# Identification of Prognostic Genes, Pooling Information Across Different Studies and Oligonucleotide Arrays

Jeffrey S. Morris

UT MD Anderson Cancer Center

Department of Biostatistics and Applied Mathematics

# Introduction: CAMDA

- "Critical Assessment of Microarray Data Analysis"
  - 4 lung cancer data sets (most NSCLC, ADC)
  - Clinical and microarray data, similar questions addressed
  - **Challenge**: Integrate information across studies to yield insights into cancer biology.
- The papers:
  - Harvard: (Battacharjee, et al. *PNAS* 2001)
    - 186 patients (139 ADC), Affymetrix arrays
  - Michigan: (Beer, et al. *Nature Med.* 2002)
    - 86 patients (all ADC), Affymetrix arrays
  - Stanford: (Garber, et al. *PNAS* 2001)
    - 62 patients (35 ADC), Glass arrays
  - Ontario: (Wigle, et al. *Cancer Res.* 2002)
    - 39 patients (19 ADC), Glass arrays

# Introduction

- Our Focus: ADC patients w/ surv. data
  - Michigan – 86 patients, Harvard -- 125
- GOALS:
  1) Pool information across studies to identify prognostic genes in lung ADC.
     - Offer prognostic information above and beyond that offered by known clinical predictors (e.g. stage)
  2) Develop methodology for combining gene expression information across different oligonucleotide chip types

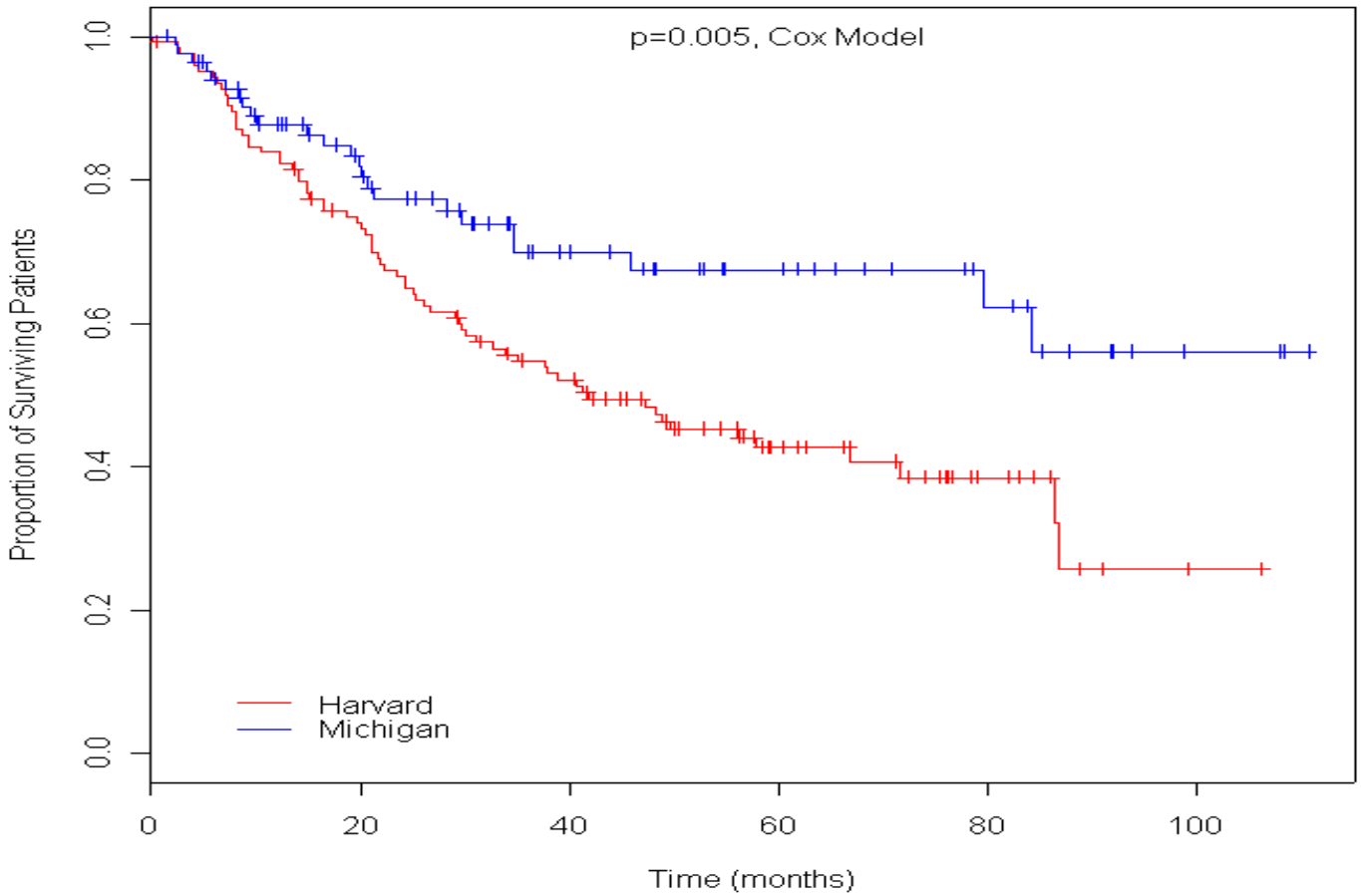
# Outline

- Details of our Analytical Methods
  - Pooling Information across Studies
  - Pooling Information across Chip Types
  - Preprocessing
  - Identifying Prognostic Genes
- Results and Interpretation
- Conclusions

# Pooling Information across Studies

- Problem: Study-to-study heterogeneity
  - Study populations or conditions may not be comparable.
- Our data:
  - Nearly identical gender/age/stage/smoking
  - Similar follow-up time distributions
  - Different survival distributions, even after adjusting for available covariates

# Pooling Information across Studies




Harvard patients – worse prognosis

# Pooling Information across Studies

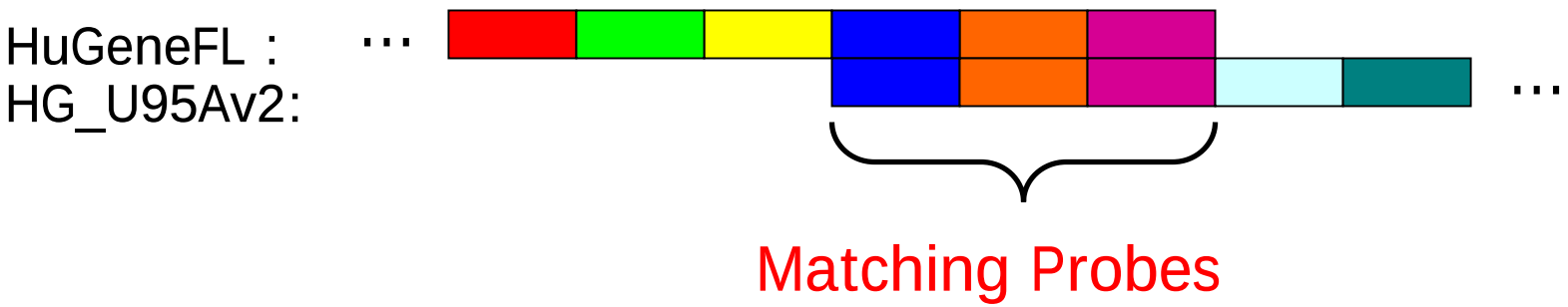
- Meta-analysis approaches in literature:
  - Bayesian Hierarchical Models
    - See e.g. Stangl (1996)
  - Frailty Models
    - See e.g. Therneau and Grambsch(2000), Ch9
- Our approach: Include fixed offset for institution in survival models

