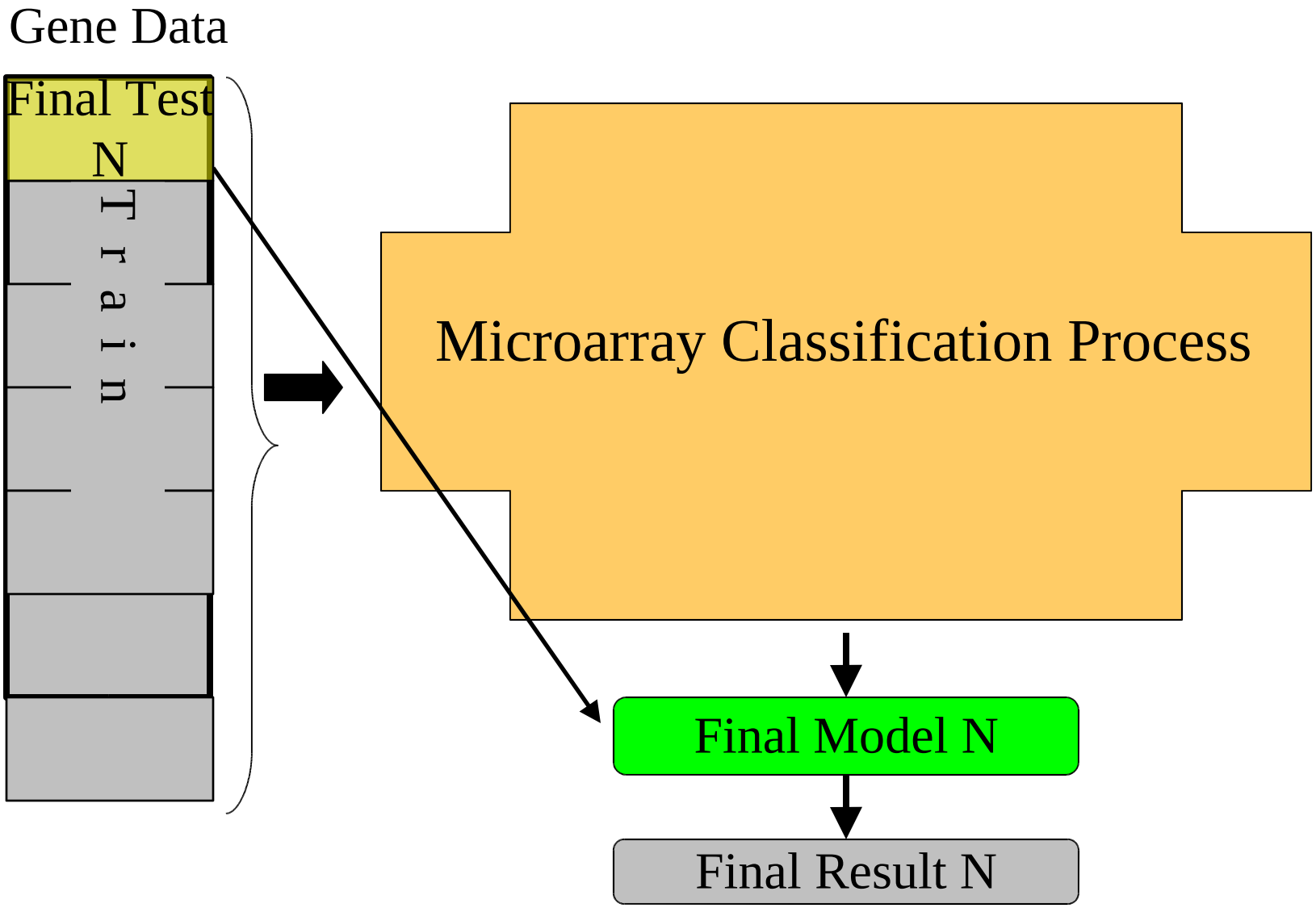
Evaluation inside X-validation



Evaluation inside X-validation, 2



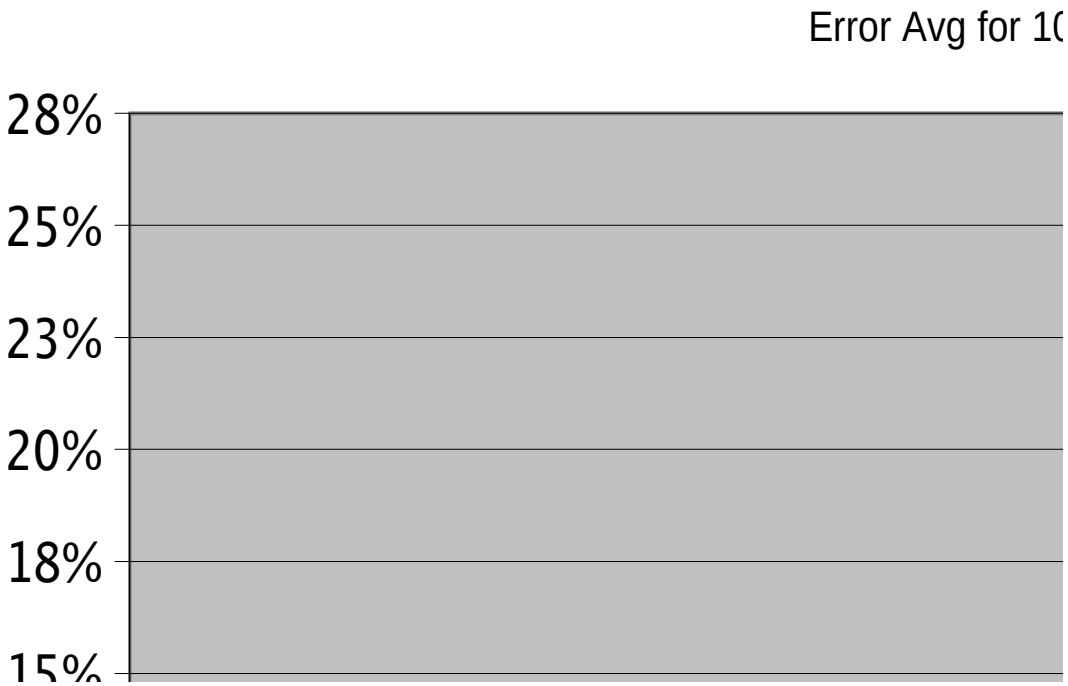
Evaluation inside X-validation, N



Iterative Wrapper approach to selecting the best gene set

- Model with top 100 genes is not optimal
- Test models using 1,2,3, …, 10, 20, 30, 40, ..., N top genes with cross-validation.
- Gene selection:
