Pooling Data Across Microarray Experiments Using Different Versions of Affymetrix Oligonucleotide Arrays

Jeffrey S. Morris,

Li Zhang, Chunlei Wu, Keith Baggerly and Kevin Coombes

UT MD Anderson Cancer Center

Houston, TX, USA

Combining Information across Microarray Studies

- Many publically available microarray data sets
- Can combine information across studies to:

1. Validate results from individual studies
   - Find intersection of differentially expressed genes
   - Build model using one study, validate using another
2. Discover new biological insights by analyses pooling data across studies.
   - Potential for increased statistical power
   - Important since many individual studies are underpowered.

Pooling Data across Studies

- **Challenge**: In general, microarray data from different studies not comparable
- Clinical differences
  - Different study populations
- Technical differences
  - Laboratory differences: sample collection and storage, microarray protocol
  - Different platforms: cDNA/oligo, different versions of same technology (e.g. Affy chips)

Pooling Data across Studies

- Approaches in Existing Literature:

1. Include study effects in model
   - Gene-specific study effects
   - SVD, Distance-weighted Discrimination

   Drawback: First-order corrections not enough
2. Model unitless summary measures
   - standardized log fold-change
   - t-statistics
   - probabilities of +/0/- expression

   Drawback: Implicit assumptions about comparability of clinical populations across studies

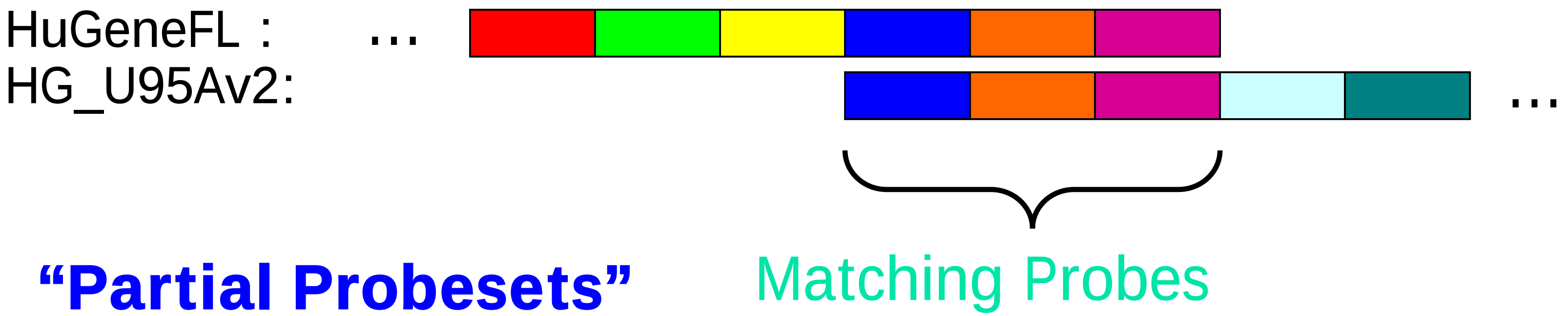
Pooling Data across Studies

- Sequence-related reasons for incomparability of raw expression levels across platforms:
  - Cross-hybridization
  - RNA degradation (near 5’ end)
  - Probe validity – map to RefSeq?
  - Alternative splicing
- It may be possible that, by taking these into account, we can obtain more comparable raw expression levels to use in pooled analyses
- **Our focus:** combining information across different versions of Affymetrix genechips

Overview of Affymetrix GeneChips

- **Probes:** 25-base sequences from gene of interest
- **Probesets**: set of probes corresponding to same gene.
  - Obtained from current sequence information in GenBank, Unigene, RefSeq
- Generations of human chips:
  - HuGeneFL: 5600 genes, 20 probes/gene
  - U95Av2: 10,000 genes, 16 probes/gene
  - U133A: 14,500 genes, 11 probes/gene

"Partial Probeset" Method



"Partial Probesets"

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

- Note: Any quantification method can subsequently be used (MAS, dChip, RMA, PDNN)

Example: Lung Cancer Data

- **Two studies** relating gene expression data to survival in lung cancer patients

1. **Harvard** (Bhattacharjee, et al. 2001)
   - 124 lung adenocarcinoma samples
2. **Michigan** (Beer, et al. 2002)
   - 86 lung adenocarcinoma samples

- **GOAL:** Pool data across studies to identify prognostic genes for lung cancer.

