Challenges of Translating Gene Expression Microarray Data into Clinically Useful Tests

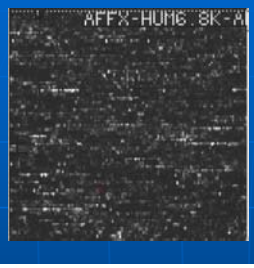
Lisa M. McShane, Ph.D.

Biometric Research Branch
Division of Cancer Treatment and Diagnosis
National Cancer Institute

April 27, 2009

1

Gene Expression Affymetrix Microarrays: Pre-process image data ⇒ gene-level data



(Figure 2 from Schadt *et al.*, *J. Cell. Biochem.* 2001)

MAS-GCOS-AGCC (Affymetrix)
- Anti-log of Tukey biweight average of adjusted log(PMij-IMij)
- Independent of other arrays

Li & Wong dChip – *PNAS* 2000, *Genome Biology* 2001
- MBEIi = θi or θi* estimated from PMij-MMij = θi φj+εij or PMij = vi + θi* φj´
- Computed across set of arrays

Irizarry et al – *Biostat & NAR* 2003
- RMAi = ei estimated from T(PMij) = ei + aj+εij
- Computed across set of arrays

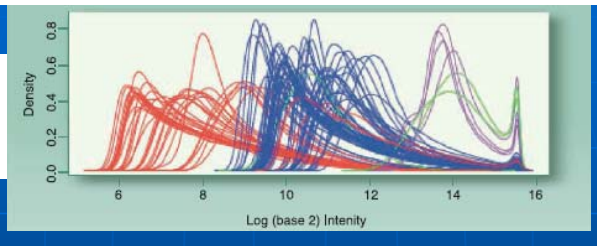
(Also GC-RMA – Wu et al – *JASA* 2004)

2

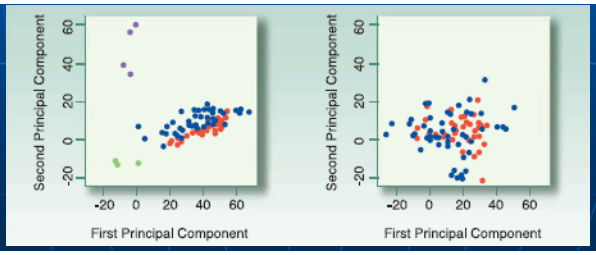
Batch Effects in Gene Expression Data

Density estimates of PM probe intensities (CEL files) for 96 NSCLC samples

Red = batch 1
Blue = batch 2
Purple &
Green = outliers?



PCA plots after RMA pre-processing with (left) and without (right) outlier CEL files



(Figure 1 from Owzar et al, *Clinical Cancer Research* 2008 using data from Beer et al., Nature Medicine 2002)

Typical Study Objectives

- Class Discovery (unsupervised)
  - Discover clusters among specimens or among genes
- Class Comparison (supervised)
  - For pre-specified classes, establish whether gene expression profiles differ overall, and identify genes responsible for differences
- Class Prediction (supervised)
  - Prediction of phenotype using information from gene expression profile

4

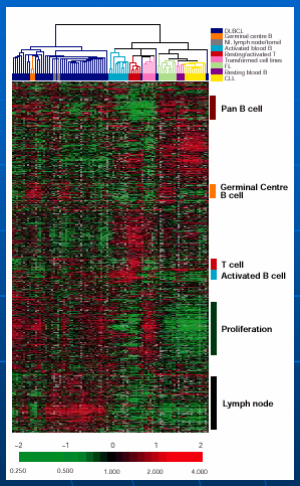
Class Discovery Examples

- Discover previously unrecognized subtypes of lymphoma.
- Cluster temporal gene expression patterns to gain insight into genetic regulation in response to a drug or toxin.

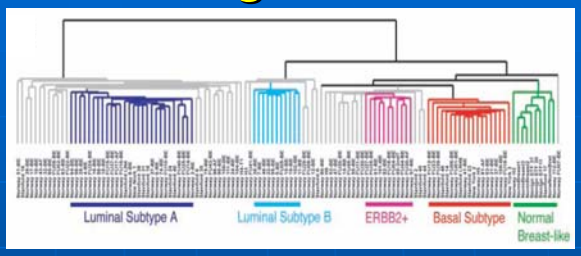
(Focus in this talk on finding subgroups of specimens rather than groups of genes.)

