GS01 0163
Analysis of Microarray Data

Keith Baggerly and Brad Broom
Department of Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
bmbroom@mdanderson.org

16 November 2010

Lecture 22: Frozen Barcodes

- How big are batches?
- Are there questions or measures that are robust to batches?
- Can we account for them using other info?
- Barcode, Zilliox and Irizarry, Nat Meth 4:911-3, 2007
- fRMA – McCall et al, Biostatistics, 11:242-53, 2010
- Barcode 2 – McCall et al, JHSPH WP 200, 2009
- TCGA

Batches and Databases

A big reason we've been harping on batch effects is that (a) they're big, and (b) as more data gets assembled, dealing with them becomes more important (e.g., TCGA).

That said, batches are only a problem as long as we don't understand them. If we understand how bad they can be and what their characteristics are, we can either account for them or declare the whole exercise a waste of time.

For some types of arrays, we now have a *lot* of data. Can we use this to learn about batches?

Channelling...

Most of the stuff I'll be talking about today comes from Rafa Irizarry's lab at Johns Hopkins. So just imagine I'm him for the remainder of the lecture.

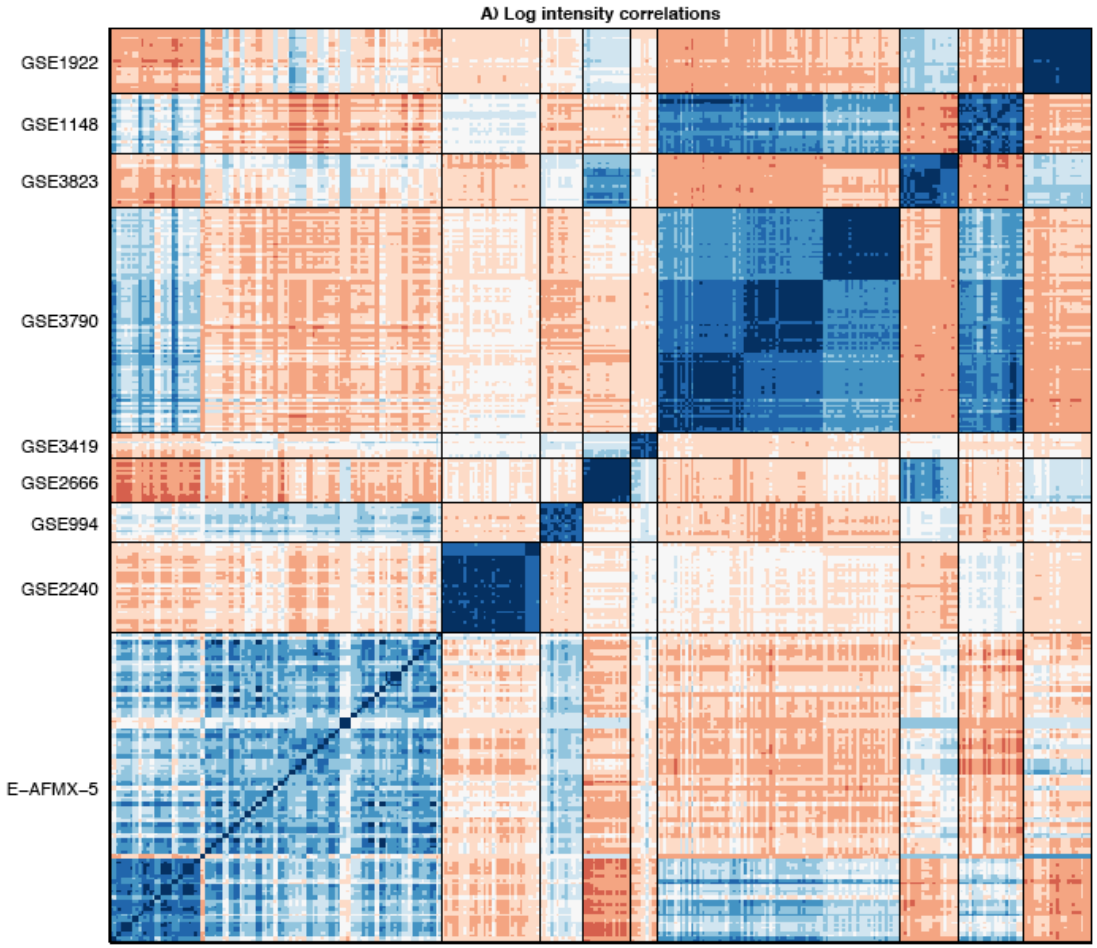
GEO and ArrayExpress

For the past several years, the Nature, Science and Cell journal families have all adhered to the MIAME standards, and required that array data be deposited in a public database: either GEO or ArrayExpress.