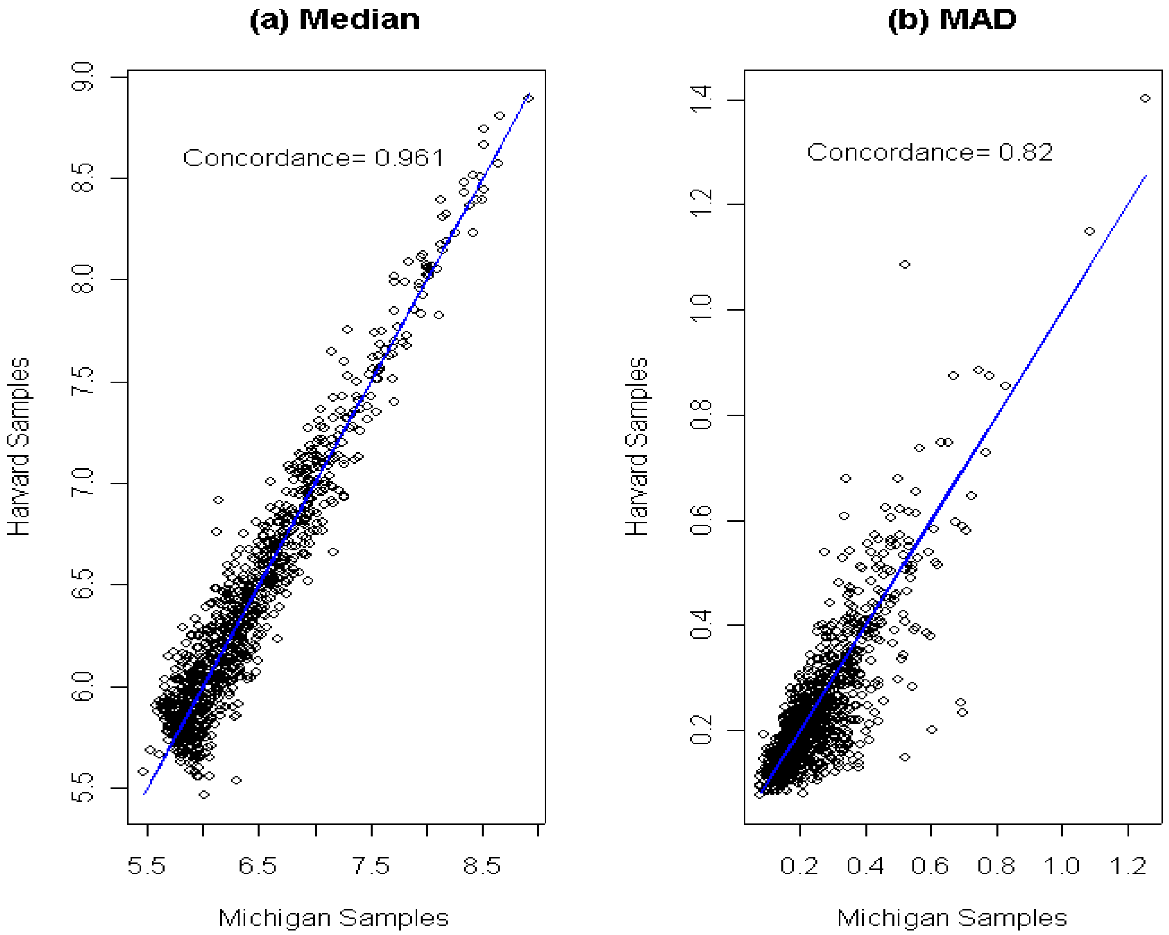
Example: CAMDA 2003 Data

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

    ~130,000 probes in 6,633 probesets
  - **Harvard**: **U95Av2**

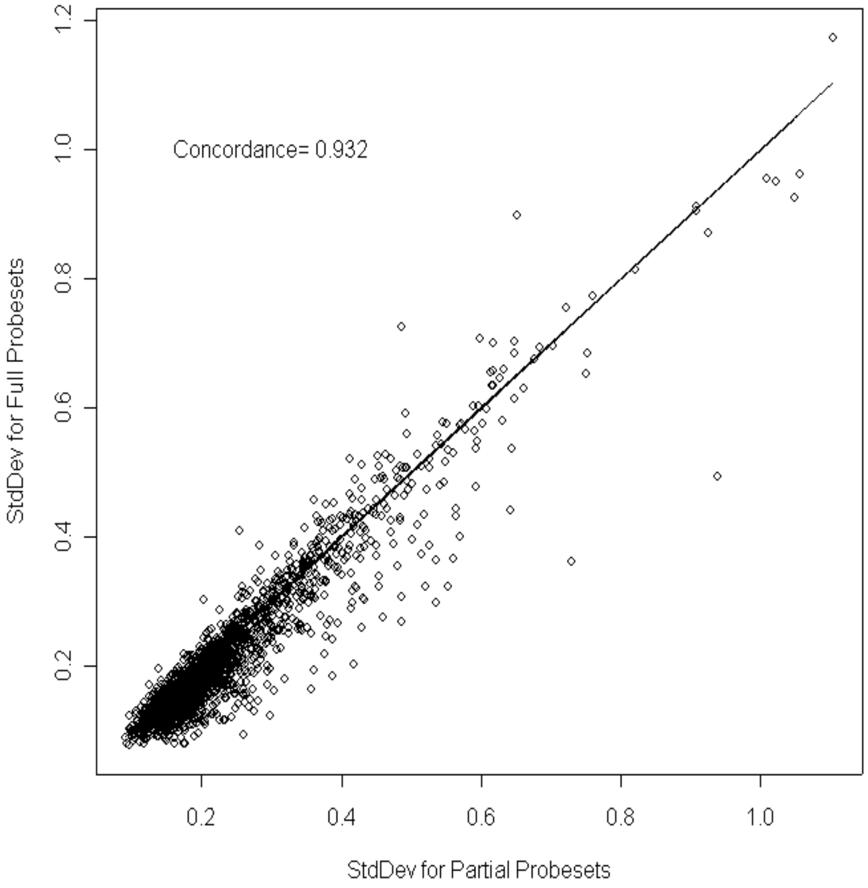
    ~200,000 probes in 12,625 probesets