5

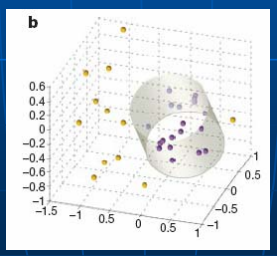
Clustering

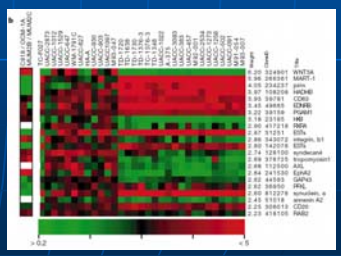


Alizadeh et al., *Nature*, 2000 (lymphoma)



Perou, *Nature* 2001; Sørlie, *PNAS* 2001; Sørlie, *PNAS* 2003 (breast cancer)





Bittner et al., *Nature*, 2000 (melanoma)

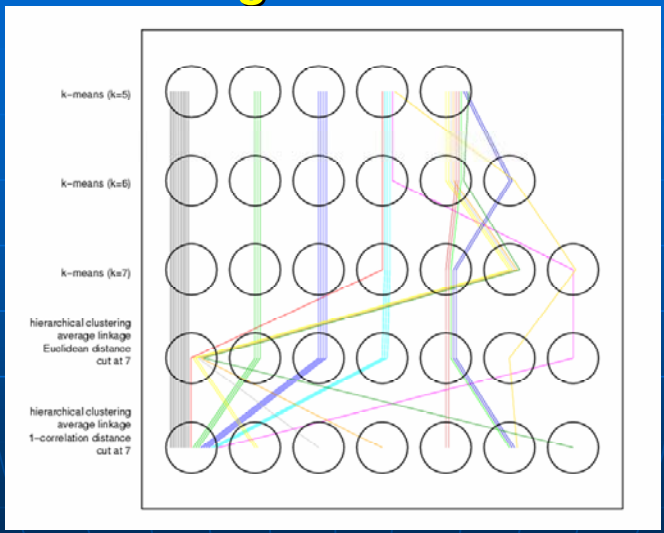
6

Interpretation of Cluster Analysis Results

- Cluster analyses always produce cluster structure
  - Where to "cut" the dendrogram?
  - Which clusters do we believe?
- Different clustering algorithms may find different structure using the same data

7

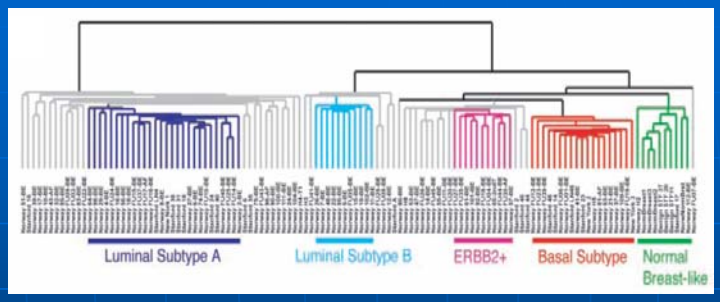
Clustering Method Matters



Bittner et al., *Nature,* 2000 (melanoma)

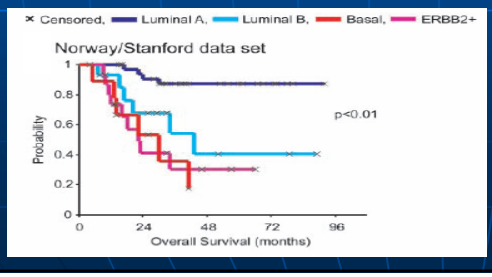
8

Clinical Utility of Clusters?



Sometimes you get lucky, but not optimal approach

122 breast cancer samples, ~500 "intrinsic" genes



Different management/ treatment for different groups?

(Perou, *Nature* 2001; Sørlie, *PNAS* 2001; Sørlie, *PNAS* 2003)

9

Class Comparison Examples

- Establish that expression profiles differ between two histologic subtypes of cancer
- Identify genes whose expression level differs between tumors that did versus did not respond to a drug

"Informative" genes are those that are differentially expressed between classes or that are statistically associated with some outcome of interest.