  - Simple: equal number of genes from each class
  - **advanced: best number from each class**
- For "randomized" algorithms (e.g. neural nets), average 10 cross-validation runs

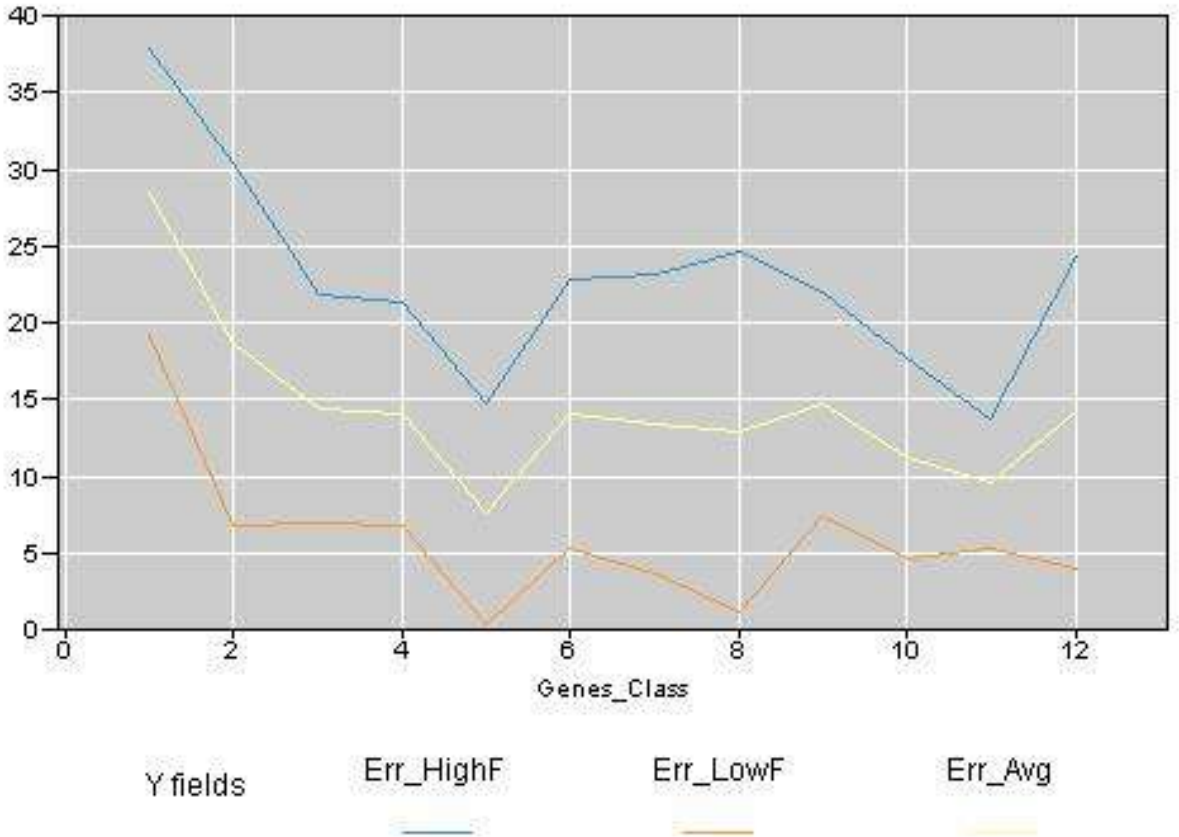
Best Gene Set: one X-validation run



Single Cross-Validation run

Best Gene Set
10 X-validation runs

- Select gene set with lowes combined Error
- good, but not optimal!