- 34,428 “matching probes” combined into 4,101 partial probesets
- After preprocessing, 1036 probesets considered in subsequent analyses
- We used PDNN (Zhang et al. 2003 *Nature Biotech*) method for quantification

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

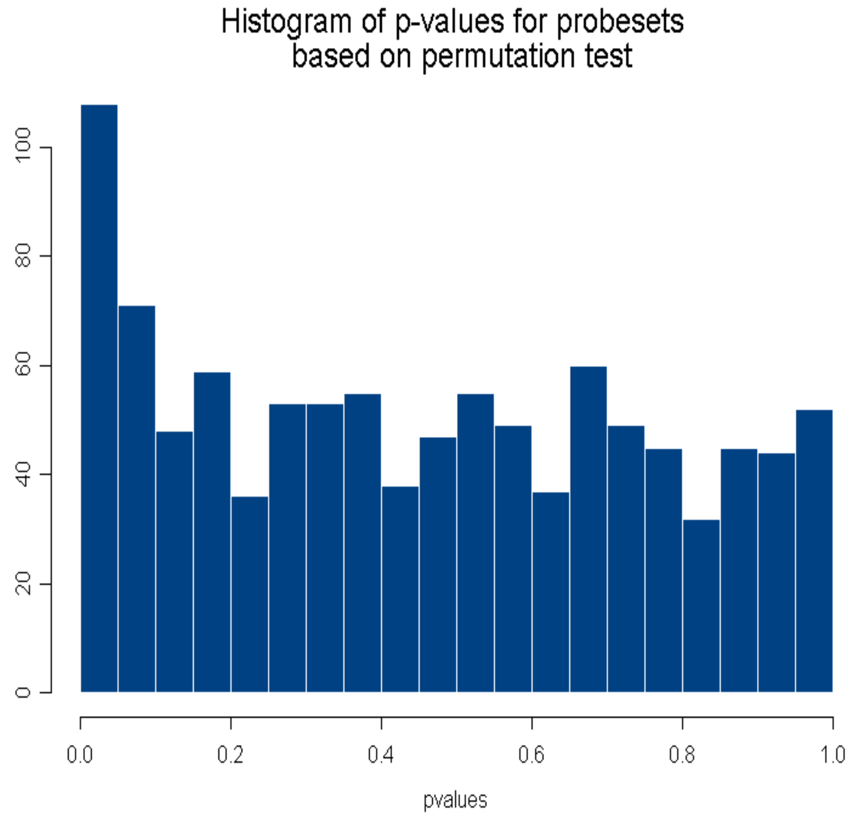
  Loss of precision?