For Affy U133A and U133+2, there are now literally hundreds of CEL files available from several labs and tissue types.

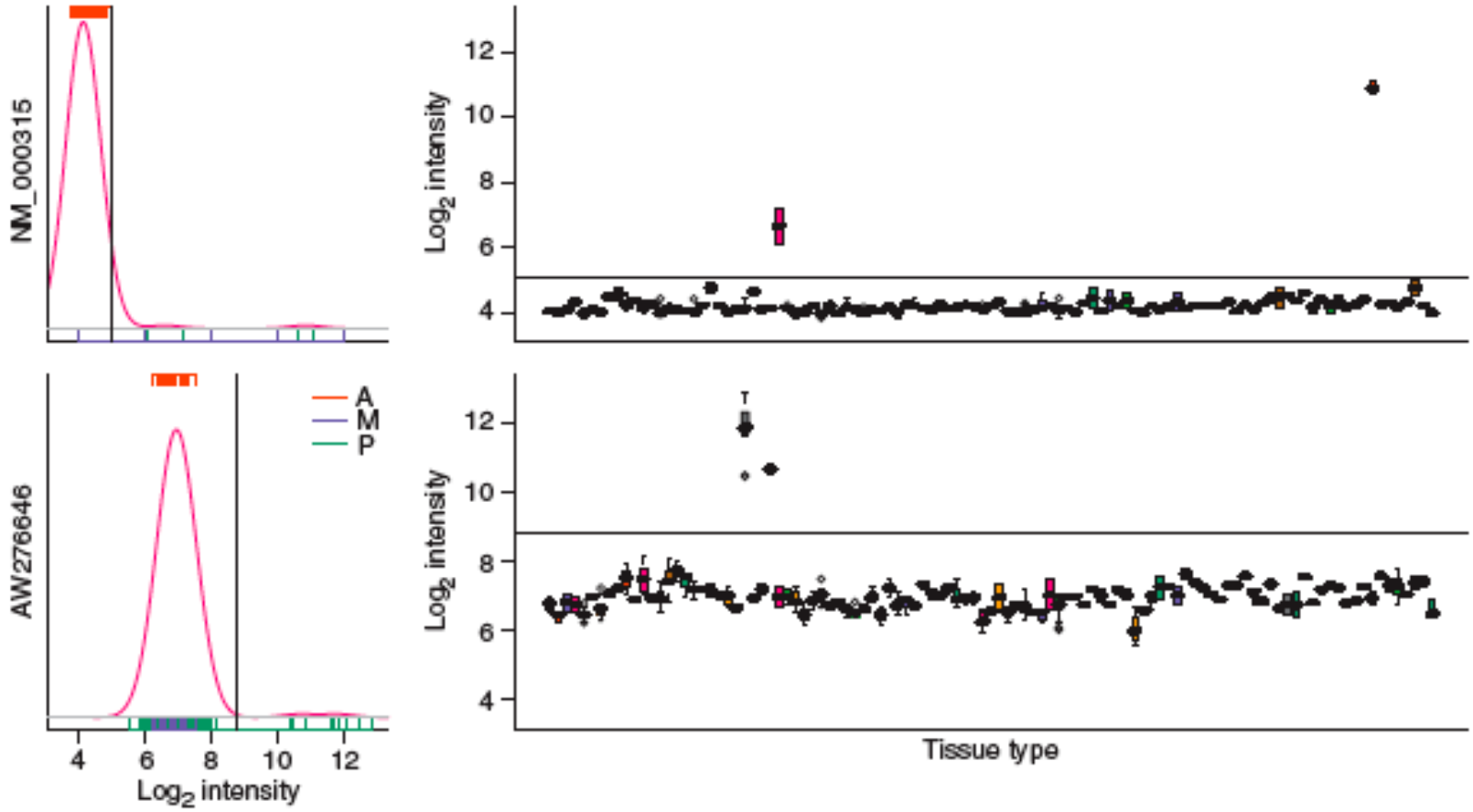
If we process all of these arrays in the same way, we may be able to get a better feel for the extent of batch effects.

Correlating GEO



Zilliox and Irizarry, Supp Fig 3A. Correlating GEO data. This is disturbing. Note E-AFMX-5.