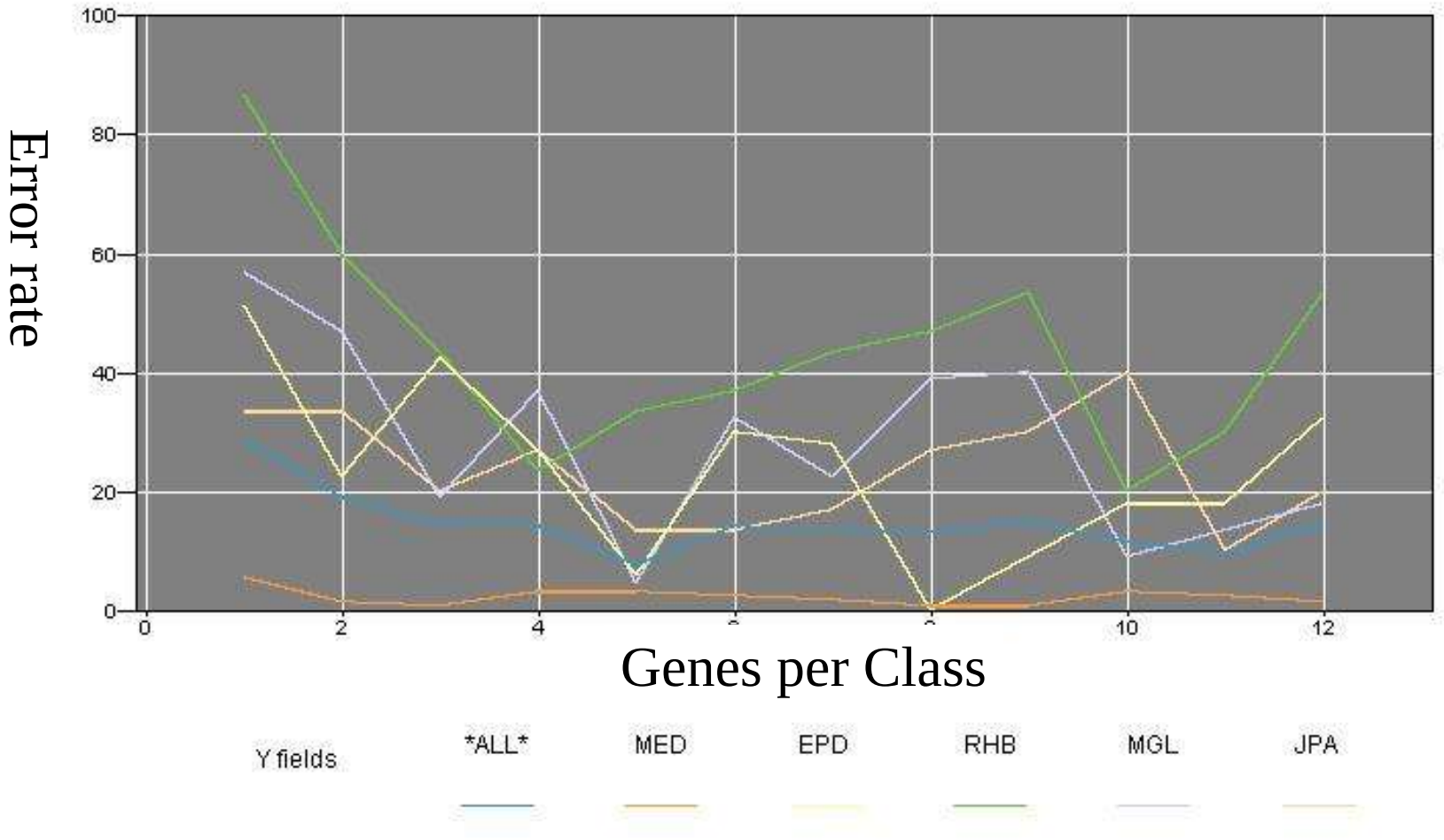
Average, high and low error rate for all classes

Multi-class classification

- Simple: One model for all classes

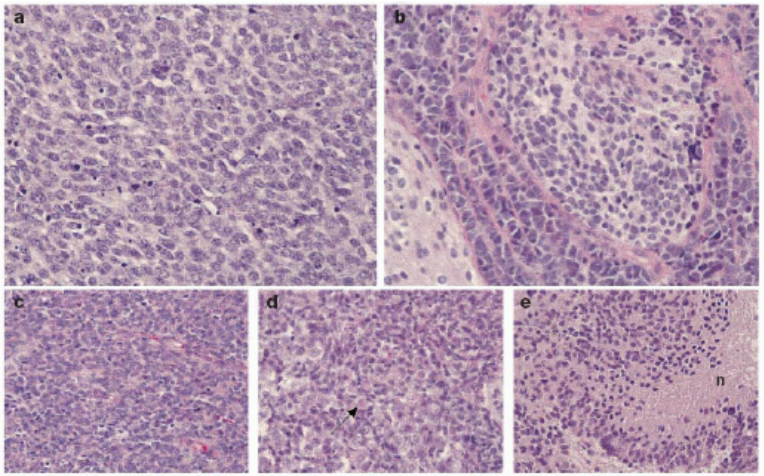
- Advanced: Separate model for each class

