GS01 0163
Analysis of Microarray Data

Keith Baggerly and Kevin Coombes
Section of Bioinformatics
Department of Biostatistics and Applied Mathematics
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
kcoombes@mdanderson.org

23 November 2004

Lecture 23: Meta-analysis

- CAMDA and Lung Cancer
- Combining Data Across Studies
- Combining Data Across Chip Types
- Model Checking
- Incorporating Clinical Information
- Testing Significance

The Longer Title...

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

in *Methods of Microarray Data Analysis IV*, Jennifer Shoemaker and Simon Lin (eds).

aka, the story of our entry in CAMDA 2003.

Why Meta-analysis?

The broad problem is simply that people have been doing microarray experiments for a while now, and in many cases the raw data is there to be had.

The hope is that this information can be leveraged by combining the information from multiple studies in such a way that we can (a) double-check the robustness of the initial results or (b) say something qualitatively new.

This latter category can be subdivided: results that could have been found initially, and results that stand out only with the combination.

CAMDA 2003

Meta-analysis was the theme of the 2003 CAMDA competition. Before this competition, the entrants had been supplied with two datasets and required to analyze one. In this case, the entrants were supplied with four datasets, and the requirement was to combine results from at least two.

The four datasets all involved microarray profiling of lung cancers (we're a bit uneasy about contrasting across organ types for now).

Goals in the Original Papers

In all four cases, the goal had been to indentify genes whose expression levels were correlated with clinical outcome, defined here as longer survival. Thus, all four datasets also have associated clinical information about survival and various patient characteristics.

Our motivation for looking at this was straightforward: we're at a cancer center, we want to improve our knowledge about how the cancer works.

We decided to stick with the idea of trying to associate gene expresssion with clinical outcome.

The Papers Themselves

Harvard: (Battacharjee, et al. PNAS 2001) 186 patients (139 ADC), Affymetrix U95Av2 arrays

The Papers Themselves

Harvard: (Battacharjee, et al. PNAS 2001) 186 patients (139 ADC), Affymetrix U95Av2 arrays

Michigan: (Beer, et al. Nature Med. 2002) 86 patients (all ADC), Affymetrix arrays HuGeneFL

The Papers Themselves

Harvard: (Battacharjee, et al. PNAS 2001) 186 patients (139 ADC), Affymetrix U95Av2 arrays

Michigan: (Beer, et al. Nature Med. 2002) 86 patients (all ADC), Affymetrix arrays HuGeneFL

Stanford: (Garber, et al. PNAS 2001) 62 patients (35 ADC), Glass arrays

The Papers Themselves

Harvard: (Battacharjee, et al. PNAS 2001) 186 patients (139 ADC), Affymetrix U95Av2 arrays

Michigan: (Beer, et al. Nature Med. 2002) 86 patients (all ADC), Affymetrix arrays HuGeneFL

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

How Do We Choose?

Which datasets should we combine?

Break this down into related questions:

1. What do we want to do with the data?
2. Does it make sense to try to combine the data?
3. How is this combination to be achieved?

What do we want to do?

We want to use expression levels to predict outcome.

What do we want to do?

We want to use expression levels to predict outcome.

But

What do we want to do?

We want to use expression levels to predict outcome.

But

a. as stated, this doesn't fully exploit the other clinical information available.

b. this is what the individual groups did, so the main gain to be had would be from increasing the sample size.

We'd like to expand the problem, and in so doing make some broader gains available.

The new objective

Find genes whose expression levels supply information about survival *above and beyond* that which can be derived largely from the clinical covariates.

This explicitly incorporates the clinical covariate information into the context of the problem. By doing something new, we get some additional information of a type that the individual studies did not work with.

What datasets make sense?

Given the question, what data can we work with?

In terms of predicting survival, one factor that we know will be present is the difference between institutions.

If the patient populations are qualitatively different in other ways as well it will be harder to make a valid comparison.

Check Rough Clinical Equivalence

Some aspects in which we seek similarity:

type of tumor: adenocarcinoma, SCLC, etc

stage of tumor: high grade/low grade?

length of followup

e.g. – The majority of the tumors in all cases are adenocarcinomas, and the other subtypes are very unevenly distributed.

Check Rough Clinical Equivalence

Some aspects in which we seek similarity:

type of tumor: adenocarcinoma, SCLC, etc

stage of tumor: high grade/low grade?

length of followup

e.g. – The majority of the tumors in all cases are adenocarcinomas, and the other subtypes are very unevenly distributed.

Focus on adenocarcinomas.

What else do we see in the datasets?

The Toronto dataset had comparatively short followup in a smaller number of cases – 3 events in 18 patients – so we we'rent sure it could be reliably compared. There were also difficulties in finding a consistent mapping of the gene identifiers (which turned out to be IMAGE clone ids). We decided not to use this one.

The Stanford dataset, by contrast, had a large number of patients with metastases (stage IV tumors) – 12 cases in 30 patients. Unsurprisingly, the survival profiles looked worse. We decided not to use this one.

The others were just right...