- 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

Results



- Histogram suggests there are some **significant probesets**
- FDR=0.20 corresponds pval cutoff of 0.0024 (BUM, Pounds and Morris 2003)
- **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

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)

Limitations of Partial Probeset Method

- Worked well for combining across HuGeneFL/ U95Av2
  - ~25% probes from HuGeneFL on U95Av2, with 4,101 probesets
- Not enough matching probes for use with U95Av2/U133A
  - ~6% of probes from U95Av2 also on U133A, with only 628 probesets
- Requiring matching probes strong criterion, maybe weaker criterion would suffice?

Alternative Splicing

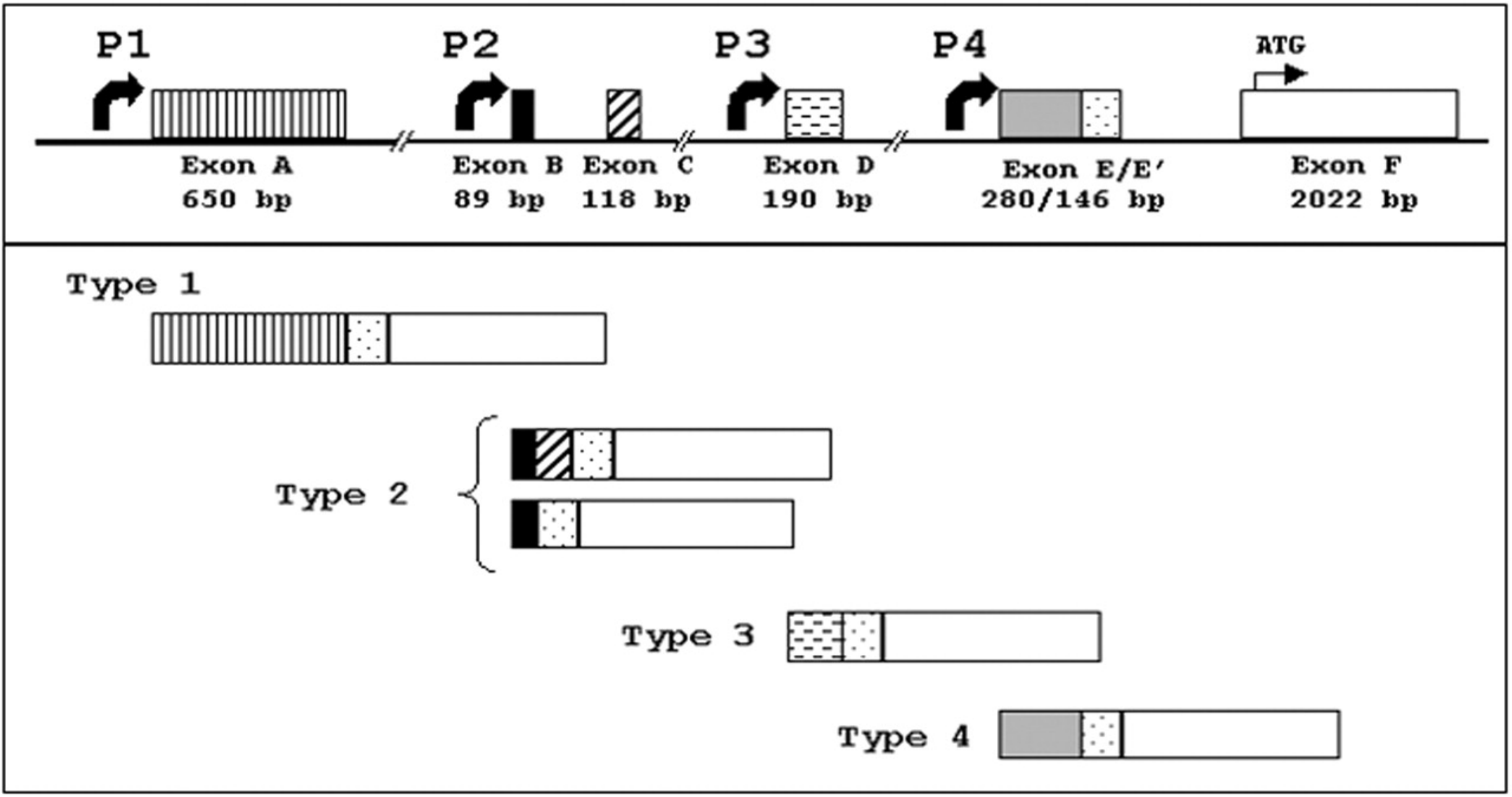


