Identification of Prognostic Genes, Combining Information Across Different Institutions and Oligonucleotide Arrays

Jeffrey S. Morris,

Guosheng Yin, Keith Baggerly, Chunlei Wu, and Li Zhang

UT MD Anderson Cancer Center

Department of Biostatistics

Introduction

- **CAMDA: "Critical Assessment of Microarray Data Analysis"**

- **CAMDA Challenge:** Pool information across studies to yield new biological insights.

- Our focus:
  1. **Adenocarcinoma** histology
  2. **Survival** outcome.
  3. **Michigan** (Beer, et al. 2002 *Nature Med*) & **Harvard** (Bhattacharjee, et al. 2001 *PNAS*)

Introduction

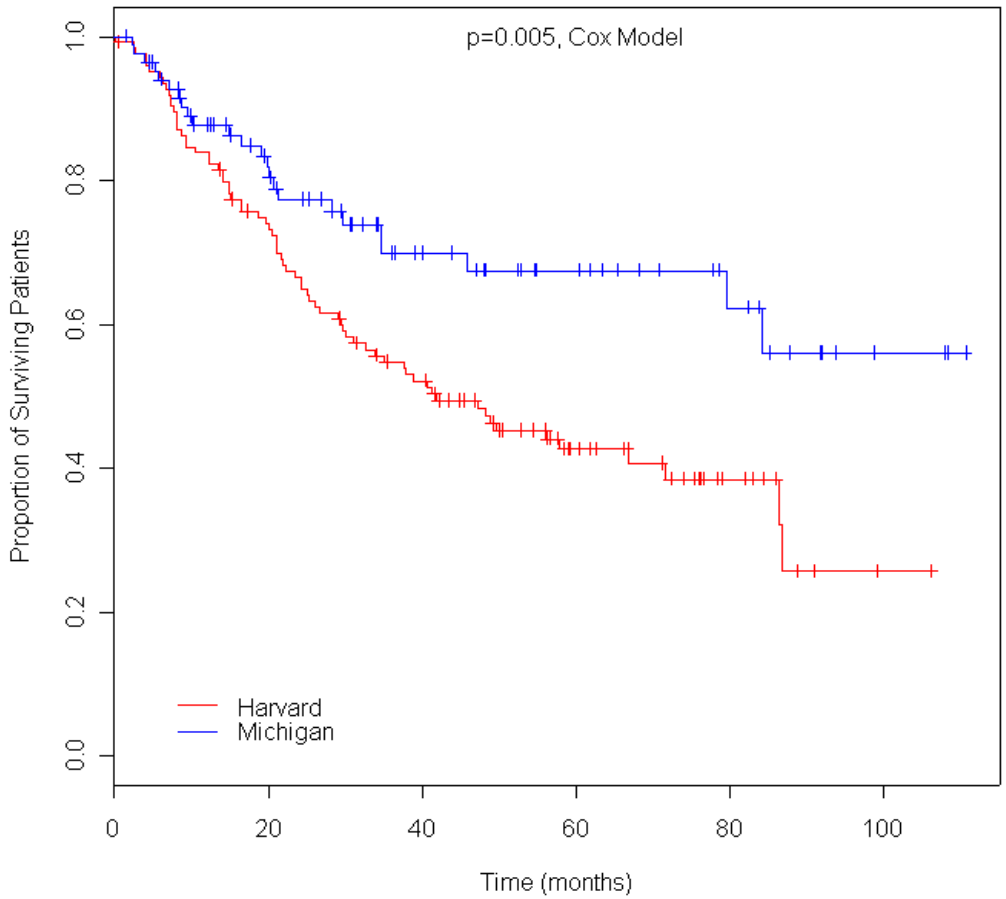
Our goals:

1. **Pool information** across different studies to **identify prognostic genes** for lung adenocarcinoma patients.

   - Offer information on patient survival ***over and above*** the information already provided by readily available clinical predictors.