# Pooling Information across Chip Types

- Michigan: HuGeneFl Chip
  - 6,633 probe sets -- 20 probe pairs each
- Harvard: HG_U95Av2 Chip
  - 12,453 probe sets – 16 probe pairs each
- Problems:
  - Different genes
  - Incomparable expression levels

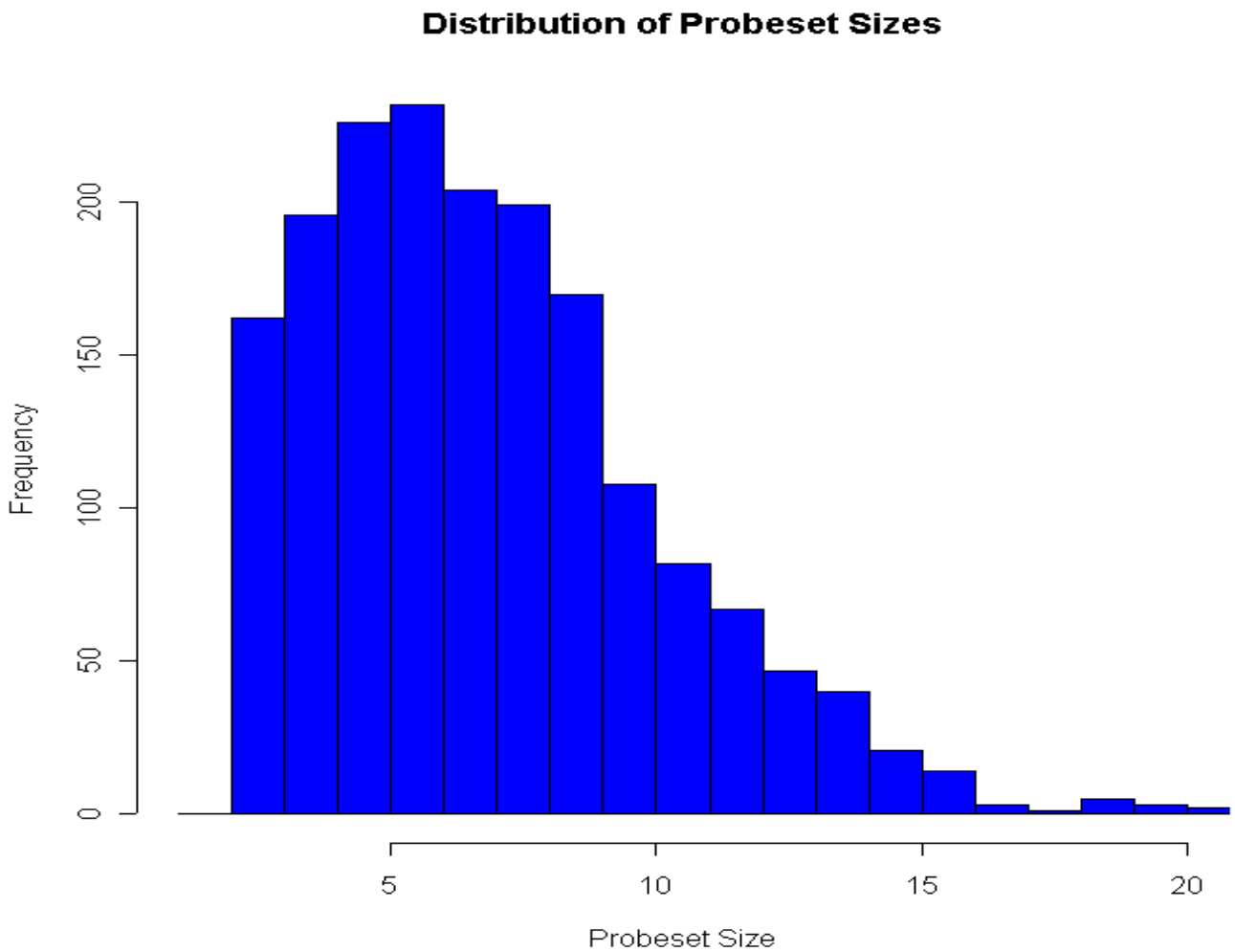
# Pooling Information across Chip Types

- Matching probes:




- Our approach:
  1. Use only matching probes
  2. Regroup into new probesets based on UNIGENE clusters (build 161) : “partial probesets”
- 4,101 probe sets with at least 3 probes

# Pooling Information across Chip Types

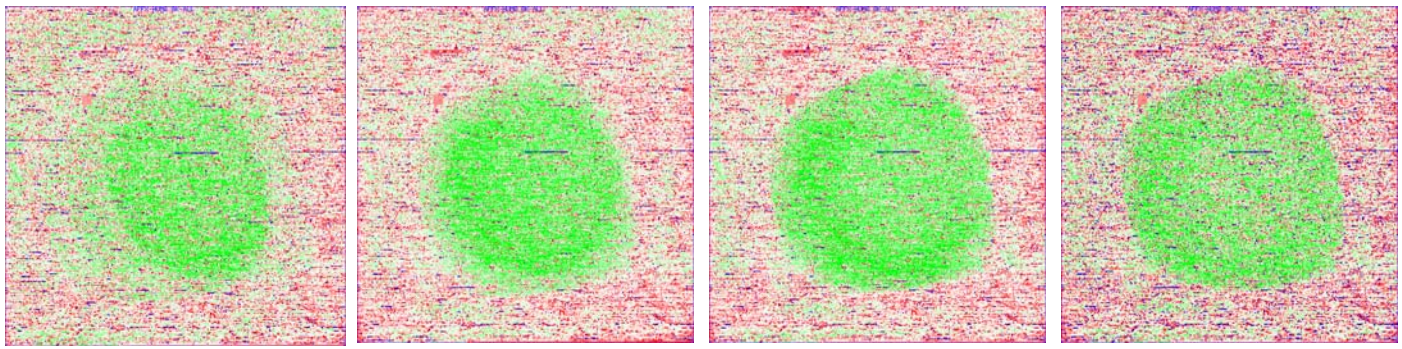


# Quantification of Expression Levels

- Gene expressions quantified by applying Li's PDNN model to our partial probesets
  - Uses probe sequence info to predict patterns of specific and nonspecific hybridization intensities
  - Allows borrowing of strength across probe sets
  - Model is not overparameterized – $O(N_{\text{probesets}})$
- Shown to outperform dChip and MAS5.0
- See Zhang, et al. (2003) *Nature Biotech* for further details on method and comparison