Similar gender/age/stage/smoking status

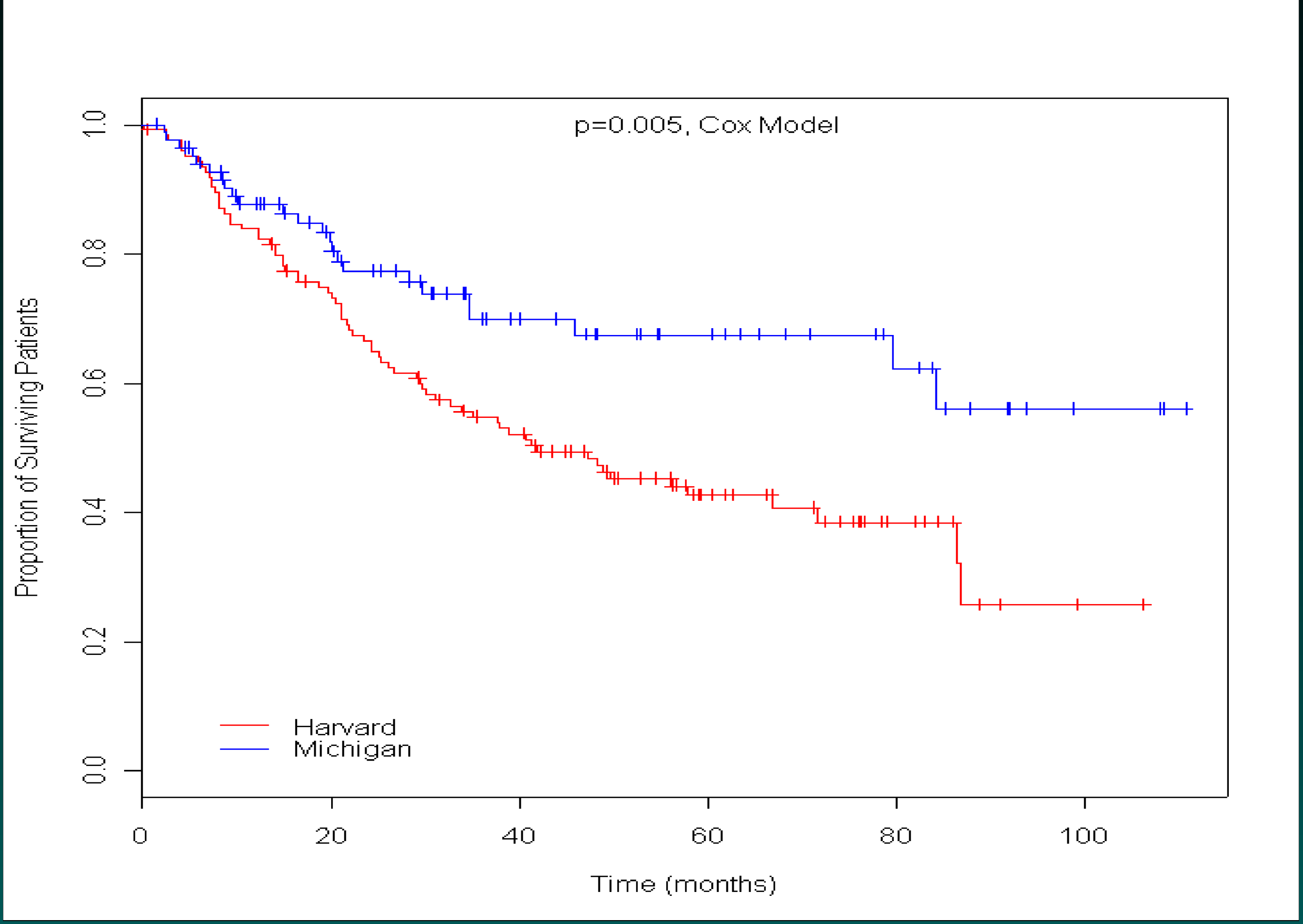
Similar follow-up time distributions

Different survival distributions

These are significantly different using standard tests, such as log rank tests or proportional hazards models (more on the latter below).

We decided to work with these two datasets, and to use a term for “institution effect” in the model.

The survival picture



Quantitative Combination?

So, how can we combine the gene expression data?

Can we put it on the same scale?

Quantitative Combination?

So, how can we combine the gene expression data?

Can we put it on the same scale?

Are the two chips measuring the same things?

Quantitative Combination?

So, how can we combine the gene expression data?

Can we put it on the same scale?

Are the two chips measuring the same things?

Initially, the answer to this last question is NO. The two different Affy chip types involved, the HuGeneFL (Michigan) and the U95Av2 (Harvard) have (a) used different definitions of what constitutes a gene, and (b) used different sets of probes (25-mers) to query the expression levels of said genes.

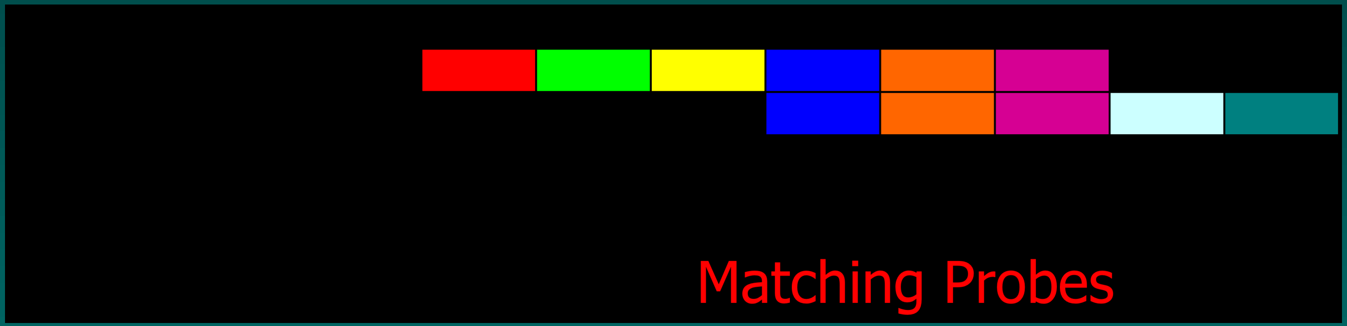
Caveat Emptor

Thus, even though Affymetrix has supplied a list matching probesets from one chip type to probesets on another, the quantitative values that we get will not be directly comparable.

What is directly comparable?

Well, if the individual 25-mer probes are the same, it’s reasonable to see those as querying the same biological targets.

Hence, our first step is to identify all of the 25-mers that are “common” to the two chip platforms, and say that when we focus at this level, the measurements may be the same across the two chip platforms.



What is a gene?

Now we have common 25-mers. But which sets of 25-mers constitute genes? The sets used for the HuGeneFL assembly? For the U95Av2 assembly?

What is a gene?

Now we have common 25-mers. But which sets of 25-mers constitute genes? The sets used for the HuGeneFL assembly? For the U95Av2 assembly?

The above is a trick question. Neither!

When Affy assembled their probesets, they used what they thought to be the best current guesses as to what constituted gene sequences – the sequences associated with different Unigene clusters.

These clusters change over time, as our knowledge of gene structure evolves.

Our approach

In order to use the best and most recent data, we take the common probe sequences, and blast them against the latest build of Unigene. This allows us to assemble our own “pseudo-probesets” which are likely to be more accurate than Affy’s own.