2. Develop methodology to **pool information across different versions of Affymetrix chips** in such a way that we obtain comparable expression levels across the different chip types.

Pooling Information Across Studies



- Comparable distributions of age, gender, stage, smoking status, and follow-up time.
- Different survival distributions
- Fixed study effect included in our survival models to account for this heterogeneity

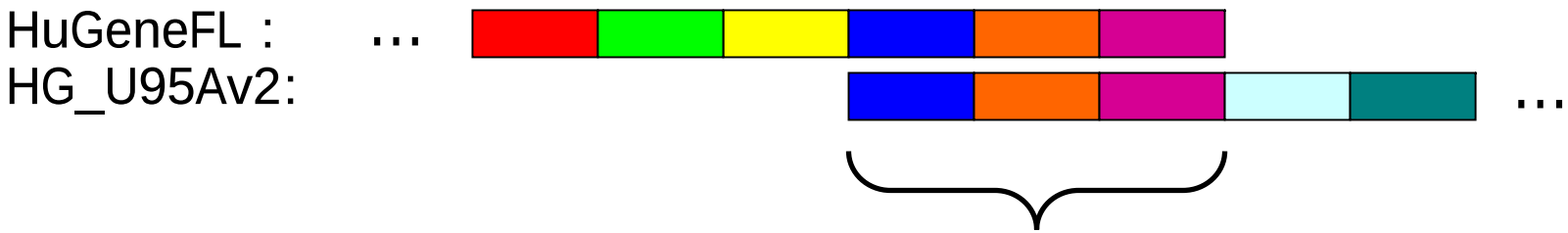
Pooling Information Across Chip Types

- Two studies used different chip types:
  - **Michigan**: **HuGeneFL**

    6,633 probesets/20 probe pairs each
  - **Harvard**: **U95Av2**

    12,453 probesets/16 probe pairs each
- Standard analyses on Affy-determined probesets not expected to yield comparable quantification

Pooling Information Across Chip Types



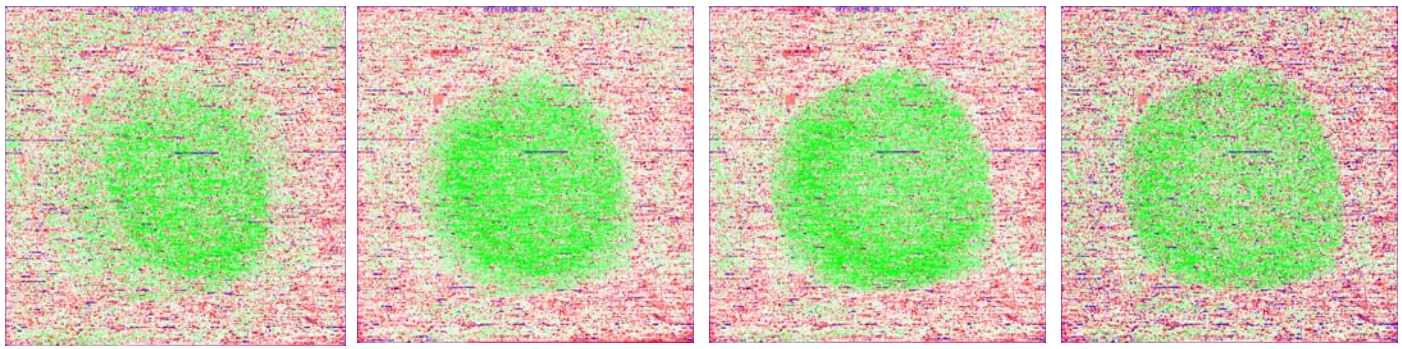
Our Solution

1. Identify "**matching probes**"
2. **Recombine** into new probesets based on UNIGENE clusters, which we refer to as "partial probesets"
3. **Eliminate** any probesets containing just one or two probes

- **Result: 4,101 partial probesets**.