Probes Across Many Arrays



Zilliox and Irizarry, Fig 1. Which genes are Present? Absent?

When is a Gene Really Present?

Assembling the barcode: 1092 samples from 118 tissue types from 40 different studies. Of the samples, 500 are breast tumors, 498 are normal tissue, and 94 are other disease.

Fit a density estimate to the data across experiments. Record all local maxima. Declare the lowest maximum to correspond to “no expression”.

Note variation in “null” levels.

Setting Cutoffs

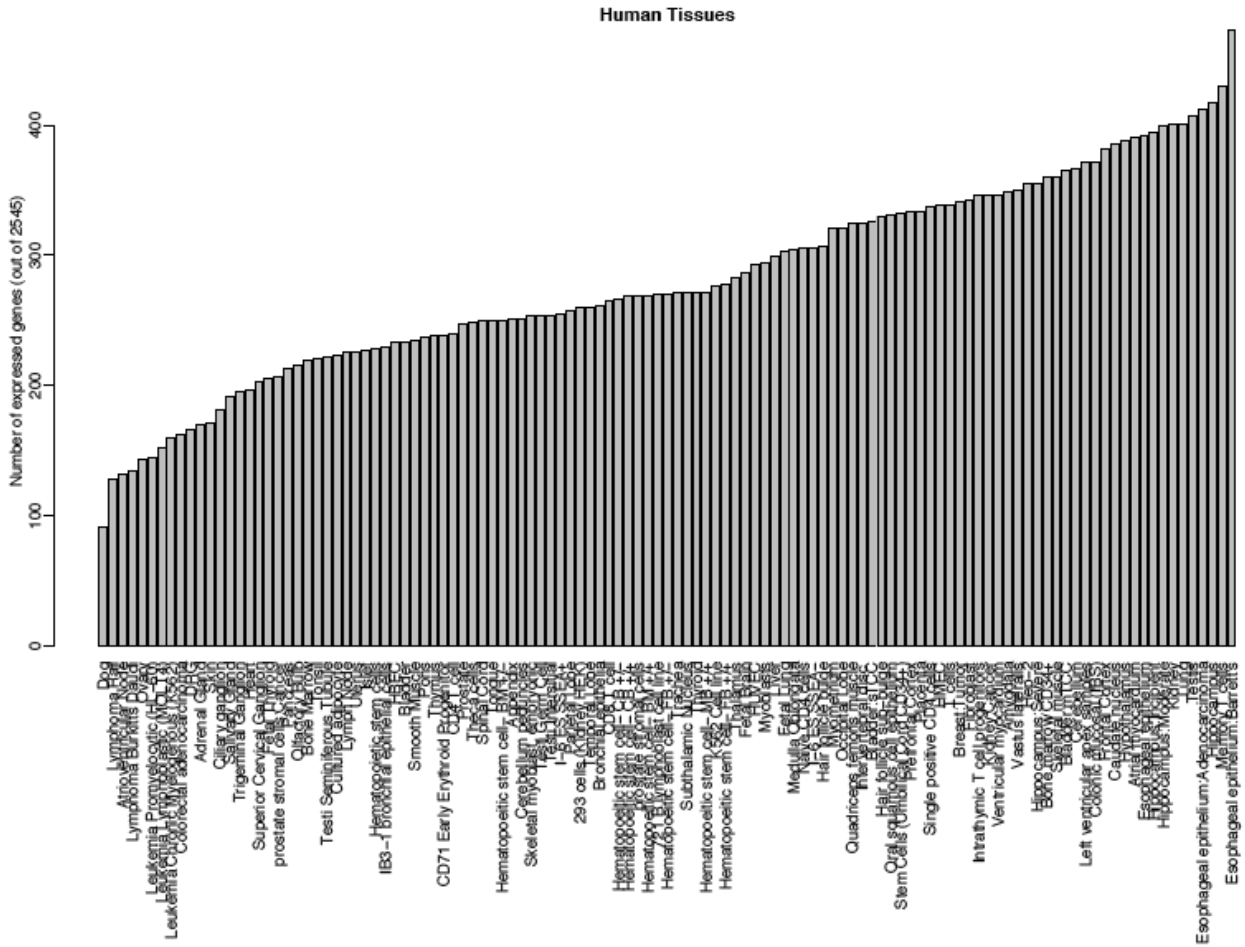
Use the lower half of the bottom mode to estimate the standard error, and flag a gene as “high” if it is more than K standard deviations above the no expression level (they used K = 6).

Keep only bimodal genes for classification (2519 for U133A).