A caveat

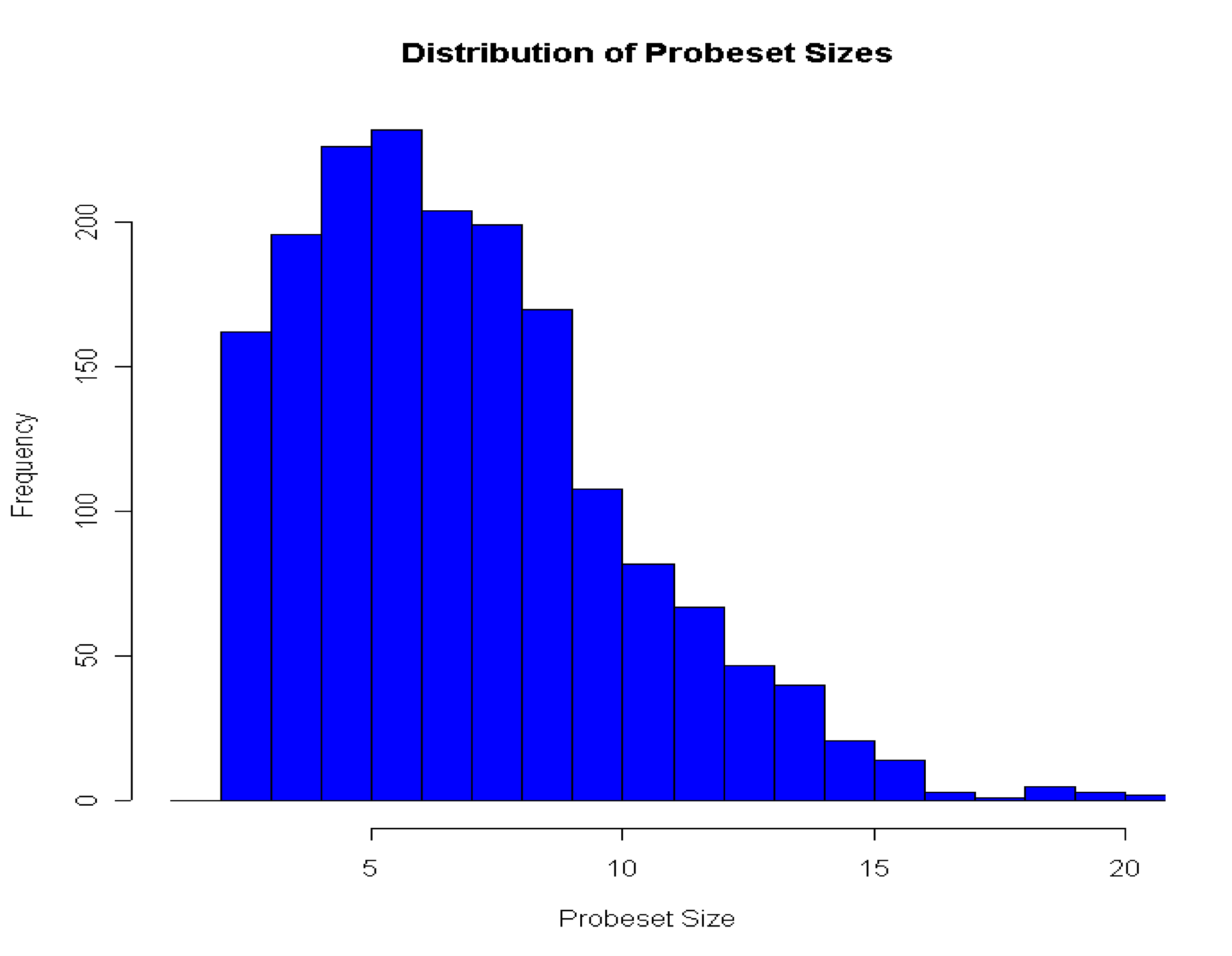
Well, somewhat. While we are measuring the same stuff, we can run into difficulties with the fact that the number of probes we have that are (a) common to the two chip types and (b) associated with the most recent definition of a gene is variable. The number of probes in a probeset (PM only) was 20 for the HuGeneFL and 16 for the U95Av2. Most of our pseudo-probesets have fewer, and we set a lower limit for acceptance of 3 probes.

Is this reasonable?

This lower limit is somewhat ad hoc, but reflects the fact that there are still bits of “crud” that can hit the chips, and these distortions hit individual probes. Thus, there is a non-negligible chance that one probe may be bad; hence more than one is required. In order to decide which of a disagreeing set is more likely, we need a tiebreaker, hence more than two.

All told, when we first assembled the combined list we got 4,101 pseudo-probesests of more than 3 probes to work with.

Our probeset sizes



Quantitative Combination?

Ok, now we have probesets.

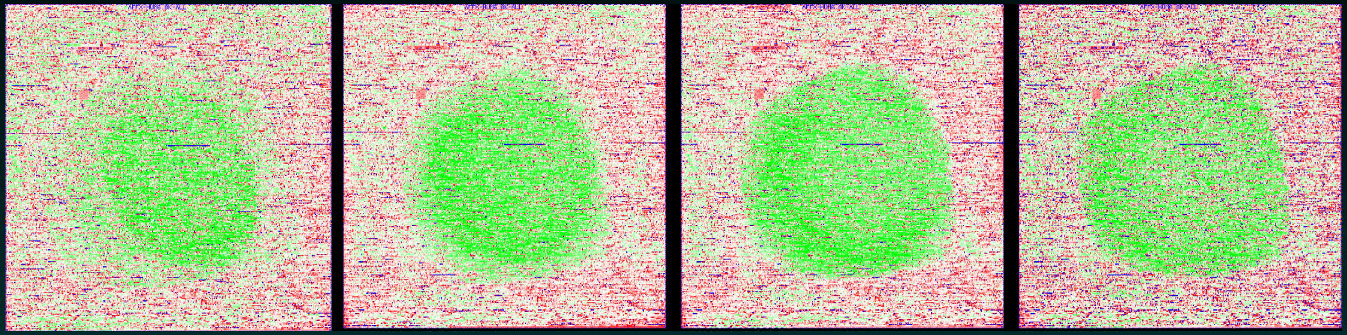
Armed with probesets, we can quantify the gene expression levels across chips.

This does require some new coding of the fitting routines, since R works with the standard CDF files, and we're redefining these mappings.

Given shared probes, we can use our method of choice, be it dChip, RMA, or PDNN (we chose the last one here).

Ok, we have an idea how to quantify all of the chips.

Should we use all of the chips available?



Some of the Michigan chips (L54, L88, L89, and L90 respectively here) show substantial spatial distortion. We want to omit these.

Data Problems

Similarly, some of the Harvard samples were run multiple times, and the average quantifications were used in the initial study. However, samples were run again if the initial sample “looked odd”, so a safer procedure would be to simply use the most recent run of a given Harvard sample.

After filtering out the “odd samples”, we were left with 200 total: 124 from Harvard, 76 from Michigan.