Quality Control

- Several poor quality arrays removed
  - Large dead spot on center of 4 Michigan chips



L54 L88 L89 L90

  - 6 other Michigan chips/2 Harvard chips removed
- Matching clinical/microarray data for 200 patients (124 H, 76 M)

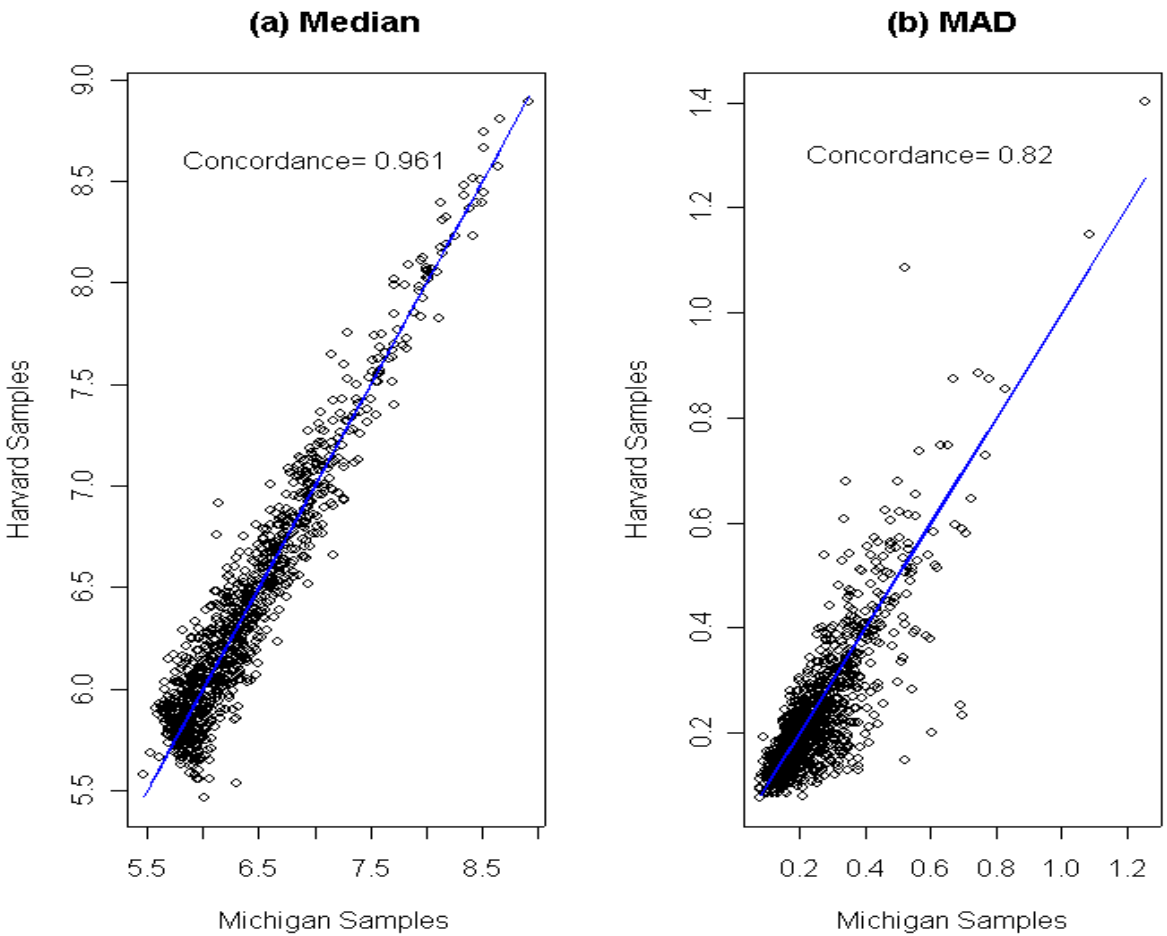
Quantification of Expression Levels

- Log-scale quantifications for each probeset obtained using PDNN model.
  - Uses Perfect Match (PM) probes only
  - Uses probe sequence info to predict patterns of specific and nonspecific hybridization intensities
  - Borrows strength across probe sets
- Shown to outperform dChip and MAS5.0
- See Zhang, et al. (2003) *Nature Biotech* for further details on method and comparison