Error rates for each class



Example: Pediatric Brain Tumor Data

- 92 samples, 5 classes (MED, EPD, JPA, MGL, RHB) from U. of Chicago Children's Hospital
- Photomicrographs of tumours (400x)



Single Neural Net

Class	1-Net Error Rate
MED	2.1%
MGL	17%
RHB	24%
EPD	9%
JPA	19%
ALL	8.3%

Bagging Improves Results

- Bagging or simple voting of N different neural nets improves the accuracy

Bagging 100 Networks

Class	1-Net Error Rate	Bag Error rate	Bag Avg Conf
MED	2.1%	2% (0)*	98%
MGL	17%	10%	83%
RHB	24%	11%	76%
EPD	9%	0	91%
JPA	19%	0	81%
ALL	8.3%	3% (2)*	92%

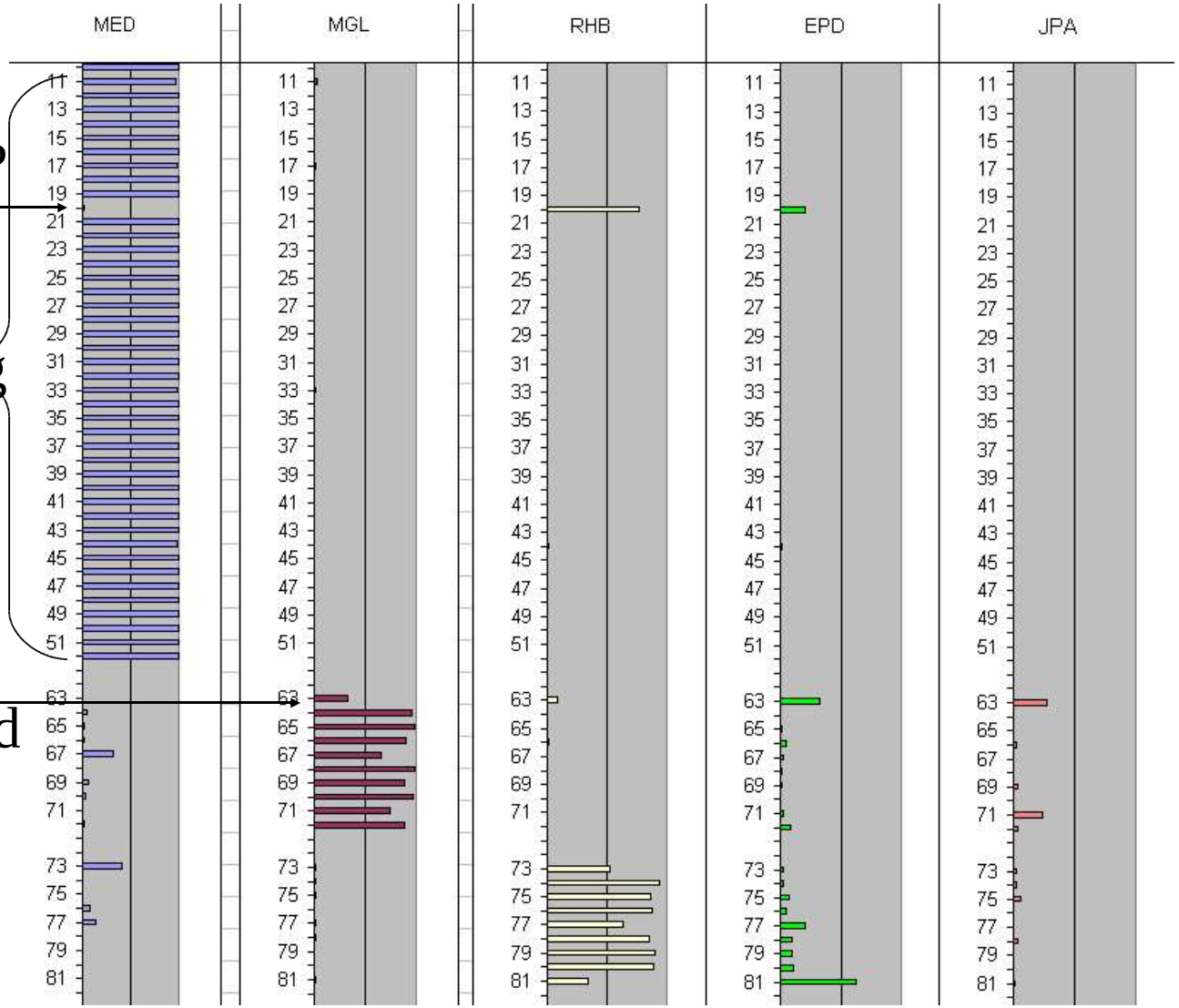
- Note: suspected error on one sample (labeled as MED but consistently classified as RHB)

Cross-validated prediction strength

misclassified?