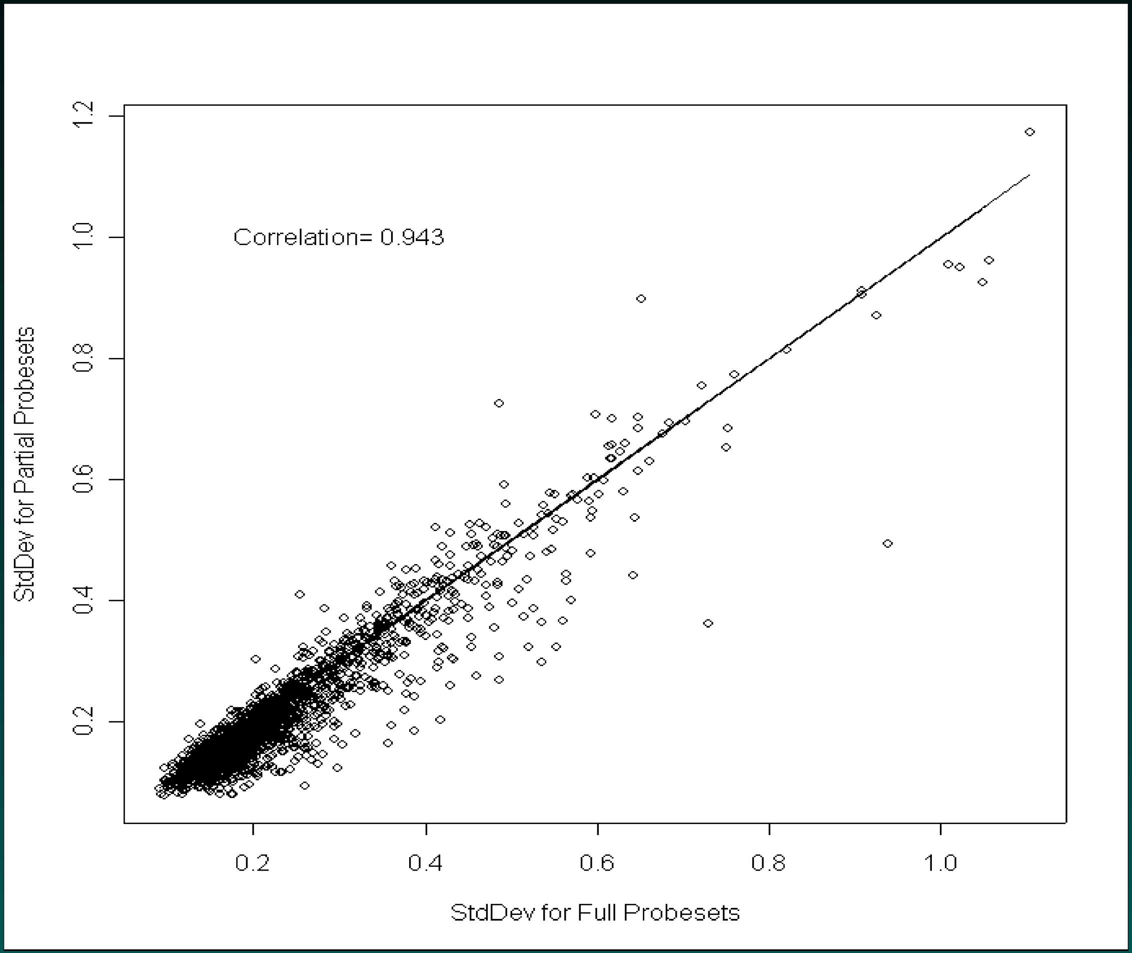
Some gene filters

Remove the weak-signal genes from consideration. Here, this was arbitrarily defined as excluding the half of the genes with the lowest median expression across samples. This step could be improved.

Normalized the chips to have a common mean and standard deviation across genes (using PDNN quantifications to start).

This filter restricts our attention further to 2,055 genes.

Does the approach work?



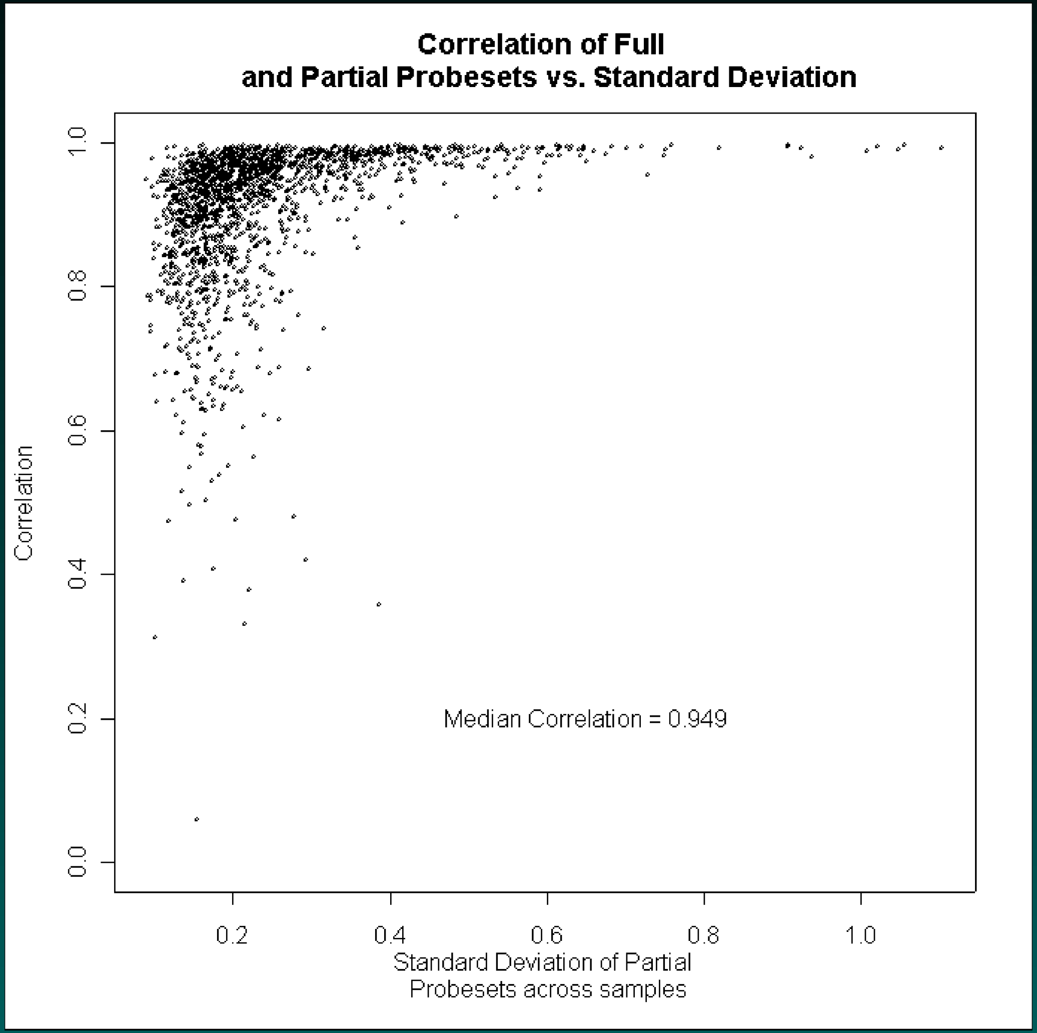
Test 1: variability across samples within a chip type (U95Av2 here); both full and partial probesets.

What are we seeing?

If there is a problem with reproducibility introduced by the method, we would expect to see several spots above the main diagonal, corresponding to more variability when the partial probesets are used.

As it happens, we didn’t see a problem, and if anything the variability went the other way. One potential reason for this stability is the way the probes were chosen for the latter chips – Affy only retained the “good” probes from the earlier chip, and these are more likely to be stable.