# Detecting Outliers

- Log-scale plots to detect outliers
  - Large spot detected on 4 Michigan chips

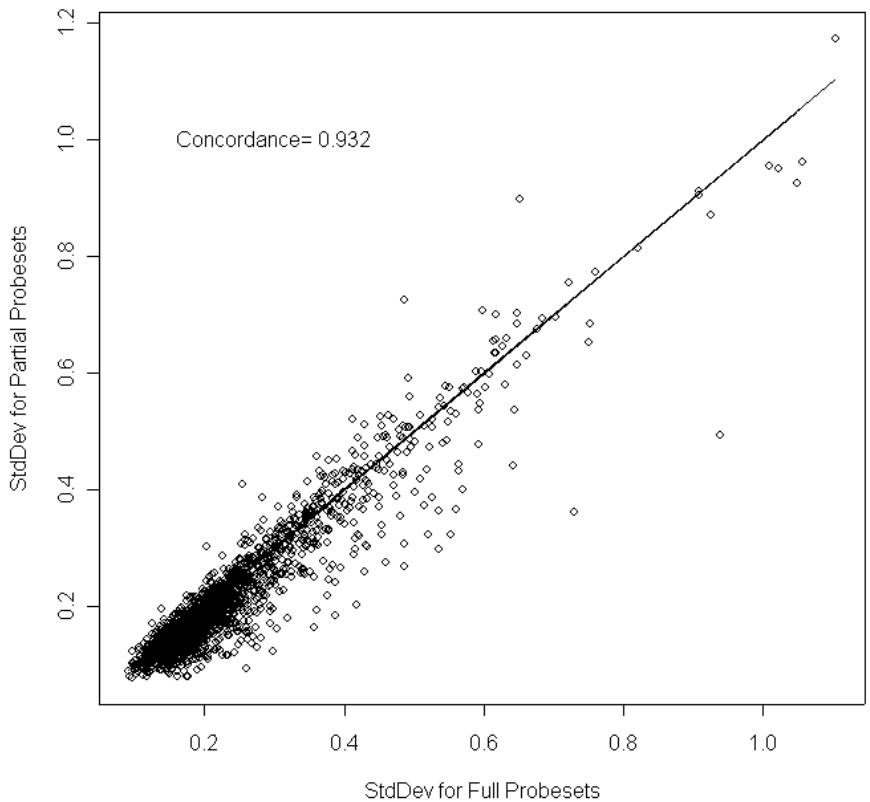


L54 L88 L89 L90

  - Other outliers: 6 from Michigan, 2 Harvard
- Other preprocessing (remove low expr./normalize)
- Matching clinical/microarray data for 200 patients (124 H, 76 M)

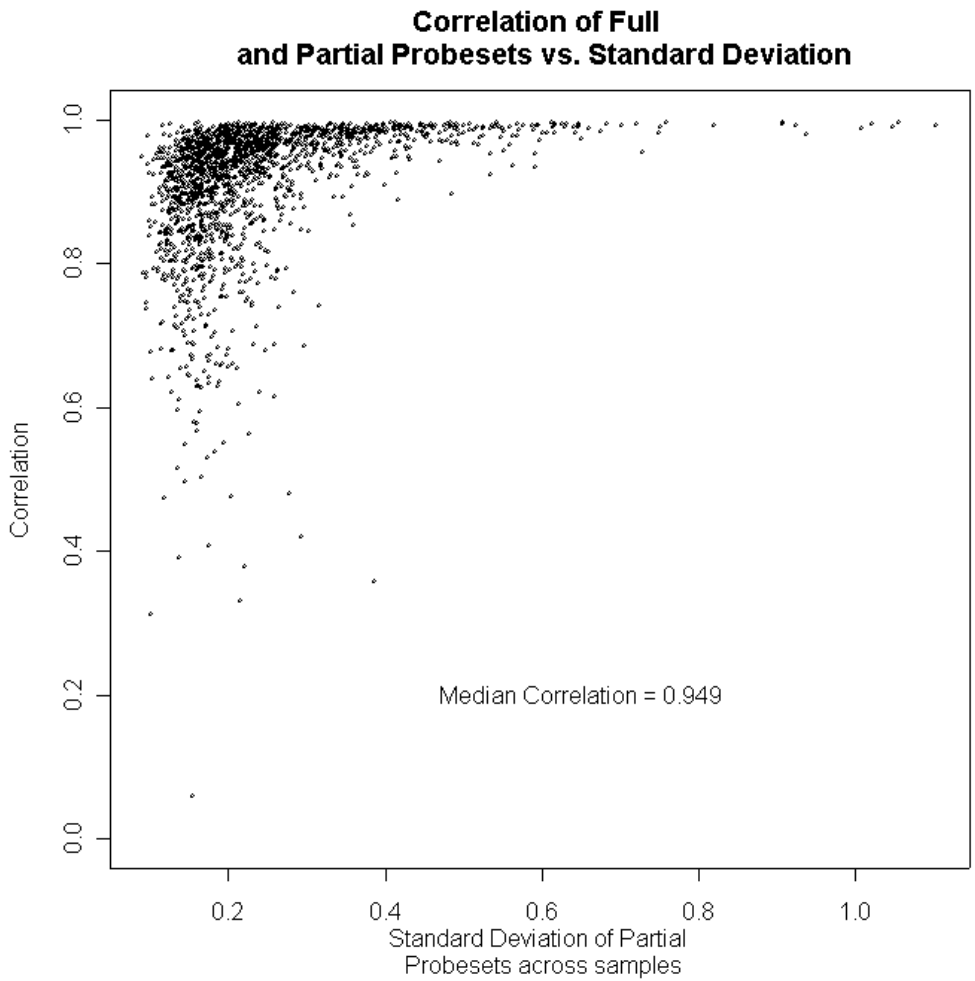
# Assessing “Partial Probeset” Method

- What do we lose by using partial probesets?