MED: Strong predictions

poorly differentiated



Outline

- DNA and Biology
- **Microarray Classification - Best practices**
- **Gene Set Analysis**
- Synthetic microarray data sets

Gene Sets – Key to Power

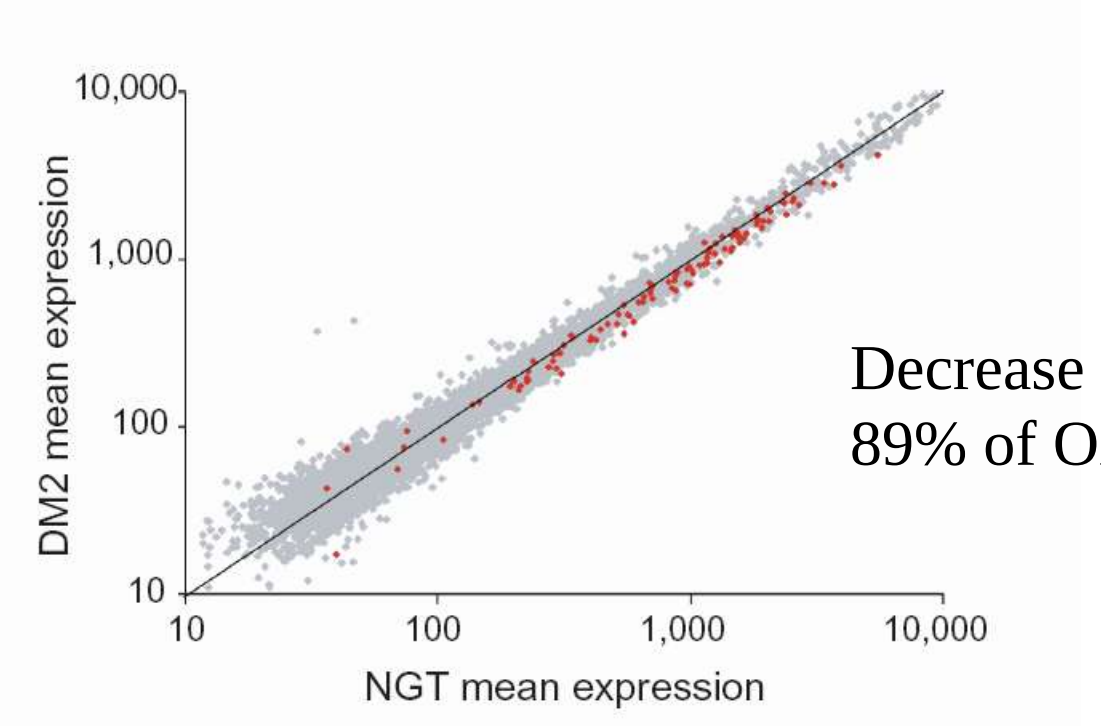
- Number of genes is a problem for single gene analysis
- Analyzing gene sets increases statistical power
- Group genes statistically (prototypes, correlations)
- Group genes using biological knowledge (pathways, Gene Ontology, medical text, …)

Gene Set Analysis

- Gene Set Enrichment
  - Mootha et al, Nature Genetics, July 2003
- Question: Genes involved in oxidative phosphorylation (~100) vs. diabetes ?

- No genes in OXPHOS group are individually significant (avg. ~20% decrease)

OXPHOS Gene set example



Decrease is consistent for 89% of OXPHOS genes

Mean expression of all genes (gray) and of OXPHOS genes (red) for people with DM2 (Type 2 diabetes mellitus) vs Normal.

Outline

- DNA and Biology
- Microarray Classification - Best practices
- Gene Set Analysis
- **Synthetic microarray data sets**

Synthetic Microarray Data

- Getting high accuracy is very important
- Current data has too few samples; the "true" labels may be unknown.
- **What methods are best for given data and what is their estimated accuracy?**
- Proposal: generate synthetic but realistic microarray data with KNOWN labels to evaluate algorithms under different conditions

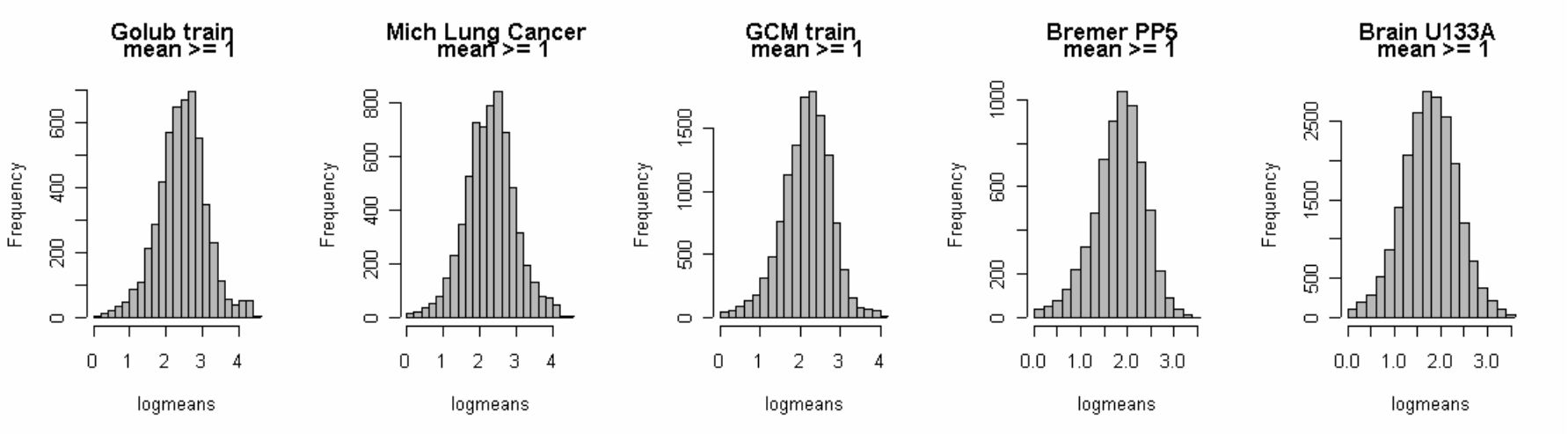
First Step: Analysis of Existing Microarray Datasets