Another test



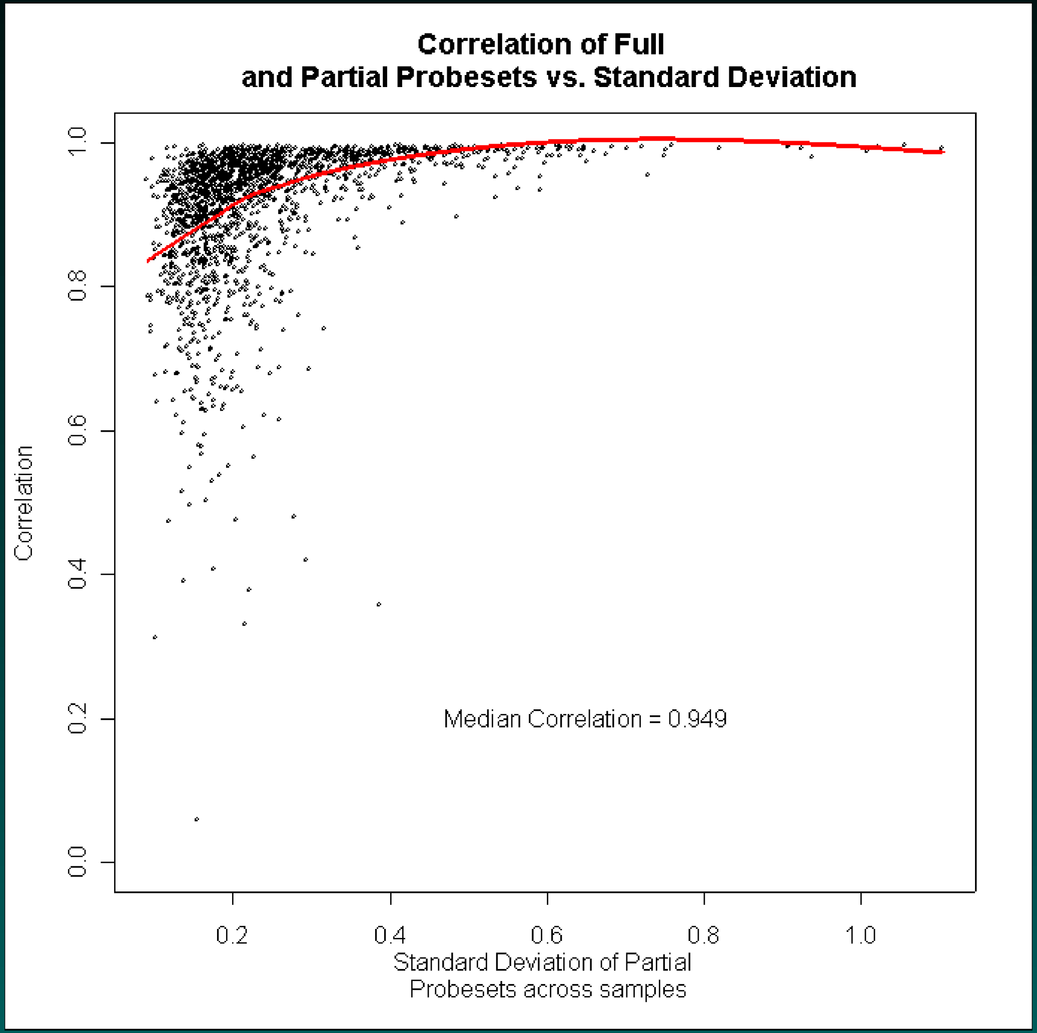
Test 2: rank ordering of expression levels

What we’re seeing

Test 2: look at the rank ordering of expression levels across samples using both the full and partial probesets.

If we’re capturing the signal adequately with just the partial probesets, the rank correlation the two quantifications should be quite high.

Another test, part 2

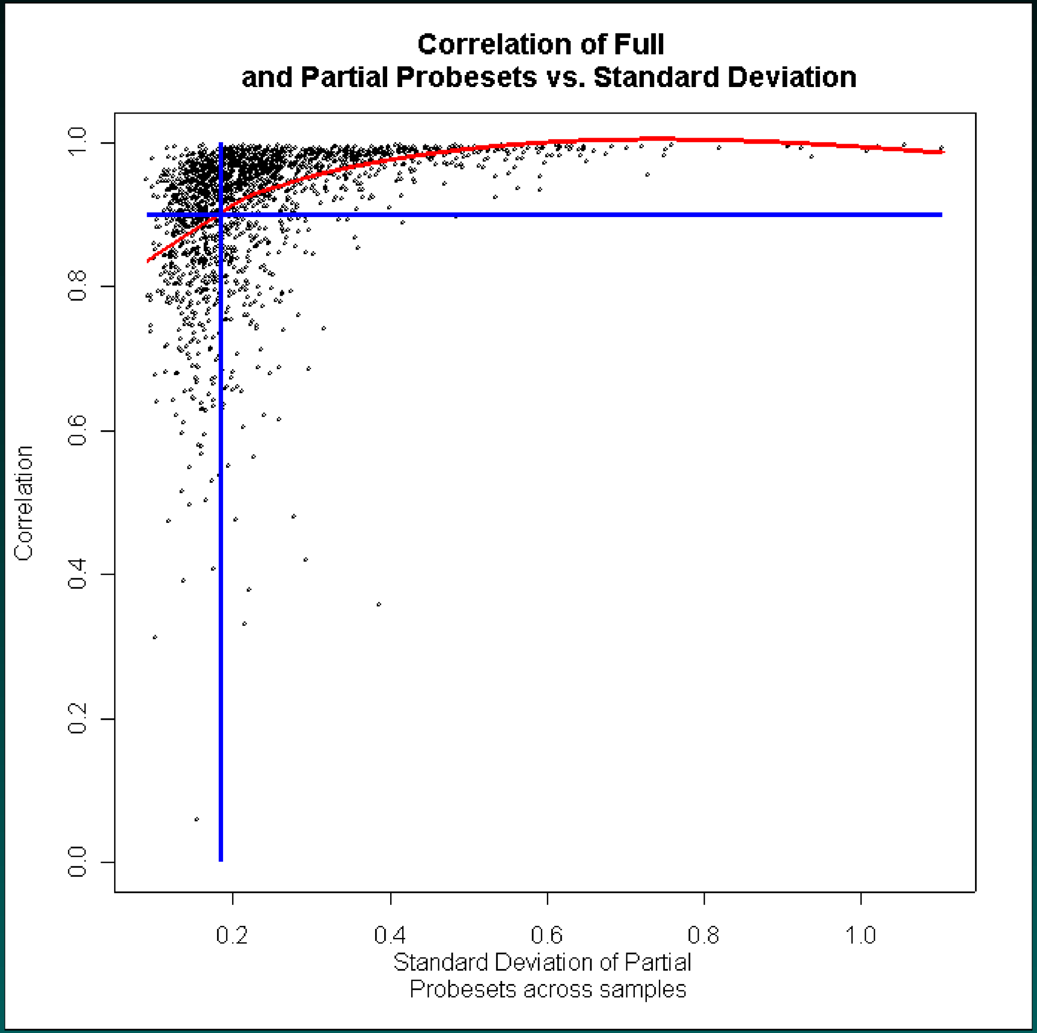


The new details

Fitting a loess curve to the data confirms that the correlations are quite high for the most part, but get worse if the standard deviation of the gene expression levels is low.