What are most of the barcode genes? About 3/4 encode membrane or extracellular proteins.

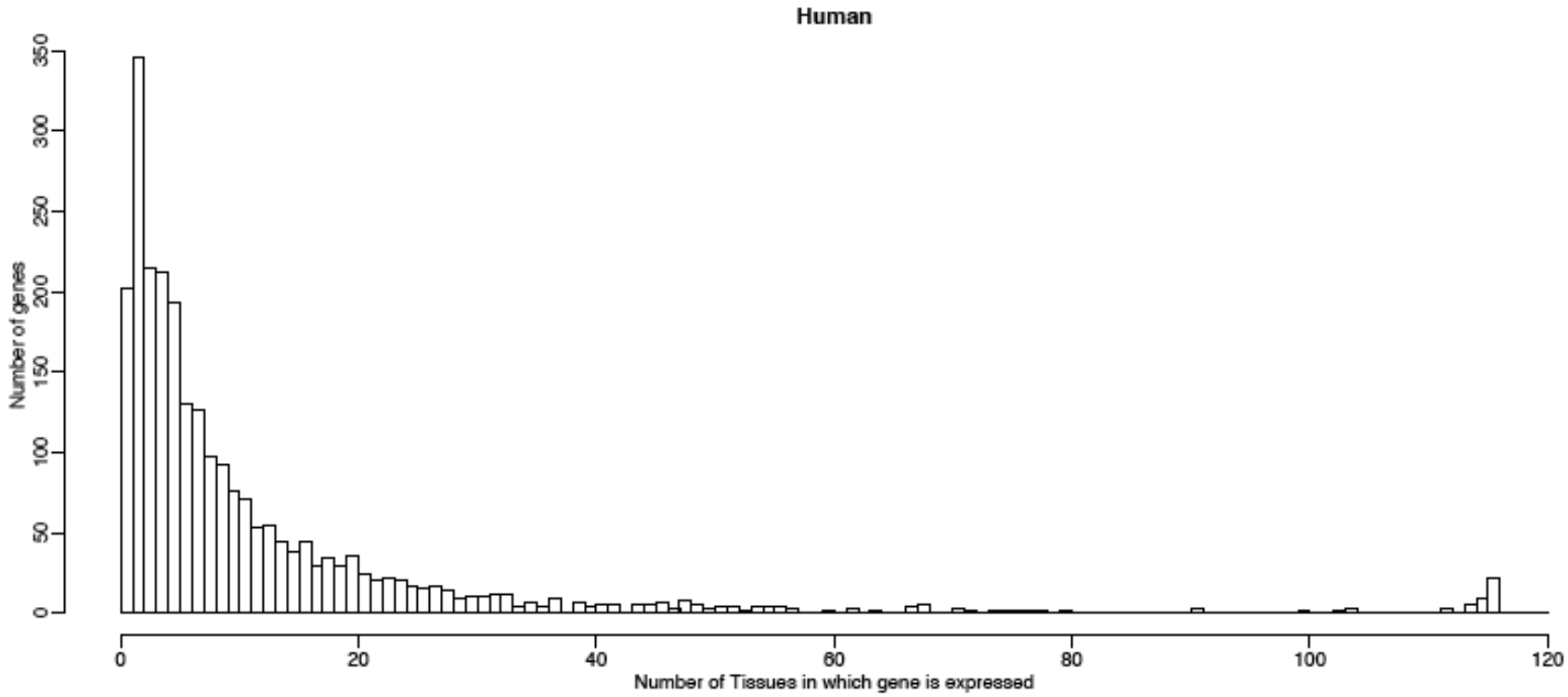
Store mean, standard deviation on a probeset by probeset basis, now that we’ve assembled them.

How Many Genes Characterize Each Tissue?