Preprocessing

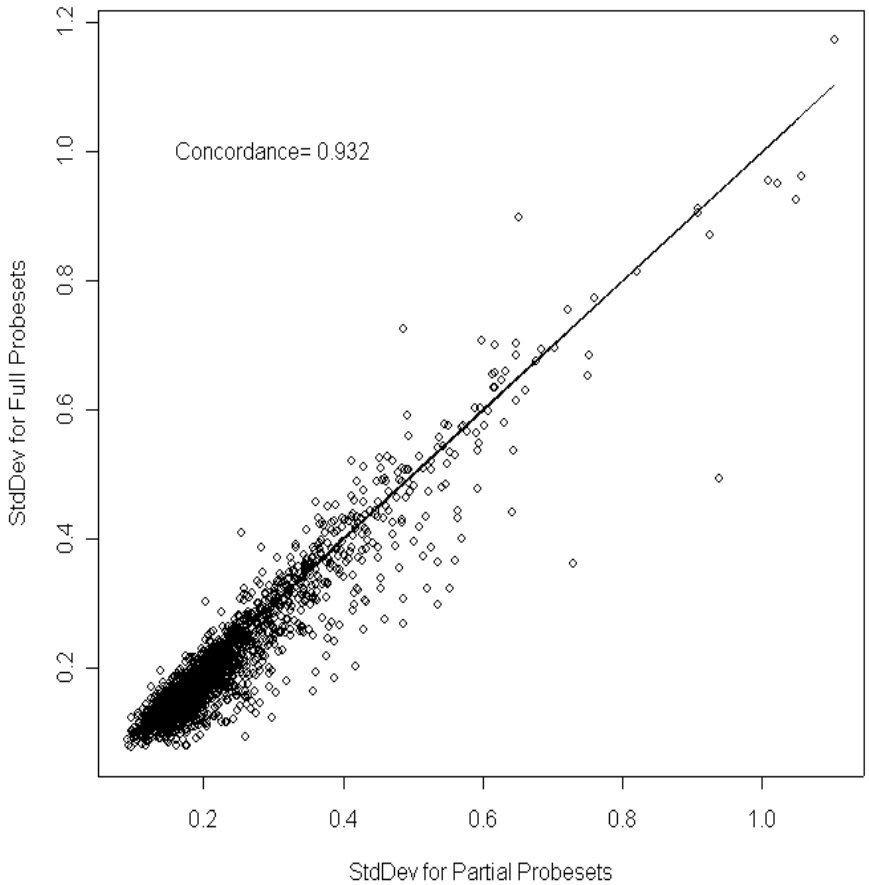
- Preprocessing steps:
  - Remove probesets with **smallest mean expression** levels across chips
  - **Normalize** log expression values within chips
  - Remove probesets with **smallest standard deviation** (<0.20) across chips
  - Remove probesets with **poor concordance** (<0.90) between partial and full probesets.
- **1036 probesets** remain after preprocessing

Assessing Our Method for Combining Information Across Chip Types



- "Partial Probeset" method appears to give **comparable expression levels** across chip types.