Dataset	Platform, Software	Tissue	Genes	Samples	Classes
Golub train	Affy HuGeneFL; MAS4	Blood	7070	38	2
Golub test	Affy HuGeneFL; MAS4	Blood	7070	34	2
Mich Lung	Affy HuGeneFL	Lung	7070	96	3
GCM train	Affy Hu6800 and Hu35KsubA	Many	16004	144	14
GCM normal	Affy Hu6800 and Hu35KsubA	Many	16004	90	13
Bremer pp5	Affy HuGeneFL; Probe Profiler	Brain	7070	92	5
Brain U133	Affy U133; Probe Profiler	Brain	22215	33	4

Different Technologies, Tissues, Normal/Abnormal

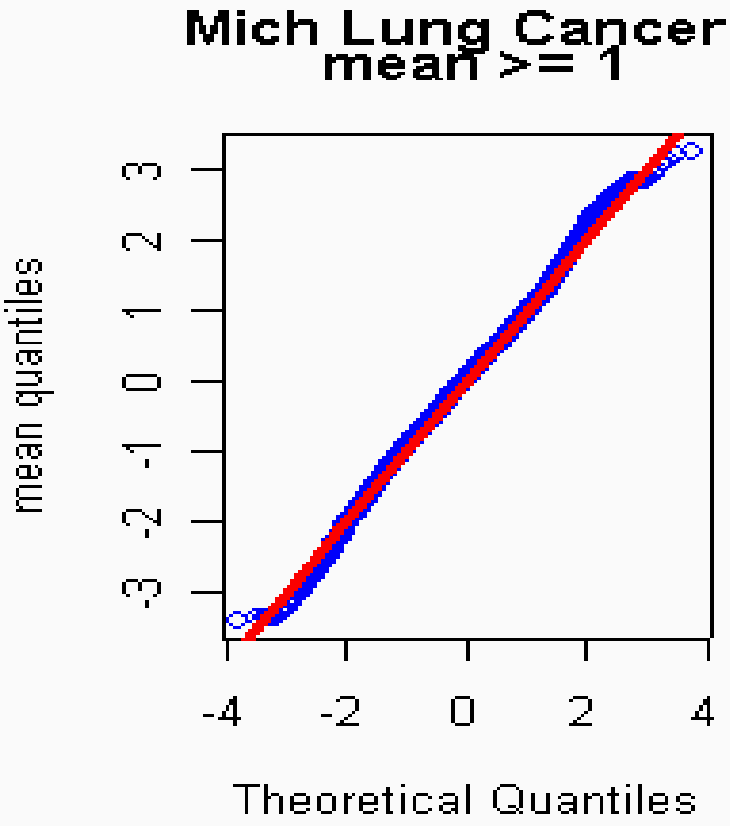
Global Distribution of Gene Means

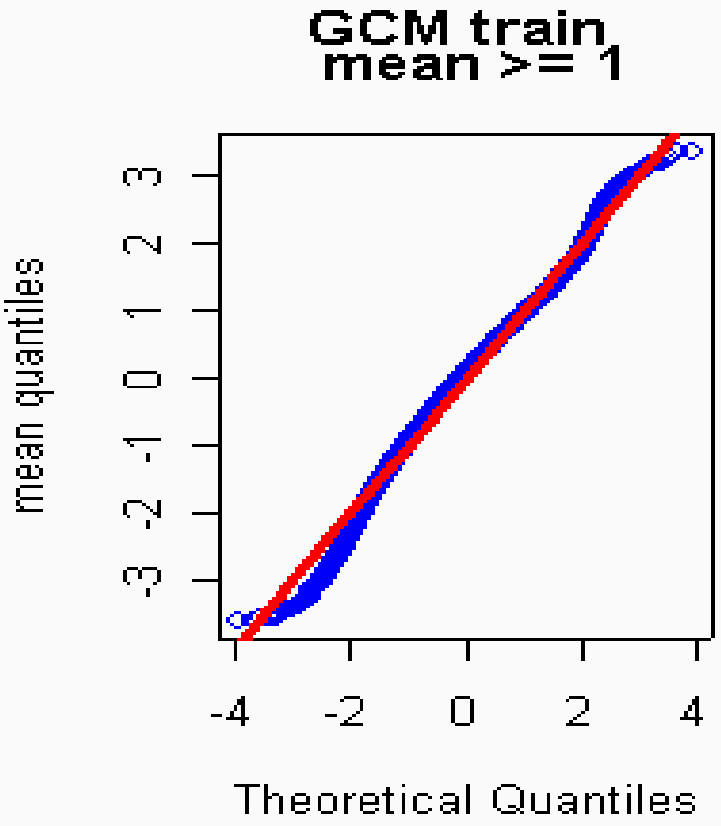
Analyzed means of ALL genes across samples
Excluded Mean < 1 (mismatch, random)