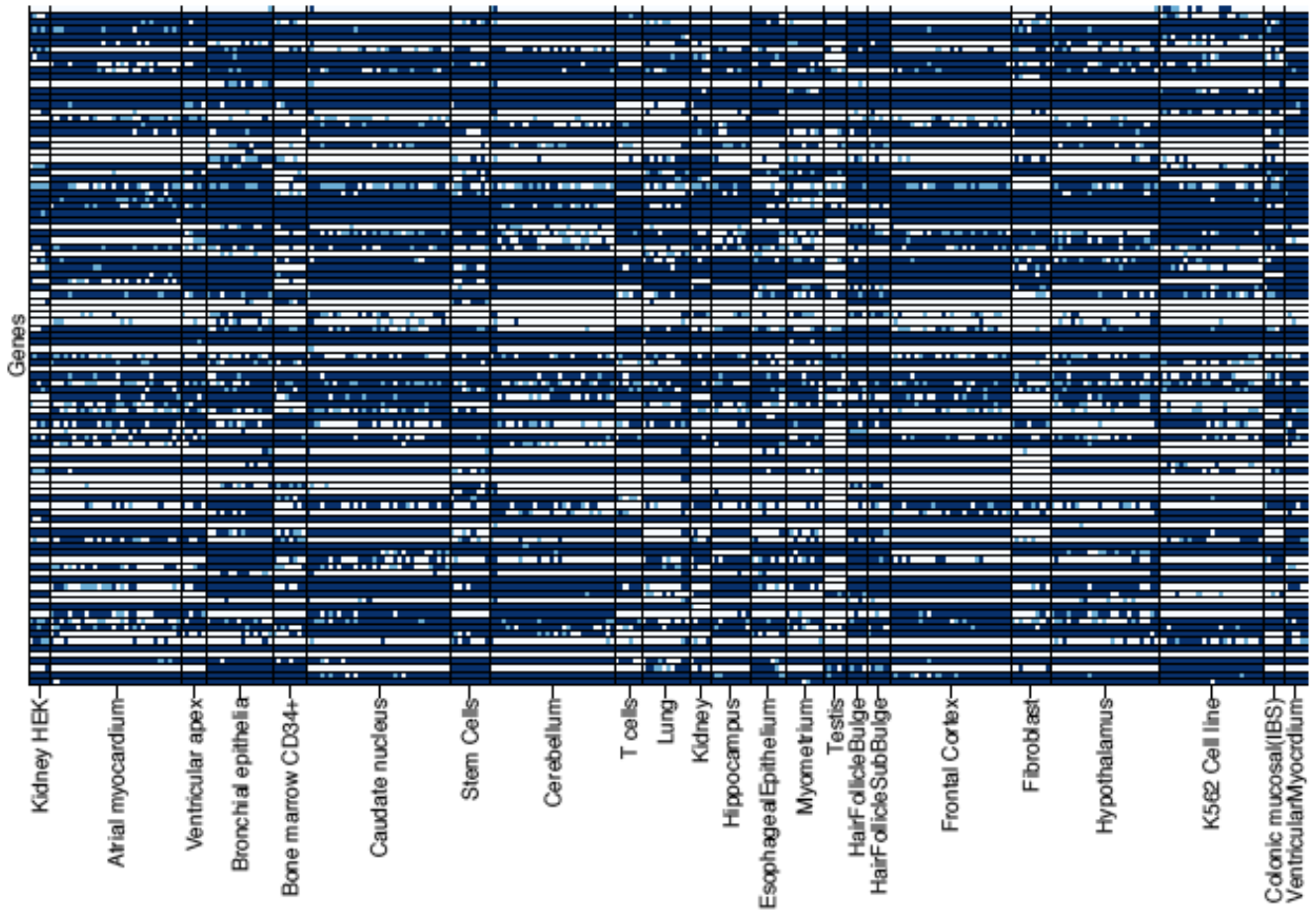
Number of high genes by tissue

How Many Tissues Is a Gene High In?