This actually makes sense. If the sd is low, the gene expression values are almost the same, and in these cases small fluctuations due to the method can have the most dramatic effects. However, these problematic genes are likely among the least interesting ones from our point of view in that we will not see very big (and hence easily measured) changes.

Another test, part 3

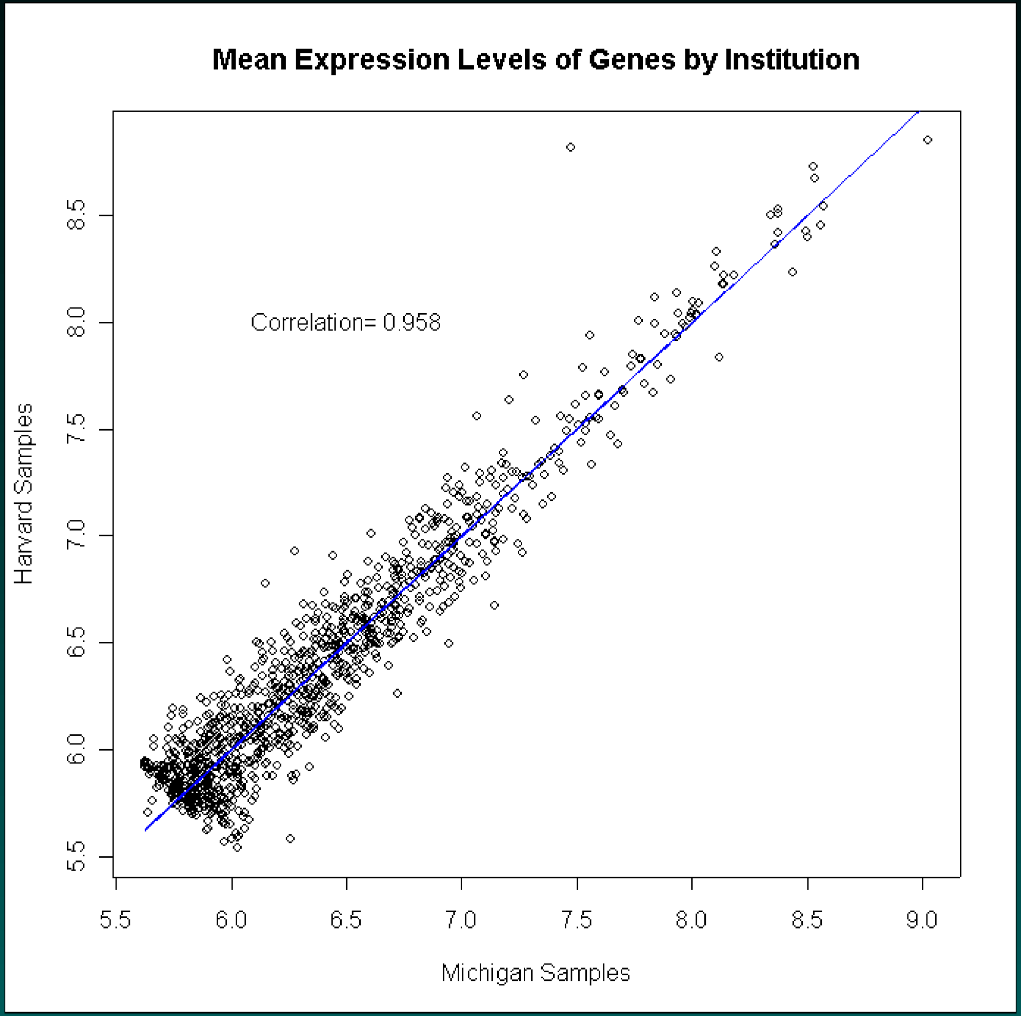


Finishing off the correlation filter

Given that we can identify the most problematic cases, we choose to filter these as well, excluding all genes where the standard deviation of gene measurements across samples is < 0.2 or where the rank correlation between full and partial measurements is < 0.9.

This further filtering reduces our list to 1,036 genes.

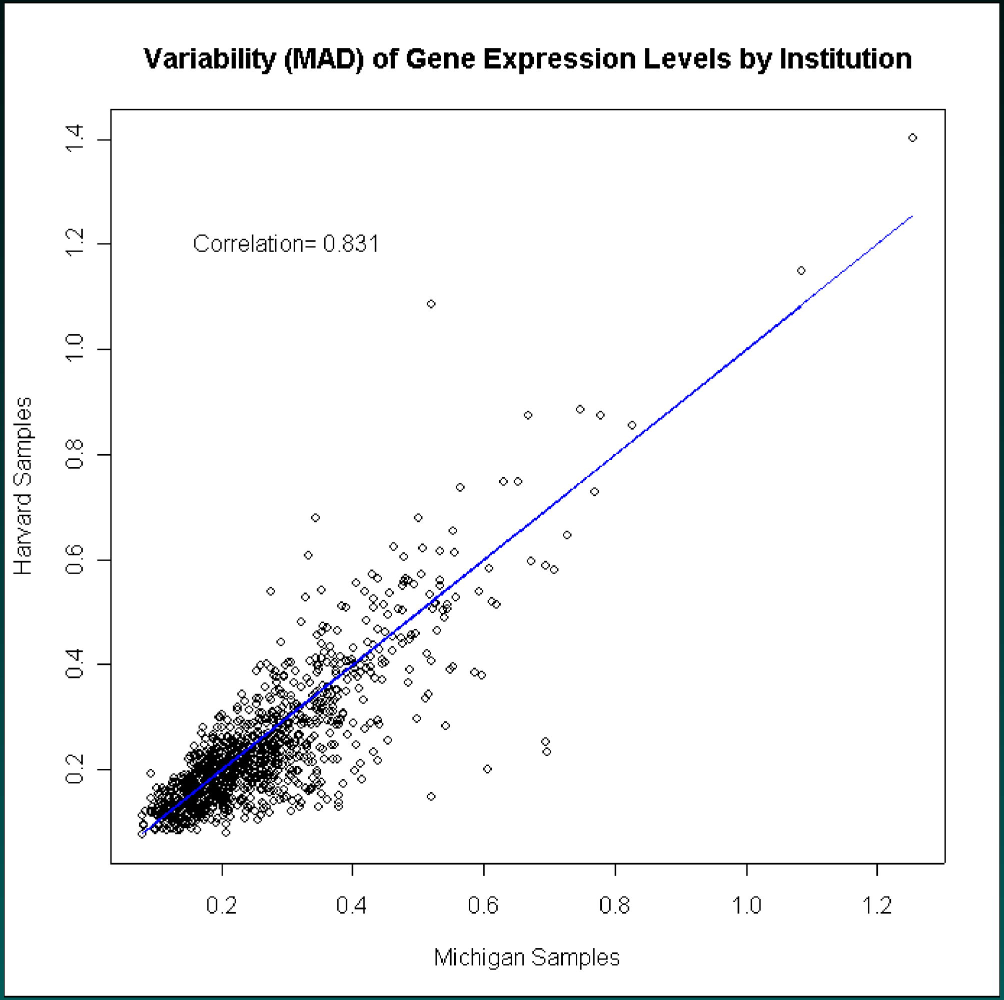
What about average expression?