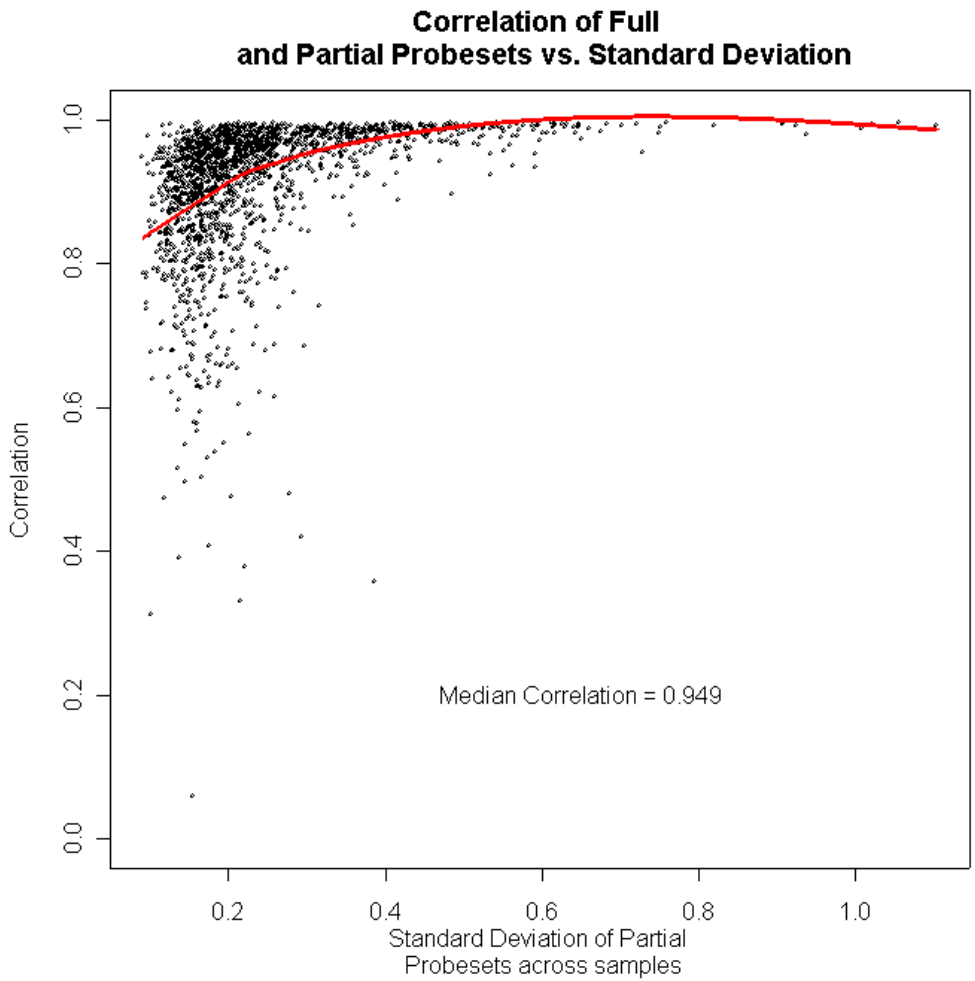

  - No evidence of precision loss
  - Standard deviations across samples similar when using full or partial probesets
  - Best probes carried forward?