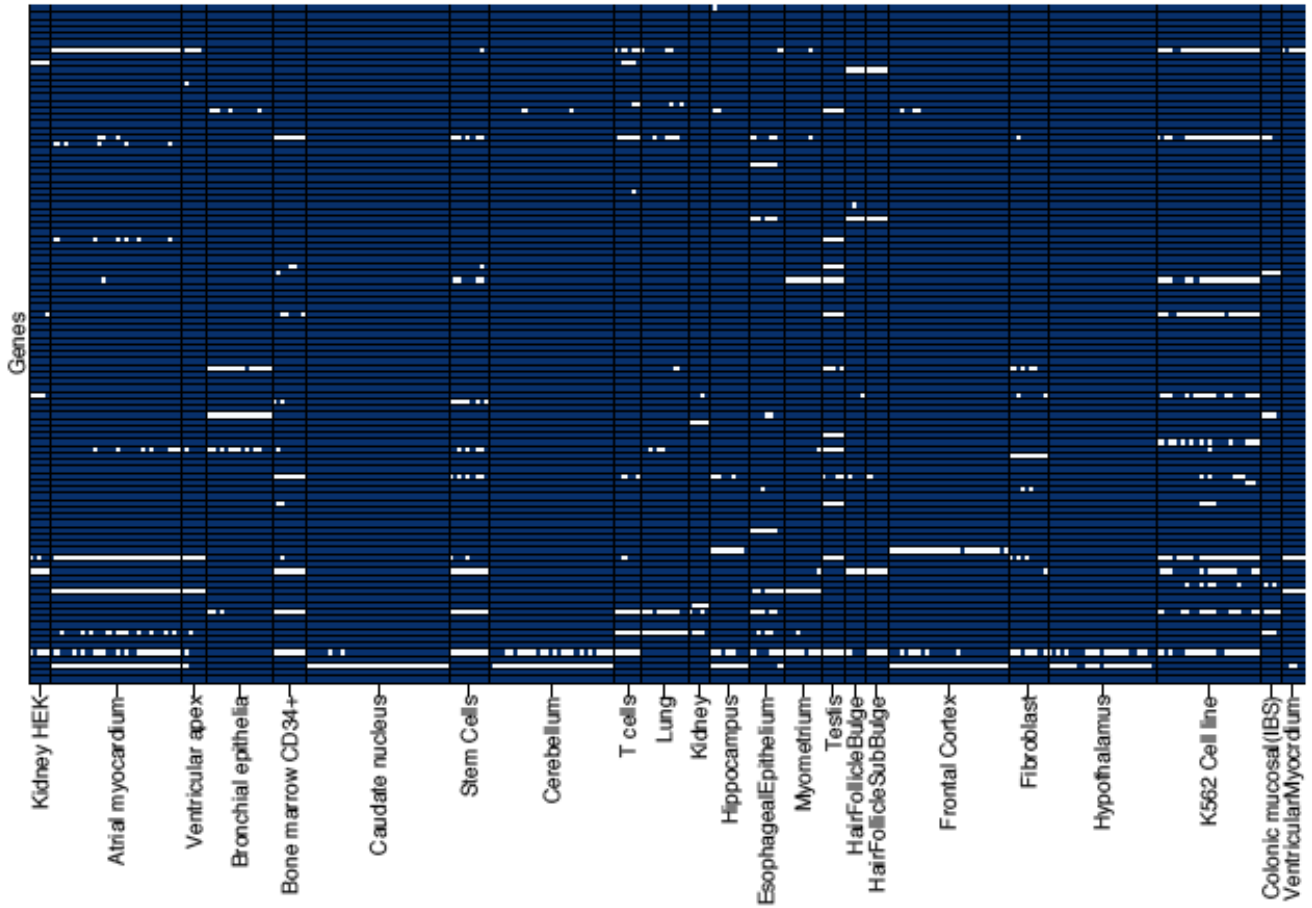
Number of high tissues by gene

How do Present/Absent Calls Do?



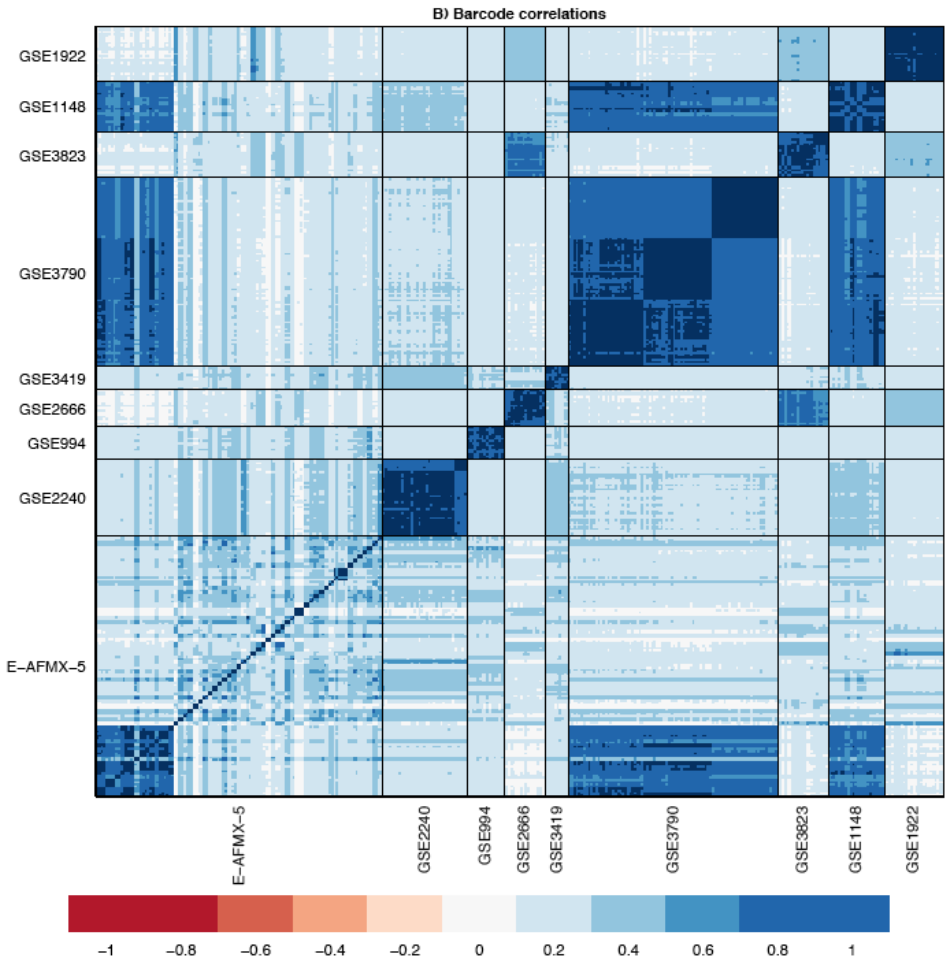
Defined one array at a time, largely driven by intensity (100 random genes).