Assessing our Method for Combining Information across Chip Types



- Median "partial probeset" size is 7, vs. 16 or 20

  Loss of precision?
- No evidence of significant precision loss
- Also, relative ordering of samples well preserved (median r=0.95, using Spearman correlation)

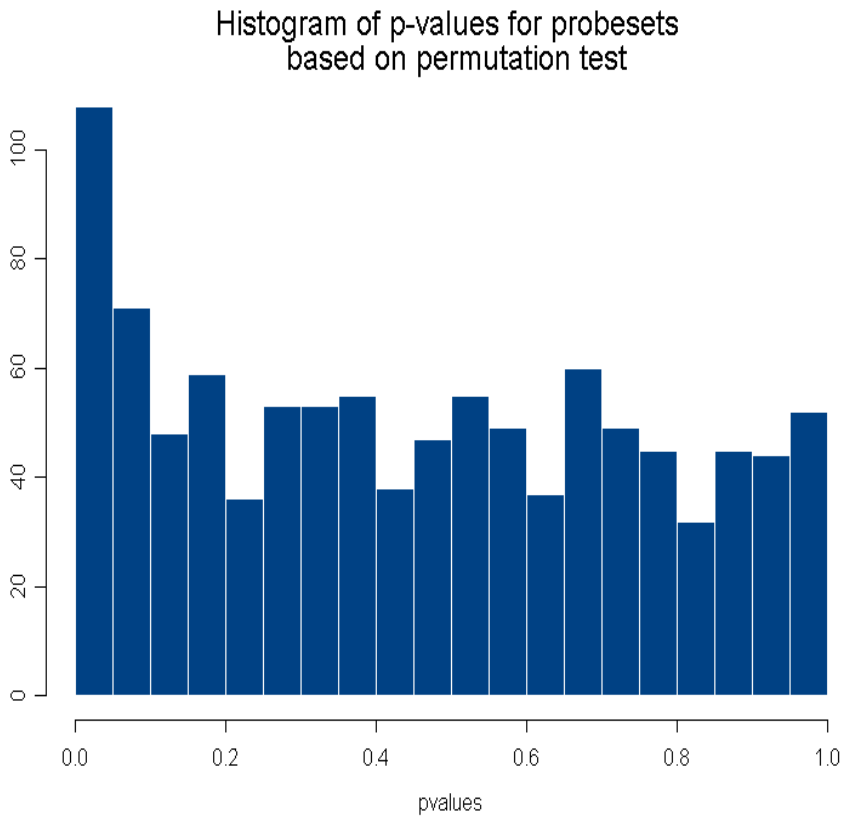
Identifying Prognostic Genes

- Series of 1036 **multivariable Cox models** fit to identify prognostic genes. Each model contained:
  - Study (Michigan=-1, Harvard=1).
  - Age (continuous factor).
  - Stage (early=0/late=1).
  - Probeset (log intensity value as continuous factor).
- Exact p-values for each probeset computed using **permutation approach**
- By using multivariate modeling, we search for genes offering prognostic information **beyond** clinical predictors

Identifying Prognostic Genes

- **BUM** method used to control **FDR<0.20**
  - Nonsignificant probesets → pvals Uniform
  - Significant probesets → more pvals near 0
  - Fit Beta-Uniform mixture to histogram of p-values
  - Model used to estimate FDR and get pval cutpoint
- Pounds and Morris, 2003 *Bioinformatics*

Results



- Histogram suggests there are some **significant probesets**
- FDR=0.20 corresponds pval cutoff of 0.0024
- **26** probesets flagged as significant

Selected Flagged Genes

Rank	Gene	β	p	Function
1	FCGRT	-2.07	<0.00001	Induced by IF-γ in treating SCLC
2	ENO2	1.46	0.00001	Marker of NSCLC
4	RRM1	1.81	0.00002	Linked to survival in NSCLC
8	CHKL	-1.43	0.00010	Marker of NSCLC
11	CPE	0.72	0.00031	Marker of SCLC
12	ADRBK1	-2.20	0.00044	Co-expressed with Cox-2 in lung ADC
16	CLU	-0.52	0.00109	Marker of SCLC
20	SEPW1	-1.29	0.00145	H202 cytotox. in NSCLC cell lines
21	FSCN1	0.66	0.00150	Marker of invasiveness in Stg 1 NSCLC
25	BTG2	-0.75	0.00232	Induced by p53 in SCLC cell lines