Diagram of C2GnT I gene organization and different mRNA variants of this gene that are **differentially expressed across tissue types**. From Falkenberg, et al. (2003) *Glycobiology* 13(6), 411-418.

Full-Length Transcript Based Probesets

- New probeset definition (FLTBP): probes match the same set of full-length mRNA sequences
- Procedure

1. Construct comprehensive library of full-length mRNA transcript sequences from RefSeq and HinvDB
2. For each probe, identify all matching full-length transcripts using Blast program

   U95Av2: 15% matched no sequence, 33% matched multiple seq.

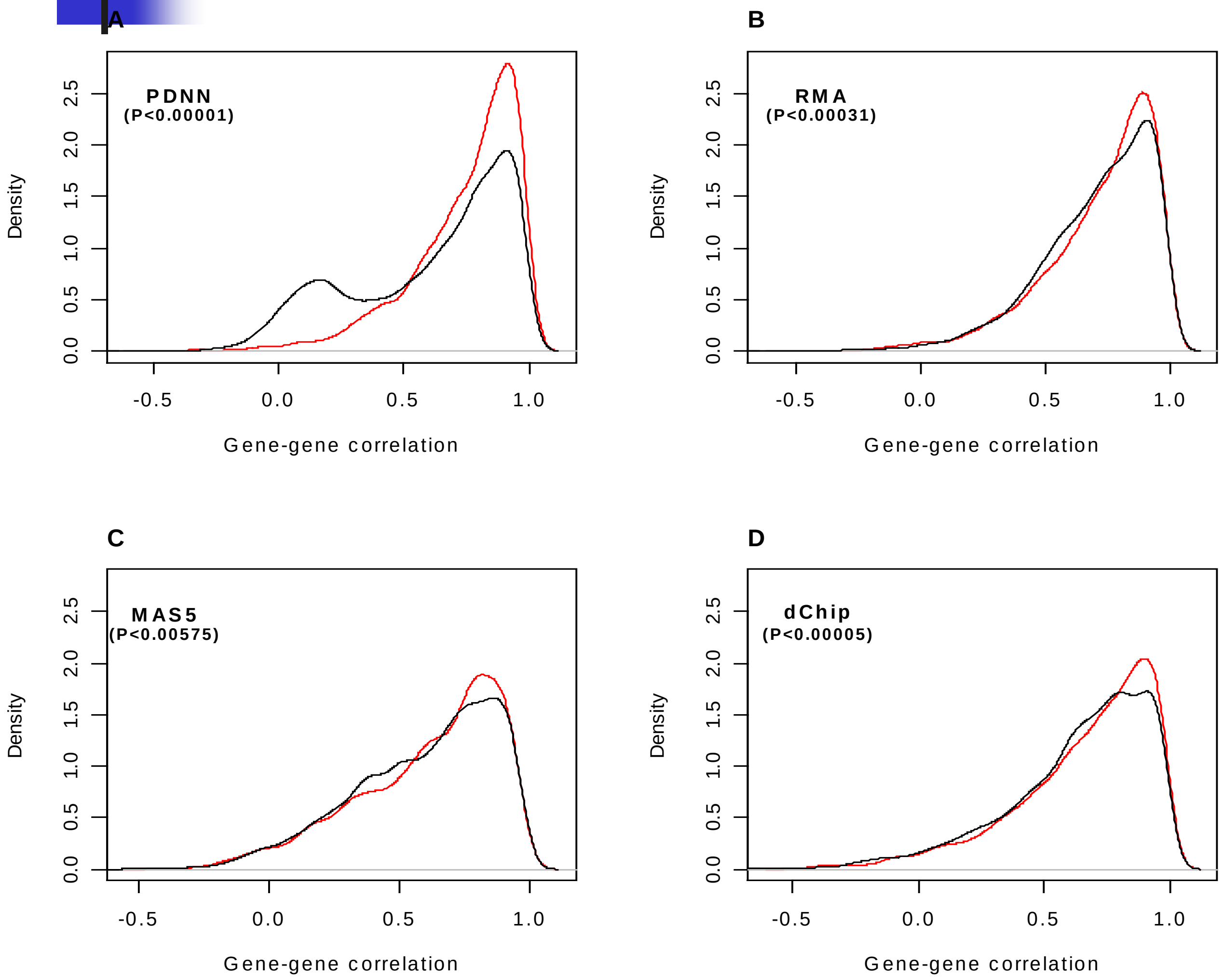
   U133A: 18% matched no sequence, 38% matched multiple seq.
3. Group probes with same matched target lists (FLTBPs)