10

Informative Gene List Instability

- Multiple testing issues
  - 10,000 non-informative genes each tested at 0.05 level of significance will produce 500 false positives
  - Typically use smaller testing level (e.g., 0.001) or more sophisticated procedures such as FDR
- Size of list dependent on stringency of multiple testing corrections
- Low power under stringent multiple testing corrections
- Co-regulation of genes

11

Beyond Gene Lists: Classifier or Multivariate Score

- Link multiplex marker measurements to clinical outcome or characteristic
- Mathematical function that associates a specimen with a class or assigns a continuous score based on inputted feature measurements
- Most scores eventually subject to cutpoints for clinical decision-making

12

Class Prediction Examples

- Predict from expression profiles which patients are likely to experience severe toxicity from a new drug versus who will tolerate it well.
- Predict which glioma patients' tumors will be sensitive to a new agent versus resistant

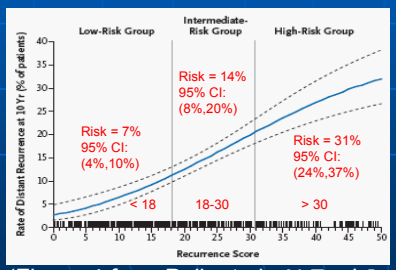
13

Gene Expression Profile

Link to clinical outcome

RISK SCORE or CLASSIFIER

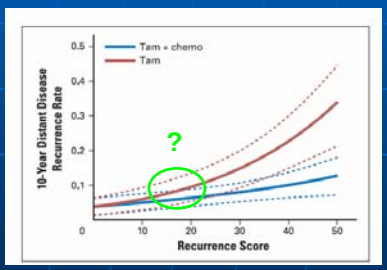
Clinical Test



(Figure 4 from Paik et al., *N Engl J Med*, Dec. 2004)

Oncotype DX is prognostic

Inform clinical decision



(Figure 4 from Paik et al., *JCO*, Aug. 2006)

TAILORx Trial: Is Oncotype DX predictive of chemo benefit?

14

Classification Methods

- Linear Predictor (for 2 classes)
  $L(\mathbf{x}) = w_1x_1 + w_2x_2 + \ldots + w_fx_f$
  is a weighted combination of important features to which a classification threshold is applied
  - Examples: Linear discriminant analysis, compound covariate predictor, weighted voting method, support vector machines with inner product kernel, perceptrons, naïve Bayes MVN mixture classifier
- Distance-based
  - Examples: Nearest neighbor, nearest centroid
- Generalizable to > 2 classes

(Simon, *Journal of Clinical Oncology* 2005)

15

Quantifying "How good is the classifier?"

- Estimate percent correct classifications ("classification accuracy")
- Survival differences or hazard ratios associated with classification (or with continuous risk score) of sufficient magnitude to be clinically meaningful
- Value added beyond standard clinico-pathologic factors

16

Classification: Avoiding Pitfalls

- When number of potential features is much larger than the number of cases, can always fit a classifier to have 100% prediction accuracy on data set used to build it
- Estimating accuracy by "plugging in" data used to select genes or build a classifier results in highly overoptimistic estimates of prediction accuracy
- Internal (e.g., cross-validation, bootstrapping) and external validation of classifier are essential ("training" and "test" sets)

17

Sample Size Considerations for Building a Classifier

- Sample size = number of cases, NOT number of genes measured
- Sample size determination for training set
  - Expected accuracy of resulting classifier is ± tolerance of (true) accuracy (Dobbin and Simon, *Biostatistics* 2007; Dobbin, Zhao and Simon, *Clin Cancer Res* 2008)
  - Few dozen to few hundred cases required depending on
    - Number of genes
    - Largest standardized fold change for informative genes
    - Prevalence of each class

18

There is no substitute for a well-designed, COMPLETELY INDEPENDENT validation study.

19

Steps to Validate Clinical Utility

- Achieve acceptable reproducibility of classification or score
  - Stringent component-wise reproducibility might not be necessary
  - Reference lab versus multiple labs
- COMPLETELY specify
  - Specimen acquisition and handling realistic for clinical use
  - Assay platform (e.g., reagents, chip, equipment)
  - Technical protocol, including quality criteria
  - Data pre-processing
  - Form of classifier or risk score, including cutpoints

20

Steps to Validate Clinical Utility

- Design prospective study
  - Patients representative of target population (e.g., age, stage)
  - Specific treatment context
  - Adequate sample size
- Pre-planned analysis to establish fitness for intended clinical use
  - Clinical outcome measure (e.g., overall survival, distant disease-free survival, tumor response)
  - Performance metrics
    - Percent accuracy
    - Survival curve separation
  - Adds value to existing indicators

21

Summary

- Considerable investment of time and resources to develop clinically useful tests
- Expertise required: clinical, laboratory, biology, statistics, computational
- Attention to clinical feasibility and affordability
- Clinical impact must be sufficiently high!

22

Acknowledgements

- Richard Simon
- Kevin Dobbin
- Lara Lusa
- Members of the Biometric Research Branch at NCI
- Members of the Cancer Diagnosis Program at NCI

23