Found: log(means) ~ normal distribution
Gene means are globally logNormal

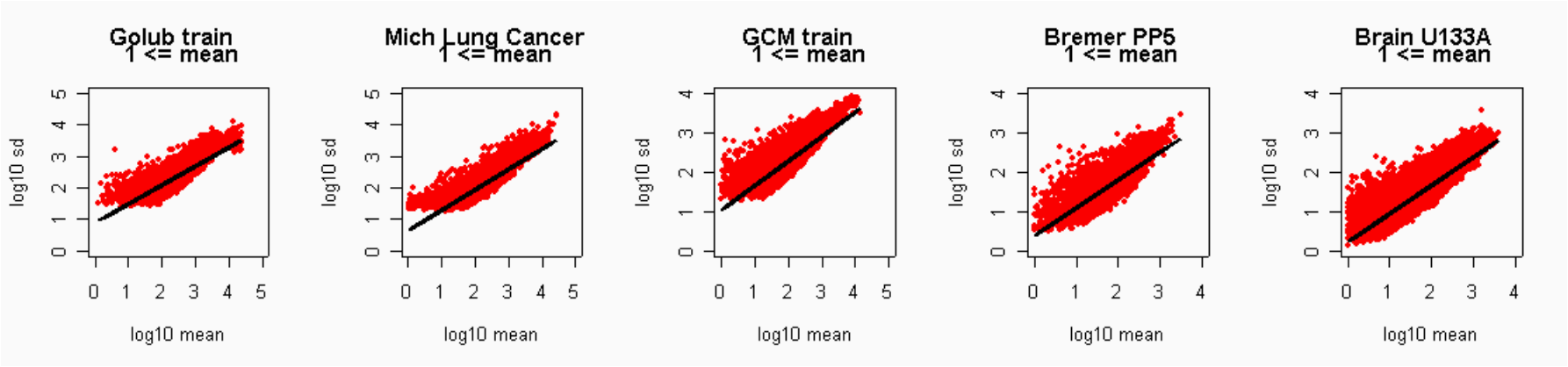
Global Gene Means (Q-Q plot)





Plot: log10(means) (blue) vs normal quantiles (red)

Gene Expression
Standard Deviation vs Mean

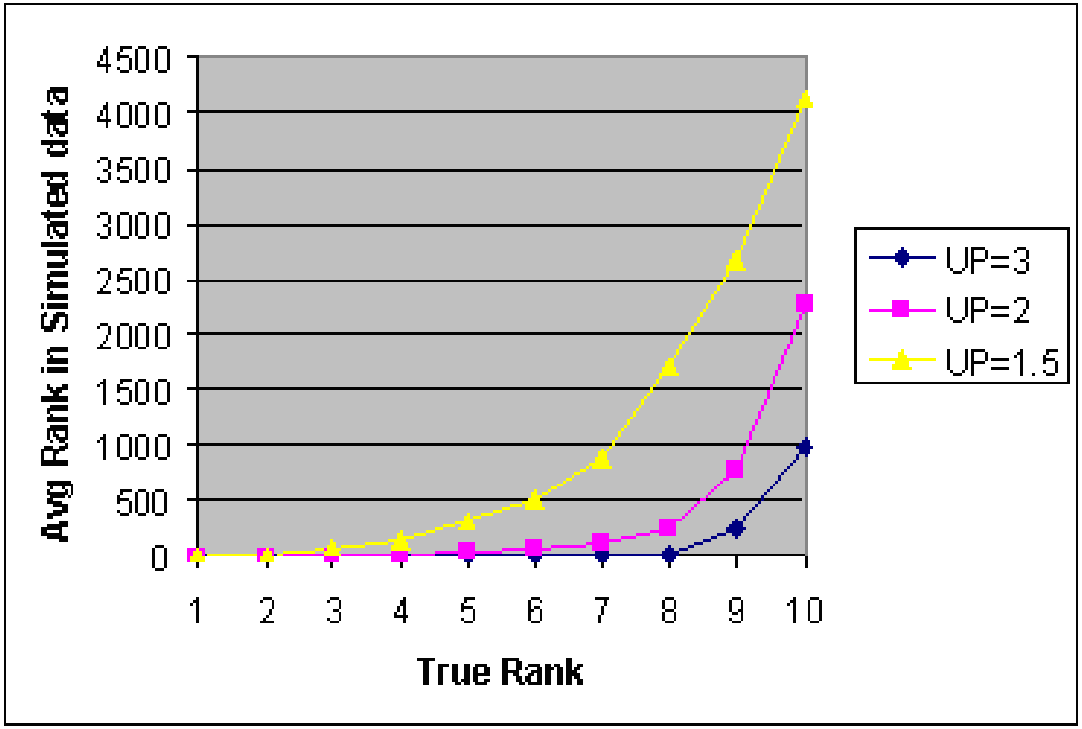


Globally

$$\log(SD) \sim a \log(means) + b$$

True Rank vs T-value Rank

- Simulated 100 microarray datasets, with 6000 genes and 30 samples (using R)
- First 10 genes up-regulated by UP factor