Selected Flagged Genes

Rank	Gene	β	p	Function
1	FCGRT	-2.07	<0.00001	Induced by IF-γ in treating SCLC
2	ENO2	1.46	0.00001	Marker of NSCLC
4	RRM1	1.81	0.00002	Linked to survival in NSCLC
8	CHKL	-1.43	0.00010	Marker of NSCLC
11	CPE	0.72	0.00031	Marker of SCLC
12	ADRBK1	-2.20	0.00044	Co-expressed with Cox-2 in lung ADC
16	CLU	-0.52	0.00109	Marker of SCLC
20	SEPW1	-1.29	0.00145	↓ H202 cytotox. in NSCLC cell lines
21	FSCN1	0.66	0.00150	Marker of invasiveness in Stg 1 NSCLC
25	BTG2	-0.75	0.00232	Induced by p53 in SCLC cell lines

Selected Flagged Genes

Rank	Gene	β	p	Function
1	FCGRT	-2.07	<0.00001	Induced by IF-γ in treating SCLC
2	ENO2	1.46	0.00001	Marker of NSCLC
4	RRM1	1.81	0.00002	Linked to survival in NSCLC
8	CHKL	-1.43	0.00010	Marker of NSCLC
11	CPE	0.72	0.00031	Marker of SCLC
12	ADRBK1	-2.20	0.00044	Co-expressed with Cox-2 in lung ADC
16	CLU	-0.52	0.00109	Marker of SCLC
20	SEPW1	-1.29	0.00145	↓ H202 cytotox. in NSCLC cell lines
21	FSCN1	0.66	0.00150	Marker of invasiveness in Stg 1 NSCLC
25	BTG2	-0.75	0.00232	Induced by p53 in SCLC cell lines

Selected Flagged Genes

Rank	Gene	β	p	Function
1	FCGRT	-2.07	<0.00001	Induced by IF-γ in treating SCLC
2	ENO2	1.46	0.00001	Marker of NSCLC
4	RRM1	1.81	0.00002	Linked to survival in NSCLC
8	CHKL	-1.43	0.00010	Marker of NSCLC
11	CPE	0.72	0.00031	Marker of SCLC
12	ADRBK1	-2.20	0.00044	Co-expressed with Cox-2 in lung ADC
16	CLU	-0.52	0.00109	Marker of SCLC
20	SEPW1	-1.29	0.00145	H202 cytotox. in NSCLC cell lines
21	FSCN1	0.66	0.00150	Marker of invasiveness in Stg 1 NSCLC
25	BTG2	-0.75	0.00232	Induced by p53 in SCLC cell lines

Selected Flagged Genes

Rank	Gene	β	p	Function
1	FCGRT	-2.07	<0.00001	Induced by IF-γ in treating SCLC
2	ENO2	1.46	0.00001	Marker of NSCLC
4	RRM1	1.81	0.00002	Linked to survival in NSCLC
8	CHKL	-1.43	0.00010	Marker of NSCLC
11	CPE	0.72	0.00031	Marker of SCLC
12	ADRBK1	-2.20	0.00044	Co-expressed with Cox-2 in lung ADC
16	CLU	-0.52	0.00109	Marker of SCLC
20	SEPW1	-1.29	0.00145	↓ H202 cytotox. in NSCLC cell lines
21	FSCN1	0.66	0.00150	Marker of invasiveness in Stg 1 NSCLC
25	BTG2	-0.75	0.00232	Induced by p53 in SCLC cell lines