What about average expression?

We wanted to do a quick check now and see if the quantification values that we’re getting overall tend to agree across institutions. The mean expression levels look quite similar (tightly clustered about the main diagonal. This would probably have shown more clearly using an MA plot.

What about expression variability?



What about expression variability?

The amount of variation in expression levels also appears pretty good across institutions, with a possible slight bias – the Michigan samples are more variable for 61% of the probesets.

So you want to find a gene...

In order to find genes that provide information above and beyond what is supplied by the clinical covariates, we fit Cox proportional hazards models

One of the earlier studies (Michigan) fit univariate Cox proportional hazards models to test the association between gene expression and outcome.

Our approach differs in that we fit a *multivariate* model. The terms we include in our model are study, age, stage, and 1 gene (so we're fitting 4 variables each time).

Computing p-values

We used a permutation based approach to compute p-values. What we permuted were the allocations of gene expression values to the individuals (each individual keeps the same study classification, age, and stage throughout). For each gene, 100000 permutations were performed, and the p-value was defined as the empirical proportion of test statistics for the gene coefficient that were bigger (in absolute value) than the value observed.

Belt and suspenders...

We also checked the p-values produced by two other methods: a model-dependent Likelihood Ratio Test (LRT), and a bootstrap test (like the permutation test, only the expression values are sampled *with replacement.*

Cox Proportional Hazards

So, what is a Cox Proportional Hazards model?

In survival analysis, we often speak of the hazard function, which gives the instantaneous “hazard” of death for an individual, defined as the chance that the individual will die in the very next time interval given that they’ve already survived this far.

Hazard: $\lambda(t) \sim \mathrm{Prob}(X < t + \Delta t | X > t)$

The Cox Model

The Cox model introduces the covariates by assuming that these serve to alter the hazard function in a mulitplicative fashion:

Cox Model: $\lambda_i(t) = \lambda_0(t) \exp(X_i \beta)$

This multiplicative assumption means that the adjustment does not change over time, and means that the relative hazards of two individuals can be defined farily simply:

$\lambda_i(t)/\lambda_k(t) = \exp((X_i - X_k)\beta)$.

$\exp(\beta) =$ change in hazard per unit change in X.

Results using only clinical data:

Factor	β	$\exp(\beta)$	Z	p-val

Study
Michigan = 0
Harvard = 1
	0.67	1.95	2.73	0.0062

Age	0.03	1.03	2.60	0.0094
Stage Early (1-2) = 0 Late (3-4) = 1	1.53	4.61	6.61	< 0.000000001

stay young...

How many genes should we keep?

Ok, we have genes and associated p-values. How many do we decide to investigate?

Identifying Prognostic Genes: use the 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

Calibrating what to expect

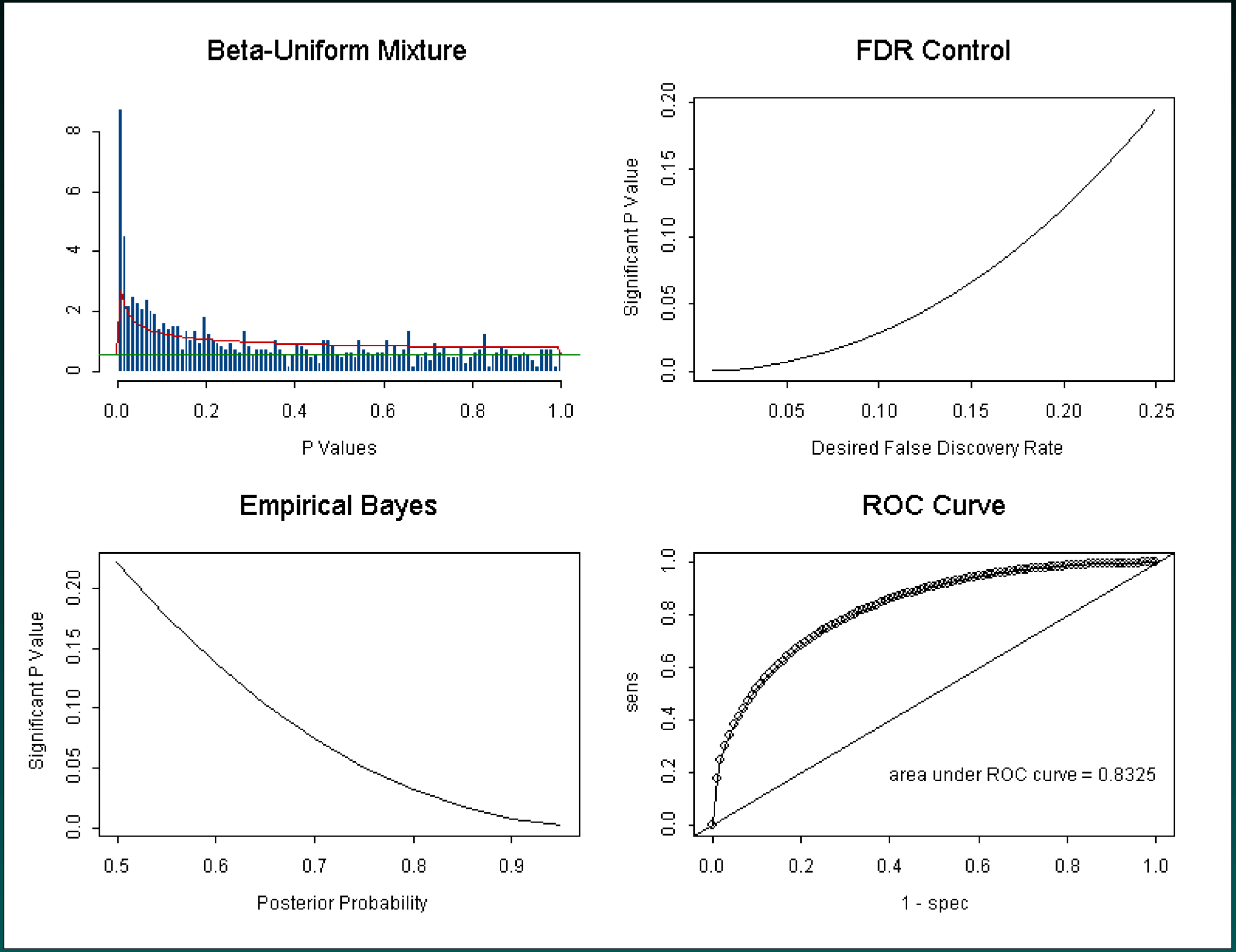
start by using Wilcoxon tests on gene expression levels to identify genes strongly associated with stage. We expect to see several genes showing up here.