# Assessing “Partial Probeset” Method



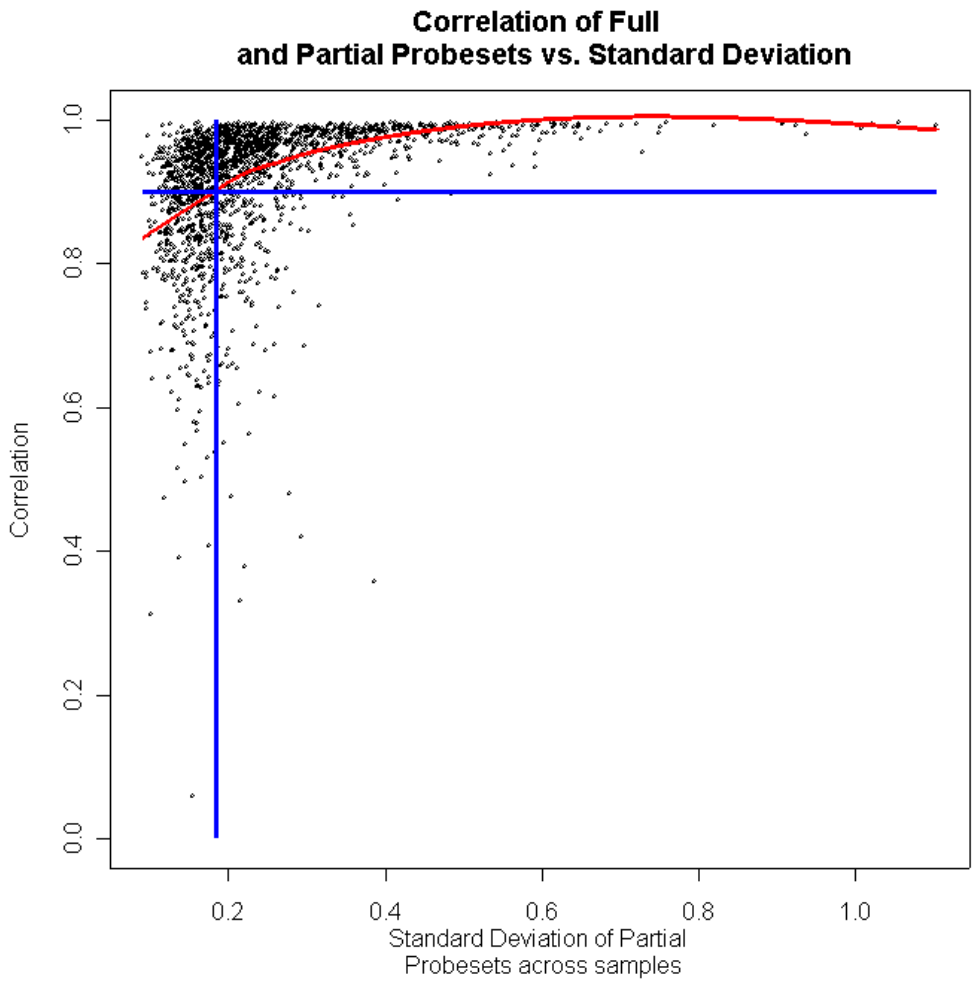

- Agreement in relative quantifications across samples

# Assessing “Partial Probeset” Method



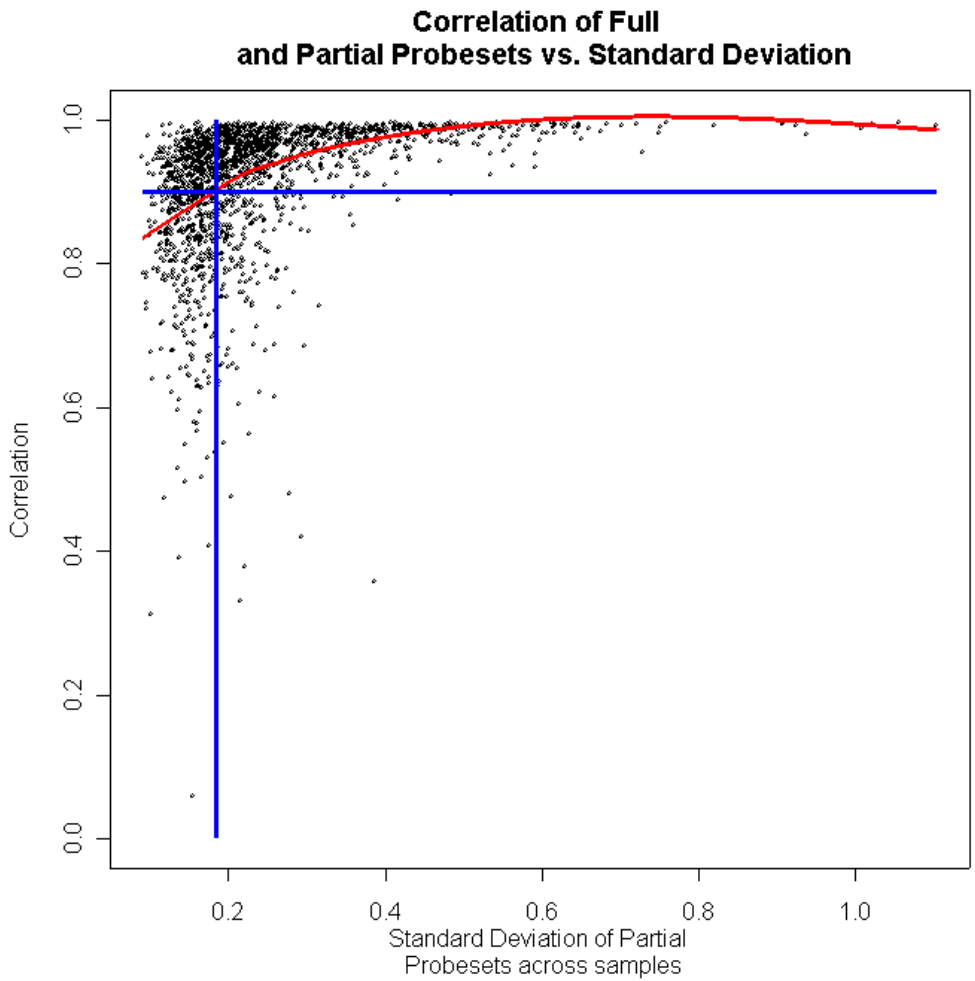

- Agreement in relative quantifications across samples
- Less variable genes worse

# Assessing “Partial Probeset” Method



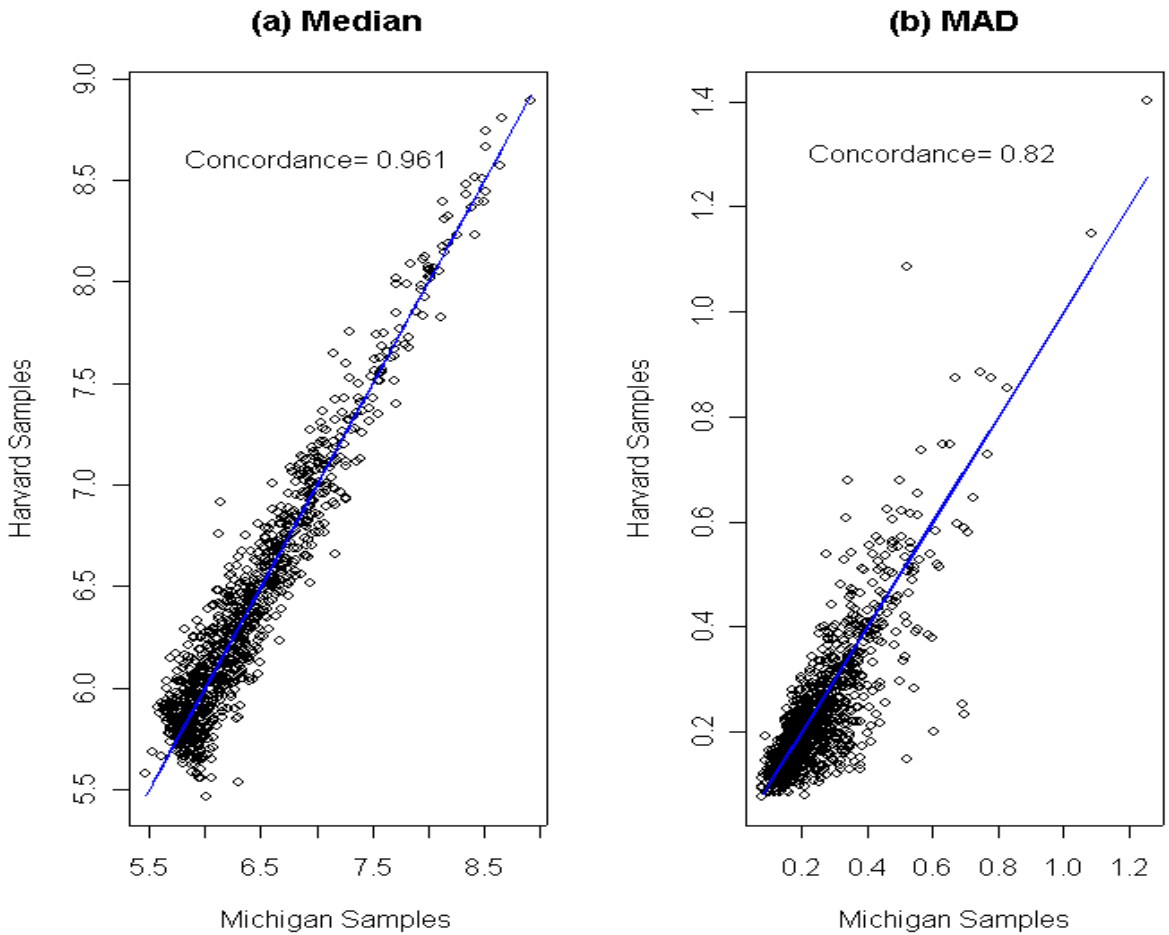

- Agreement in relative quantifications across samples
- Less variable genes worse
- Eliminate genes with *sd*<0.20 or *r*<0.90

# Assessing “Partial Probeset” Method




- Agreement in relative quantifications across samples
- Less variable genes worse
- Eliminate genes with *sd*<0.20 or *r*<0.90
- 1,036 genes