   U95Av2: 23,972 probesets, U133A: 14,148 probesets

Full-Length Transcript Based Probesets

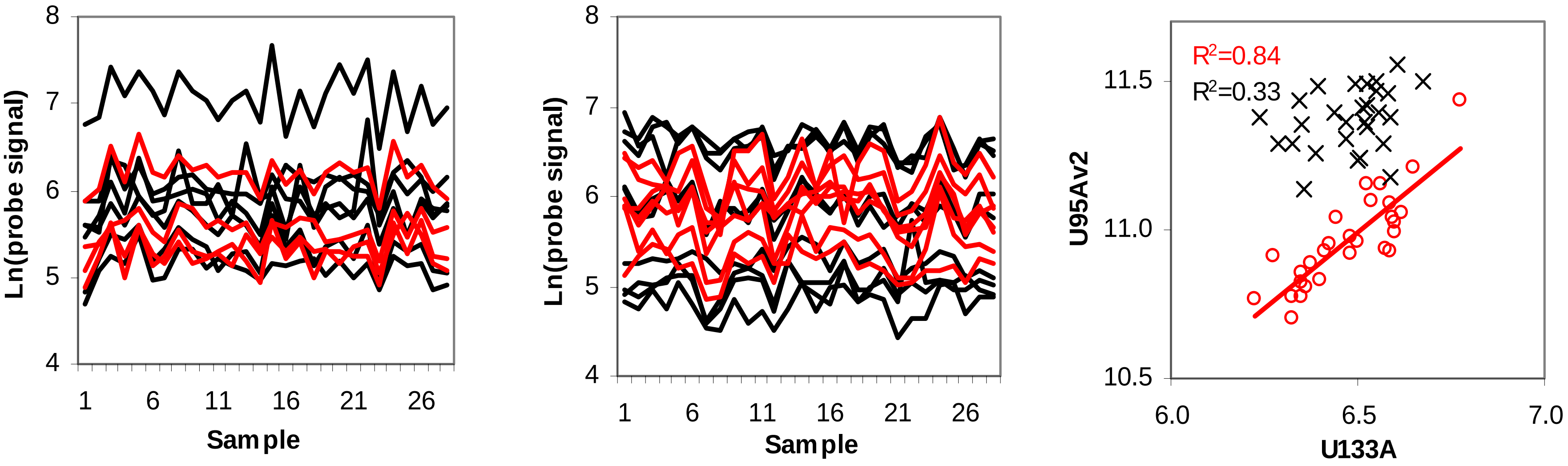
- Matching across chip types:
  - 9,642 FLTBPs match across U95Av2 and U133A
  - Affymetrix has their own method for mapping their probesets across arrays – 9,480 pairs of probesets (only about ½ map the same way as FLTBPs)
- Example: Lung cancer cell line data
  - 28 cell lines, each hybridized onto both U95Av2 and U133A arrays.
  - Paired design suggests any differences between paired measurements due to technical, not biological, sources.
  - Different quantification methods (PDNN, RMA, MAS, dChip)