Results: Stage-Related Genes

Many genes linked with stage

71 genes flagged using FDR < 0.05 ($p < 0.0064$)

The picture with stage diffs



Checking what we really want to know

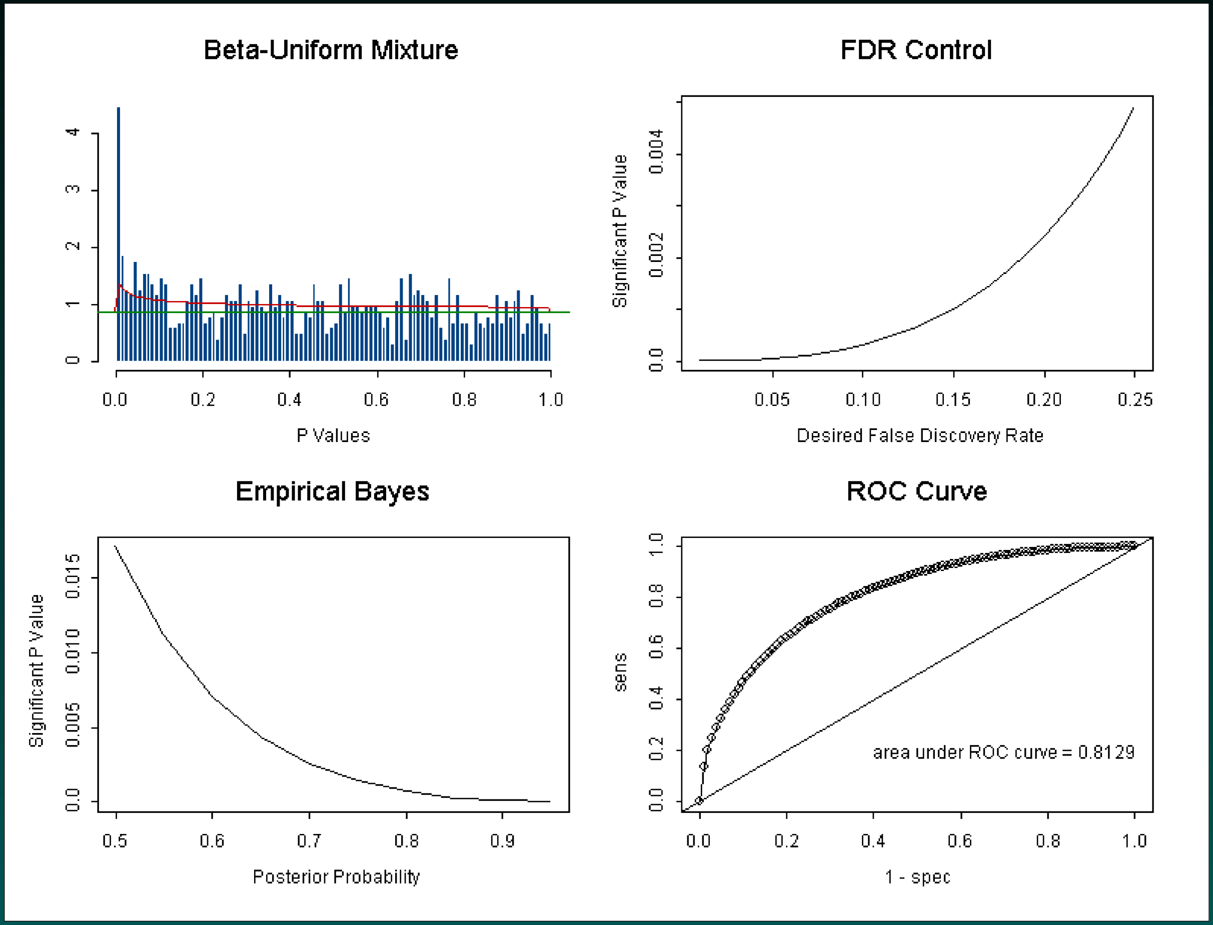
Results: Prognostic Genes

Evidence of some prognostic genes

26 flagged using FDR < 0.20

Note that we have a lot fewer genes here than for stage. Most of this is due to the restriction that these genes should provide information above and beyond that provided by the clinical covariates.

The prognostic picture



What about these genes?

Only 1 also flagged for stage

0 in top 100 genes in Michigan paper

1 cited in Harvard paper

PubMed search on the list of 26 – 14 found to be linked to lung cancer (9/14), cancer in general, or other lung disease.

A partial table

Rank	Gene	β	p	pStage	Function
1	FCGRT	-2.07	¡0.00001	0.154	Induced by IF
2	ENO2	1.46	0.00001	0.282	Marker of NS
4	RRM1	1.81	0.00002	0.321	Linked to surv
8	CHKL	-1.43	0.00010	0.979	Marker of NS
11	CPE	0.72	0.00031	0.088	Marker of SC
12	ADRBK1	-2.20	0.00044	0.484	Co-expresse
16	CLU	-0.52	0.00109	0.014	Marker of SC
20	SEPW1	-1.29	0.00145	0.028	H202 cytotox.
21	FSCN1	0.66	0.00150	0.082	Marker of inva

A partial table

Rank	Gene	β	p	Function
1	FCGRT	-2.07	$<$0.00001	Induced by IF-g in trea
2	ENO2	1.46	0.00001	Marker of NSCLC
4	RRM1	1.81	0.00002	Linked to survival in N
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 Co
16	CLU	-0.52	0.00109	Marker of SCLC
20	SEPW1	-1.29	0.00145	H_20_2 cytotox. in NSCL
21	FSCN1	0.66	0.00150	Marker of invasivenes

and some more

Rank	Gene	β	p	pStage	Function
3	NFRKB	-2.81	0.00001	0.058	Amplified in AM
7	ATIC	1.81	0.00009	0.771	Fusion partner o
13	BCL9	-1.64	0.00069	0.057	Over-expressed
15	TPS1	-0.64	0.00107	0.882	Associated with
25	BTG2	-0.75	0.00232	0.726	Inhibits cell proli

and some more

Rank	Gene	β	p	Function
3	NFRKB	-2.81	0.00001	Amplified in AML
7	ATIC	1.81	0.00009	Fusion partner of ALK wh defines subtype of ALCL
13	BCL9	-1.64	0.00069	Over-expressed in ALL
15	TPS1	-0.64	0.00107	Associated with pulmona inflammation
25	BTG2	-0.75	0.00232	Inhibits cell prolif in primary mouse embryo fibroblasts lacking functional p53

Discussion

All in all, the results worked out rather well. We found some novel things, and they seem to make sense.

Further, some of our gains could not have been realized using just one of the two studies. Of the 26 we found, we would only find 8 using the Harvard data alone, or 1 using the Michigan data alone.

Many of our gains derive from the fact that we are combining the data at the raw (CEL file) as opposed to summary (probeset quantification) levels. The downside is that this is harder. The upside is that the biology is there.