# Assessing “Partial Probeset” Method




- Median/MAD expression levels for genes similar across institution

# Identifying Prognostic Genes

- After preprocessing: 1036 genes, 200 samples
- Identify genes related to survival
  - **After** adjusting for known clinical predictors
- Fit series of multivariable Cox models:
  - Include study, age, stage, plus 1 gene
  - P-value for each gene using permutation test (Also done using LRT and bootstrap)
- Also identify genes differentially expressed by stage
  - Wilcoxon test for each gene

# Identifying Prognostic Genes

- After preprocessing:
  - 1036 genes, 200 samples
- Identify genes related to survival
  - **After** adjusting for known clinical predictors
  - Explain variability above and beyond c.p.
- Fit series of multivariable Cox models:
  - Include study, age, stage, plus 1 gene
  - P-value for each gene using permutation test

# Identifying Prognostic Genes: Cox Regression Modeling

- ***Hazard*** : $\lambda(t) \sim \text{Prob}(X < t + \Delta t \mid X > t)$
- Cox Model: $\lambda_i(t) = \lambda_0(t) \exp(X_i \beta)$
  - $X_i$ = Vector of covariates for subject $i$
  - $\beta$ = Vector of regression coefficients
- Key Assumption: Proportional Hazards
  - Hazard ratio between subjects with different covariates does not vary over time.
  - $\lambda_i(t)/\lambda_k(t) = \exp\{ (X_i - X_k) \beta \}$
  - Exp($\beta$) =Change in hazard per unit change in X

# Identifying Prognostic Genes: Cox Regression Modeling

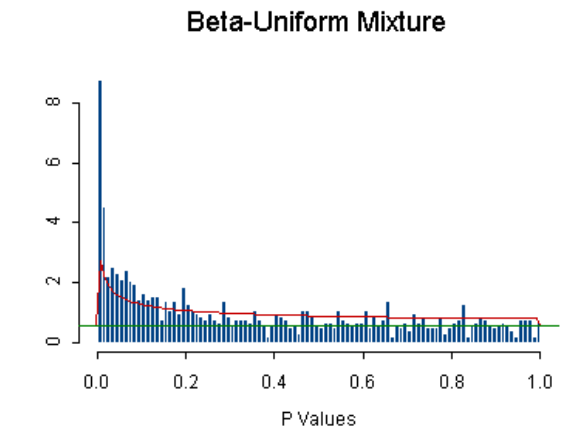
- Best Clinical Model:

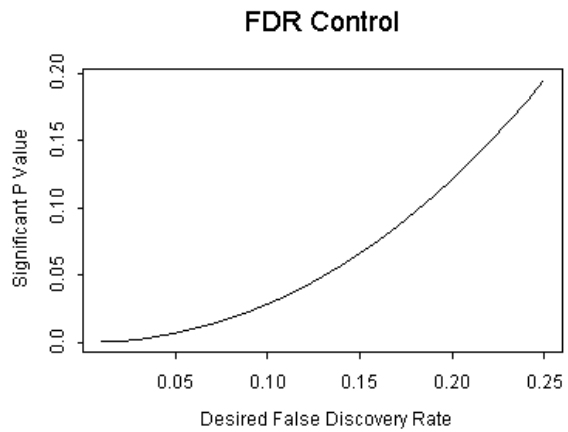
| Factor | $\beta$ | Exp($\beta$) | Z | p |
|---|---|---|---|---|
| **Study**<br>Michigan = 0<br>Harvard = 1 | 0.67 | 1.95 | 2.73 | 0.0062 |
| **Age** | 0.03 | 1.03 | 2.60 | 0.0094 |
| **Stage**<br>Early (1-2) = 0<br>Late (3-4) = 1 | 1.53 | 4.61 | 6.61 | <0.000000001 |

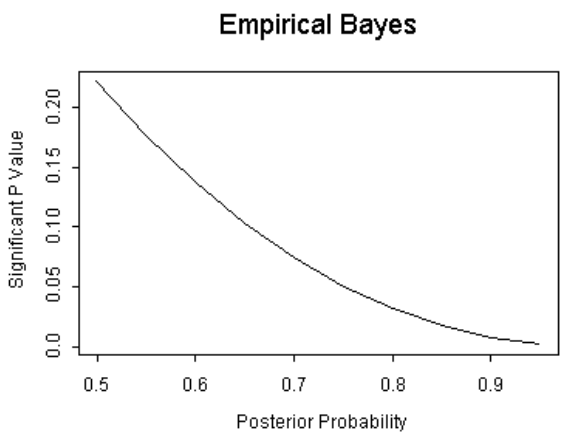
# Identifying Prognostic Genes: BUM Method

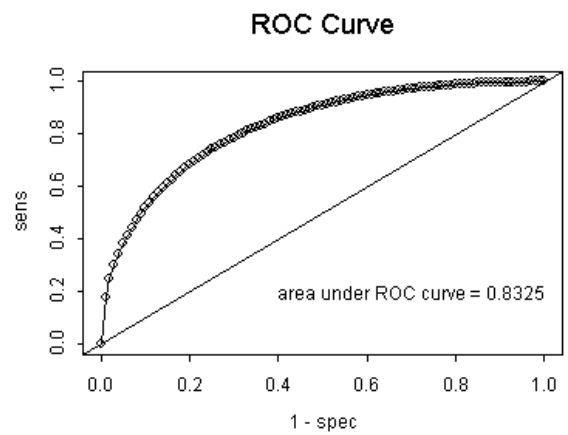
- No prognostic genes → pvals Uniform
- Prognostic genes → smaller pvals
- Fit Beta-Uniform mixture to histogram of p-values – "BUM" method
  - (Pounds and Morris, 2003 *Bioinformatics*)
- Method can be used to identify prognostic genes while controlling FDR