Results



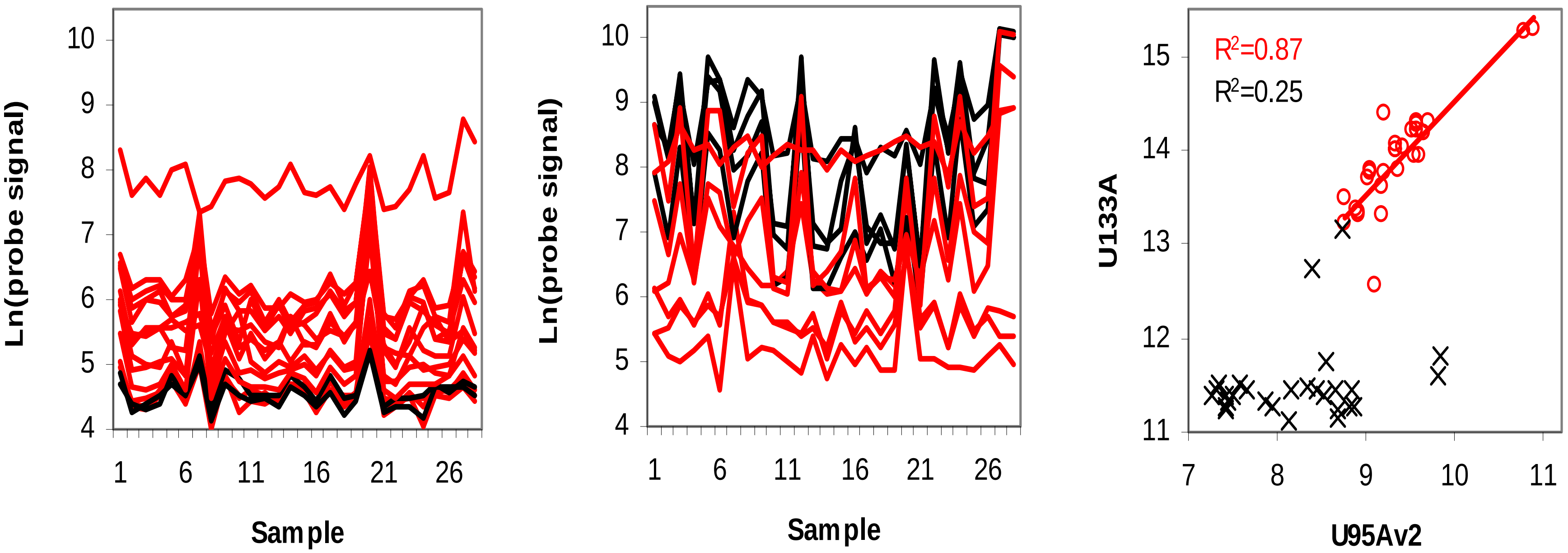
- Density estimate of chip-to-chip correlations for each gene
- Positive shift for FLTBP suggests better correlations
- Improvement greatest for PDNN
- Correlation still not perfect