How do Barcode Present/Absent Calls Do?



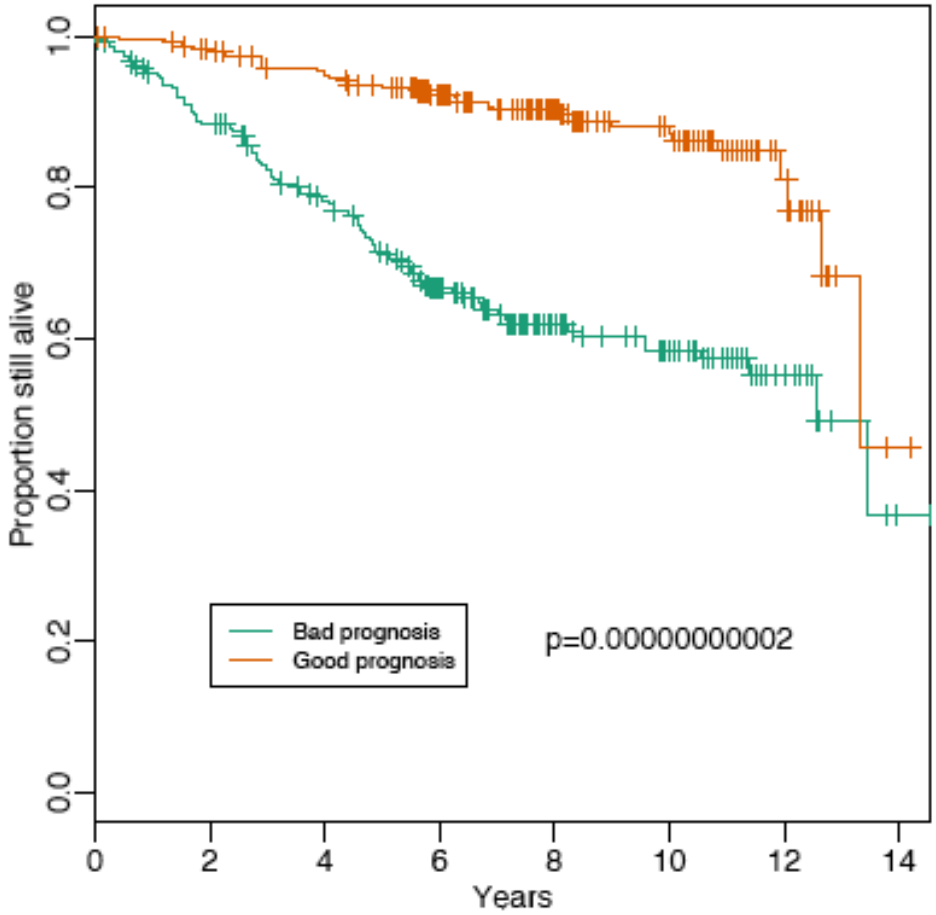
Defined cross-study. Note broad blueness.

Barcode Correlations?



Zilliox and Irizarry, Supp Fig 3B. Barcoding GEO data. Much better.

Barcode Survival



Zilliox and Irizarry, Supp Fig 5A. Barcoding normal and tumor breast, then coding tumors.