# Results: Stage-Related Genes



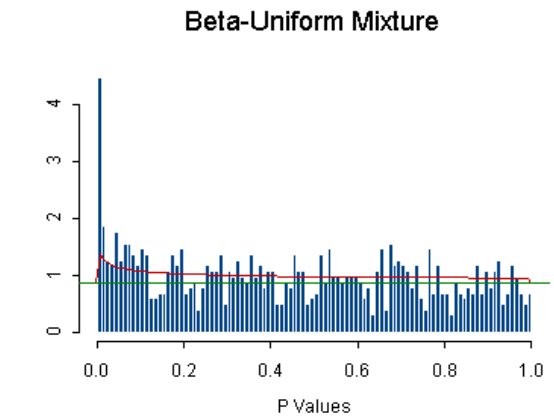



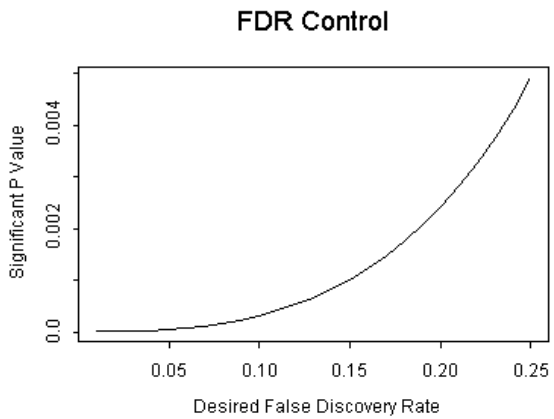



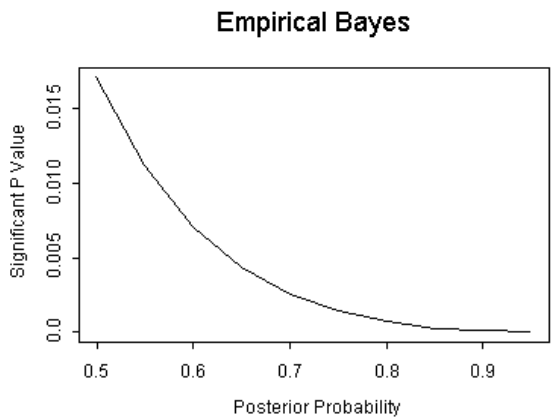



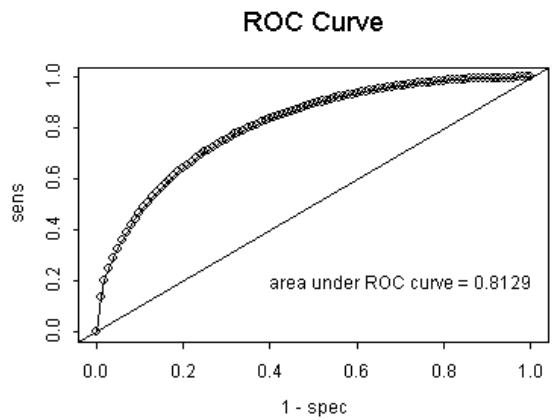

- Many genes linked with stage
- 71 genes flagged using FDR<0.05 (p<0.0064)

# Results: Prognostic Genes




- Evidence of prognostic genes
- 26 flagged using FDR<0.20
- Only 1 also flagged for stage
- 0 in top 100 genes in Michigan paper
- 1 cited in Harvard paper

# Selected Flagged Genes

| Rank | Gene | $\beta$ | p | $p_{Stage}$ | Function |
|---|---|---|---|---|---|
| 1 | FCGRT | -2.07 | <0.00001 | 0.154 | Induced by IF-$\gamma$ in treating SCLC |
| 2 | ENO2 | 1.46 | 0.00001 | 0.282 | Marker of NSCLC |
| 4 | RRM1 | 1.81 | 0.00002 | 0.321 | Linked to survival in NSCLC |
| 8 | CHKL | -1.43 | 0.00010 | 0.979 | Marker of NSCLC |
| 11 | CPE | 0.72 | 0.00031 | 0.088 | Marker of SCLC |
| 12 | ADRBK1 | -2.20 | 0.00044 | 0.484 | Co-expressed with Cox-2 in PUC |
| 16 | CLU | -0.52 | 0.00109 | 0.014 | Marker of SCLC |
| 20 | SEPW1 | -1.29 | 0.00145 | 0.028 | ↓ H202 cytotox. in NSCLC cell lines |
| 21 | FSCN1 | 0.66 | 0.00150 | 0.082 | Marker of invasiveness in Stage 1 NSCLC |

# Selected Flagged Genes

| Rank | Gene | $\beta$ | p | $p_{Stage}$ | Function |
|---|---|---|---|---|---|
| 1 | FCGRT | -2.07 | <0.00001 | 0.154 | Induced by IF-$\gamma$ in treating SCLC |
| 2 | ENO2 | 1.46 | 0.00001 | 0.282 | Marker of NSCLC |
| 4 | RRM1 | 1.81 | 0.00002 | 0.321 | Linked to survival in NSCLC |
| 8 | CHKL | -1.43 | 0.00010 | 0.979 | Marker of NSCLC |
| 11 | CPE | 0.72 | 0.00031 | 0.088 | Marker of SCLC |
| 12 | ADRBK1 | -2.20 | 0.00044 | 0.484 | Co-expressed with Cox-2 in PUC |
| 16 | CLU | -0.52 | 0.00109 | 0.014 | Marker of SCLC |
| 20 | SEPW1 | -1.29 | 0.00145 | 0.028 | ↓H202 cytotox. in NSCLC cell lines |
| 21 | FSCN1 | 0.66 | 0.00150 | 0.082 | Marker of invasiveness in Stage 1 NSCLC |

# Selected Flagged Genes

| Rank | Gene | $\beta$ | p | $p_{Stage}$ | Function |
|---|---|---|---|---|---|
| 1 | FCGRT | -2.07 | <0.00001 | 0.154 | Induced by IF-$\gamma$ in treating SCLC |
| 2 | ENO2 | 1.46 | 0.00001 | 0.282 | Marker of NSCLC |
| 4 | RRM1 | 1.81 | 0.00002 | 0.321 | Linked to survival in NSCLC |
| 8 | CHKL | -1.43 | 0.00010 | 0.979 | Marker of NSCLC |
| 11 | CPE | 0.72 | 0.00031 | 0.088 | Marker of SCLC |
| 12 | ADRBK1 | -2.20 | 0.00044 | 0.484 | Co-expressed with Cox-2 in PUC |
| 16 | CLU | -0.52 | 0.00109 | 0.014 | Marker of SCLC |
| 20 | SEPW1 | -1.29 | 0.00145 | 0.028 | ↓ H202 cytotox. in NSCLC cell lines |
| 21 | FSCN1 | 0.66 | 0.00150 | 0.082 | Marker of invasiveness in Stage 1 NSCLC |

# Selected Flagged Genes

| Rank | Gene | $\beta$ | p | $p_{Stage}$ | Function |
|---|---|---|---|---|---|
| 1 | FCGRT | -2.07 | <0.00001 | 0.154 | Induced by IF-$\gamma$ in treating SCLC |
| 2 | ENO2 | 1.46 | 0.00001 | 0.282 | Marker of NSCLC |
| 4 | RRM1 | 1.81 | 0.00002 | 0.321 | Linked to survival in NSCLC |
| 8 | CHKL | -1.43 | 0.00010 | 0.979 | Marker of NSCLC |
| 11 | CPE | 0.72 | 0.00031 | 0.088 | Marker of SCLC |
| 12 | ADRBK1 | -2.20 | 0.00044 | 0.484 | Co-expressed with Cox-2 in PUC |
| 16 | CLU | -0.52 | 0.00109 | 0.014 | Marker of SCLC |
| 20 | SEPW1 | -1.29 | 0.00145 | 0.028 | ↓ H202 cytotox. in NSCLC cell lines |
| 21 | FSCN1 | 0.66 | 0.00150 | 0.082 | Marker of invasiveness in Stage 1 NSCLC |