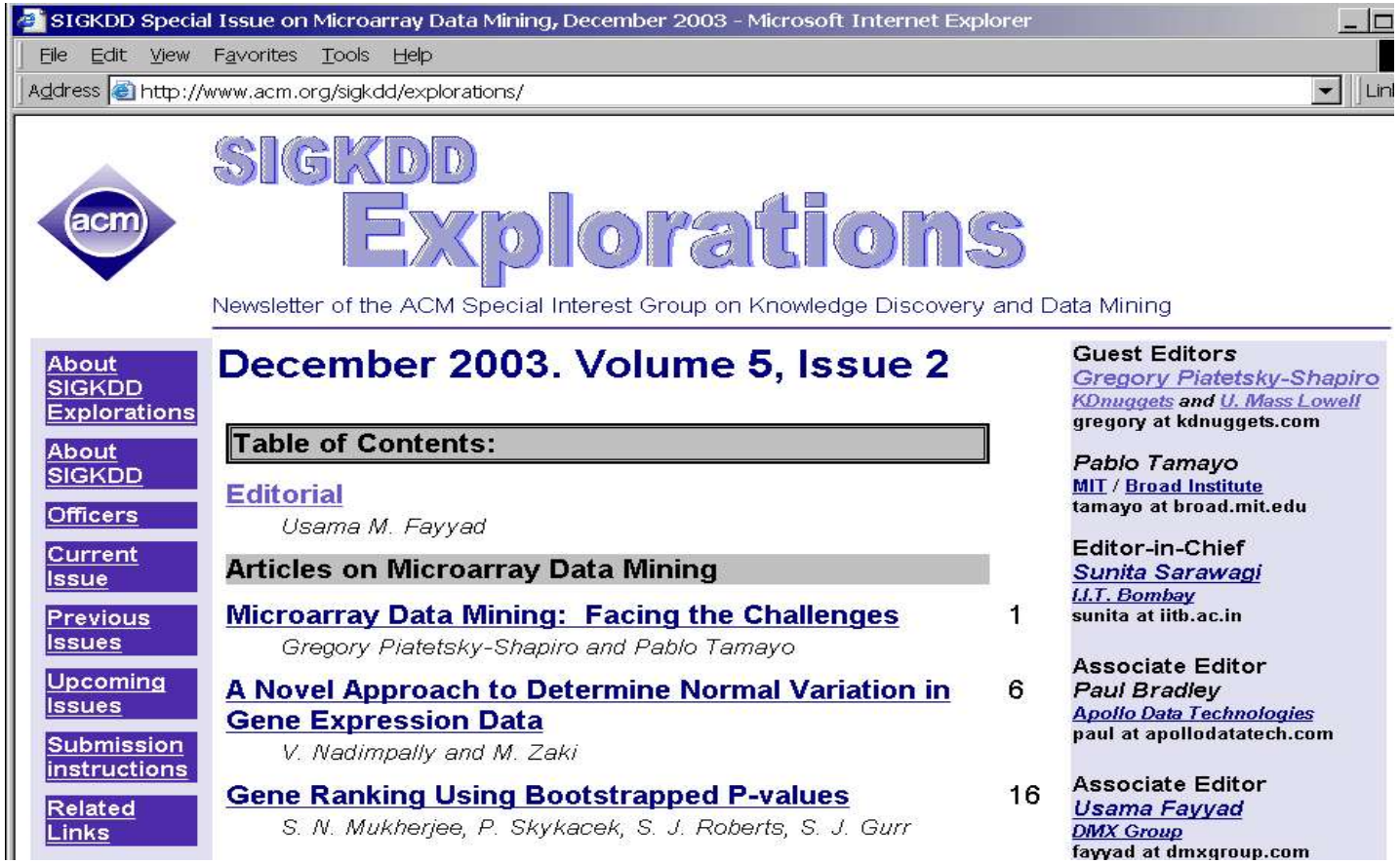
Future Research Directions

- Why is gene distribution log normal?
- Develop a synthetic microarray data generator
- Which gene selection methods are most robust and under what conditions?
- Evaluate algorithm accuracy for a given test set by creating similar synthetic sets with known answers

Acknowledgements

- Eric Bremer, Children's Hospital (Chicago) & Northwestern U.
- Tom Khabaza, SPSS
- Sridhar Ramaswamy, MIT/Broad Institute
- Pablo Tamayo, MIT/Broad Institute

SIGKDD Explorations Special Issue on Microarray Data Mining



www.acm.org/sigkdd/explorations/issue5-2.htm

Merci!

Further resources on Data Mining: www.KDnuggets.com

Microarrays:

www.KDnuggets.com/websites/microarray.html

Data Mining Course (one semester)

www.KDnuggets.com/dmcourse

Questions?