Selected Flagged Genes

Rank	Gene	β	p	Function
1	FCGRT	-2.07	<0.00001	Induced by IF-γ in treating SCLC
2	ENO2	1.46	0.00001	Marker of NSCLC
4	RRM1	1.81	0.00002	Linked to survival in NSCLC
8	CHKL	-1.43	0.00010	Marker of NSCLC
11	CPE	0.72	0.00031	Marker of SCLC
12	ADRBK1	-2.20	0.00044	Co-expressed with Cox-2 in lung ADC
16	CLU	-0.52	0.00109	Marker of SCLC
20	SEPW1	-1.29	0.00145	↓ H202 cytotox. in NSCLC cell lines
21	FSCN1	0.66	0.00150	Marker of invasiveness in Stg 1 NSCLC
25	BTG2	-0.75	0.00232	Induced by p53 in SCLC cell lines

Selected Flagged Genes

Rank	Gene	β	p	Function
1	FCGRT	-2.07	<0.00001	Induced by IF-γ in treating SCLC
2	ENO2	1.46	0.00001	Marker of NSCLC
4	RRM1	1.81	0.00002	Linked to survival in NSCLC
8	CHKL	-1.43	0.00010	Marker of NSCLC
11	CPE	0.72	0.00031	Marker of SCLC
12	ADRBK1	-2.20	0.00044	Co-expressed with Cox-2 in lung ADC
16	CLU	-0.52	0.00109	Marker of SCLC
20	SEPW1	-1.29	0.00145	↓ H202 cytotox. in NSCLC cell lines
21	FSCN1	0.66	0.00150	Marker of invasiveness in Stg 1 NSCLC
25	BTG2	-0.75	0.00232	Induced by p53 in SCLC cell lines

Selected Flagged Genes

Rank	Gene	β	p	Function
1	FCGRT	-2.07	<0.00001	Induced by IF-γ in treating SCLC
2	ENO2	1.46	0.00001	Marker of NSCLC
4	RRM1	1.81	0.00002	Linked to survival in NSCLC
8	CHKL	-1.43	0.00010	Marker of NSCLC
11	CPE	0.72	0.00031	Marker of SCLC
12	ADRBK1	-2.20	0.00044	Co-expressed with Cox-2 in lung ADC
16	CLU	-0.52	0.00109	Marker of SCLC
20	SEPW1	-1.29	0.00145	H202 cytotox. in NSCLC cell lines
21	FSCN1	0.66	0.00150	Marker of invasiveness in Stg 1 NSCLC
25	BTG2	-0.75	0.00232	Induced by p53 in SCLC cell lines

Results

- Our gene list has almost no overlap with other publications of these data. Reasons:

1. We addressed a **different research question**
   - **Us**: ID Genes offering prognostic info beyond clinical
   - **Michigan**: Univariate Cox models fit; results used to construct dichotomous “risk index”
   - **Harvard**: Cluster analysis done; clusters linked to survival; found genes driving the clustering
2. Pooling across studies yielded **significant gains in statistical power**.
   - Most genes (17/26) in our study are not flagged if we analyze 2 data sets separately (i.e. no pooling)

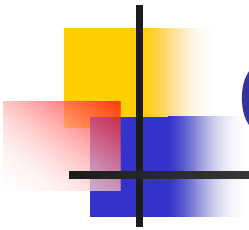
Conclusions

- New method for pooling info across studies using different versions of Affymetrix chips.
  - Recombine **matched probes** into **new probesets** using Unigene clusters.
  - Method appears to obtain **comparable** expression levels across chips without sacrificing much **precision** or significantly altering the **relative ordering** of the samples.

Conclusions

- Multivariate Cox models used to identify **new genes** offering prognostic information for lung adenocarcinoma patients.
  - Prognostic information **over and above** prognostic information provided by known clinical predictors.
  - Many of these genes seem **biologically interesting**.
  - It appears **increased statistical power** provided by the pooling helped in finding these new results.

- Pooling across studies:
  Great technical challenges, great gains to be realized
- CAMDA 2004 (http://www.camda.duke.edu/camda04)



Collaborators/Acknowledgements

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