# Selected Flagged Genes

| Rank | Gene | $\beta$ | p | $p_{Stage}$ | Function |
|---|---|---|---|---|---|
| 1 | FCGRT | -2.07 | <0.00001 | 0.154 | Induced by IF-$\gamma$ in treating SCLC |
| 2 | ENO2 | 1.46 | 0.00001 | 0.282 | Marker of NSCLC |
| 4 | RRM1 | 1.81 | 0.00002 | 0.321 | Linked to survival in NSCLC |
| 8 | CHKL | -1.43 | 0.00010 | 0.979 | Marker of NSCLC |
| 11 | CPE | 0.72 | 0.00031 | 0.088 | Marker of SCLC |
| 12 | ADRBK1 | -2.20 | 0.00044 | 0.484 | Co-expressed with Cox-2 in PUC |
| 16 | CLU | -0.52 | 0.00109 | 0.014 | Marker of SCLC |
| 20 | SEPW1 | -1.29 | 0.00145 | 0.028 | ↓ H202 cytotox. in NSCLC cell lines |
| 21 | FSCN1 | 0.66 | 0.00150 | 0.082 | Marker of invasiveness in Stage 1 NSCLC |

# Selected Flagged Genes

| Rank | Gene | $\beta$ | p | $p_{Stage}$ | Function |
|---|---|---|---|---|---|
| 1 | FCGRT | -2.07 | <0.00001 | 0.154 | Induced by IF-$\gamma$ in treating SCLC |
| 2 | ENO2 | 1.46 | 0.00001 | 0.282 | Marker of NSCLC |
| 4 | RRM1 | 1.81 | 0.00002 | 0.321 | Linked to survival in NSCLC |
| 8 | CHKL | -1.43 | 0.00010 | 0.979 | Marker of NSCLC |
| 11 | CPE | 0.72 | 0.00031 | 0.088 | Marker of SCLC |
| 12 | ADRBK1 | -2.20 | 0.00044 | 0.484 | Co-expr with Cox-2 in lung adenocarc |
| 16 | CLU | -0.52 | 0.00109 | 0.014 | Marker of SCLC |
| 20 | SEPW1 | -1.29 | 0.00145 | 0.028 | ↓ H202 cytotox. in NSCLC cell lines |
| 21 | FSCN1 | 0.66 | 0.00150 | 0.082 | Marker of invasiveness in Stage 1 NSCLC |

# Selected Flagged Genes

| Rank | Gene | $\beta$ | p | $p_{Stage}$ | Function |
|---|---|---|---|---|---|
| 1 | FCGRT | -2.07 | <0.00001 | 0.154 | Induced by IF-$\gamma$ in treating SCLC |
| 2 | ENO2 | 1.46 | 0.00001 | 0.282 | Marker of NSCLC |
| 4 | RRM1 | 1.81 | 0.00002 | 0.321 | Linked to survival in NSCLC |
| 8 | CHKL | -1.43 | 0.00010 | 0.979 | Marker of NSCLC |
| 11 | CPE | 0.72 | 0.00031 | 0.088 | Marker of SCLC |
| 12 | ADRBK1 | -2.20 | 0.00044 | 0.484 | Co-expressed with Cox-2 in PUC |
| 16 | CLU | -0.52 | 0.00109 | 0.014 | Marker of SCLC |
| 20 | SEPW1 | -1.29 | 0.00145 | 0.028 | H202 cytotox. in NSCLC cell lines |
| 21 | FSCN1 | 0.66 | 0.00150 | 0.082 | Marker of invasiveness in Stage 1 NSCLC |

# Selected Flagged Genes

| Rank | Gene | $\beta$ | p | $p_{Stage}$ | Function |
|---|---|---|---|---|---|
| 3 | NFRKB | -2.81 | 0.00001 | 0.058 | Amplified in AML |
| 7 | ATIC | 1.81 | 0.00009 | 0.771 | Fusion partner of ALK which defines subtype of ALCL |
| 13 | BCL9 | -1.64 | 0.00069 | 0.057 | Over-expressed in ALL |
| 15 | TPS1 | -0.64 | 0.00107 | 0.882 | Associated with pulmonary inflammation |
| 25 | BTG2 | -0.75 | 0.00232 | 0.726 | Inhibits cell proliferation in primary mouse embryo fibroblasts lacking functional p53 |

# Selected Flagged Genes

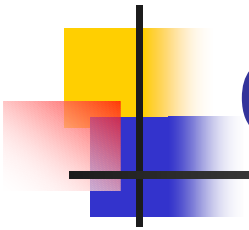
| Rank | Gene | $\beta$ | p | $p_{Stage}$ | Function |
|---|---|---|---|---|---|
| 3 | NFRKB | -2.81 | 0.00001 | 0.058 | Amplified in AML |
| 7 | ATIC | 1.81 | 0.00009 | 0.771 | Fusion partner of ALK which defines subtype of ALCL |
| 13 | BCL9 | -1.64 | 0.00069 | 0.057 | Over-expressed in ALL |
| 15 | TPS1 | -0.64 | 0.00107 | 0.882 | Associated with pulmonary inflammation |
| 25 | BTG2 | -0.75 | 0.00232 | 0.726 | Inhibits cell proliferation in primary mouse embryo fibroblasts lacking functional p53 |

# Selected Flagged Genes

| Rank | Gene | $\beta$ | p | $p_{Stage}$ | Function |
|---|---|---|---|---|---|
| 3 | NFRKB | -2.81 | 0.00001 | 0.058 | Amplified in AML |
| 7 | ATIC | 1.81 | 0.00009 | 0.771 | Fusion partner of ALK which defines subtype of ALCL |
| 13 | BCL9 | -1.64 | 0.00069 | 0.057 | Over-expressed in ALL |
| 15 | TPS1 | -0.64 | 0.00107 | 0.882 | Associated with pulmonary inflammation |
| 25 | BTG2 | -0.75 | 0.00232 | 0.726 | Inhibits cell proliferation in primary mouse embryo fibroblasts lacking functional p53 |

# Summary/Conclusions

- Pooling information across studies:
  - Fixed effect to model study heterogeneity
  - Method using matched probes to combine information across chip types
- Identified prognostic genes
  - Predictive above and beyond clinical predictors
  - Interesting biological results
- Was pooling worth it?
  - Yes, 17/26 genes would not have been identified by this analysis without pooling studies



# Collaborators/Acknowledgements

- Collaborators:
  - Li Zhang
  - Guosheng Yin
  - Keith Baggerly
  - Chunlei Wu
- Acknowledgements:
  - Kevin Coombes, David Stivers, Lianchun Xiao, and Sang-Joon Lee