Example: Sample Gene 1



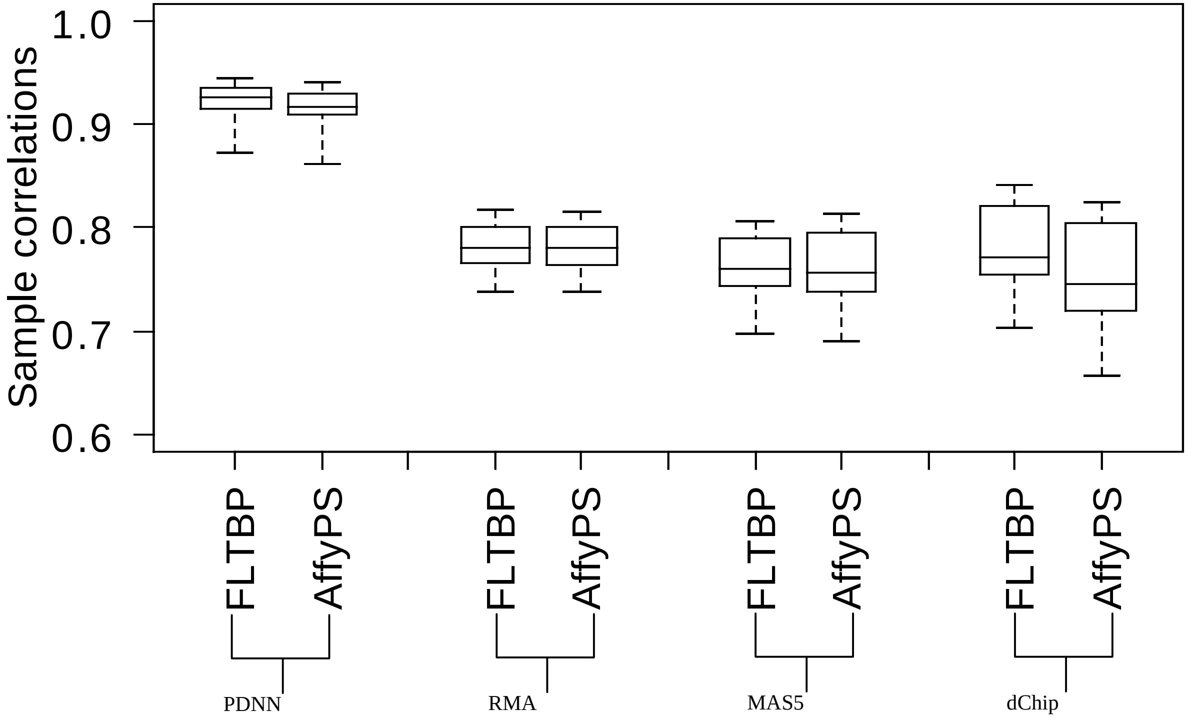
- Plot of probe signals for two chip types (Red=FLTBP)
- Scatterplot of log-expression values for each sample across the two chip types (Black=all probes, Red=FLTBP)
- Correlation across chips significantly improved with FLTBP

Example: Sample Gene 2



- Again, significantly higher correlation using FLTBP than using Affymetrix’ definition

Results



- Boxplot of chip-to-chip correlations (over genes) for each sample
- PDNN resulted in higher correlations

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.
  - Worked well combining information across HuGeneFL/U95Av2, but not U95Av2/U133A

Conclusions

- Discussed new probeset definition based on full-length transcript sequences.
  - Removes effect of known alternative splicing
  - Yields stronger between-chip correlations than Affymetrix standard definitions
- Pooling information across studies is difficult – there is still more work to be done – but worth the effort.

References

- Morris JS, Yin G, Baggerly KA, Wu C, and Zhang L (2005). Pooling Information Across Different Studies and Oligonucleotide Microarray Chip Types to Identify Prognostic Genes for Lung Cancer. *Methods of Microarray Data Analysis IV,* eds. JS Shoemaker and SM Lin, pp. 51-66, New York: Springer-Verlag.

- Wu C, Morris JS, Baggerly KA, Coombes KR, Minna JD, and Zhang L (2005). A probe-to-transcripts mapping method for cross-platform comparisons of microarray data taking into account the effects of alternative splicing. Under review.

- Morris JS, Wu C, Coombes KR, Baggerly KA, Wang J, and Zhang L (2005). Alternative Probeset Definitions for Combining Microarray Data Across Studies Using Different Versions of Affymetrix Oligonucleotide Arrays. To appear in *Meta-Analysis in Genetics*, edited by Rudy Guerra and David Allison, Chapman-Hall