Does it Work?

In general, this is pretty robust, and pretty simple.

That said, we've had several ovarian samples classified as breast. This may reflect the preponderance of breast samples used in assembling the classifier, or that we didn't score our samples in the same way.

How Did They Score Samples?

They like RMA. But.

They have over 1000 CEL files to work with. What problem do you think they encountered?

What are other problems with RMA?

How can we use RMA for a single sample (eventually required for clinical applications)?

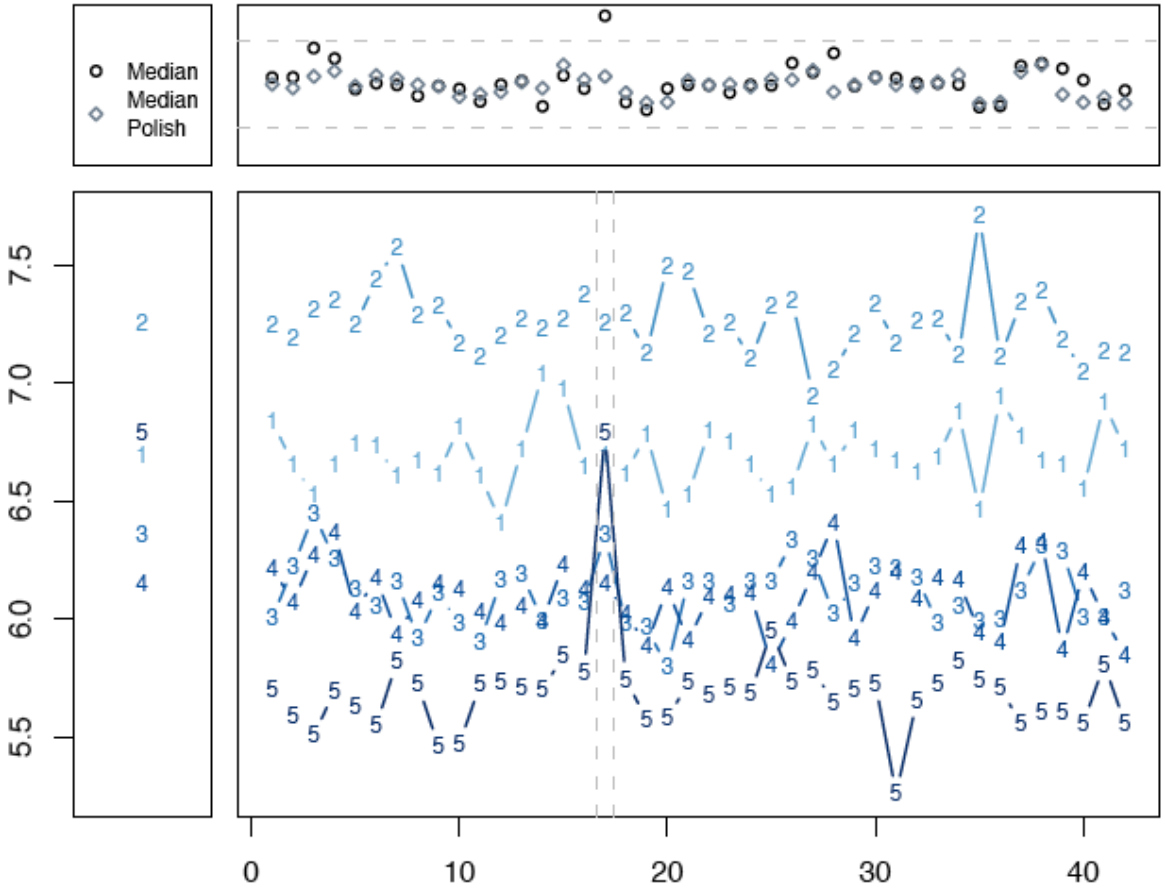
What's Good About RMA?

The model is simple, and it learns.

$$Y_{ijn} = \theta_{in} + \phi_{jn} + \epsilon_{ijn},$$

for probe j in probeset n on array i. We're interested in the sample-specific intensities θ, but we have to deal with the probe effects ϕ and probeset errors ϵ_{ijn}.

Picturing Benefits



McCall et al, Fig 1. Robust learning. But how much learning do we need?

Start With a Reference Set

Katz et al., BMC Bioinformatics, 7:464, 2006.

Run RMA on a single large database, and record (a) the overall quantile vector, and (b) the estimated probe effects $\hat{\phi}$. These values are now fixed, or *frozen*.

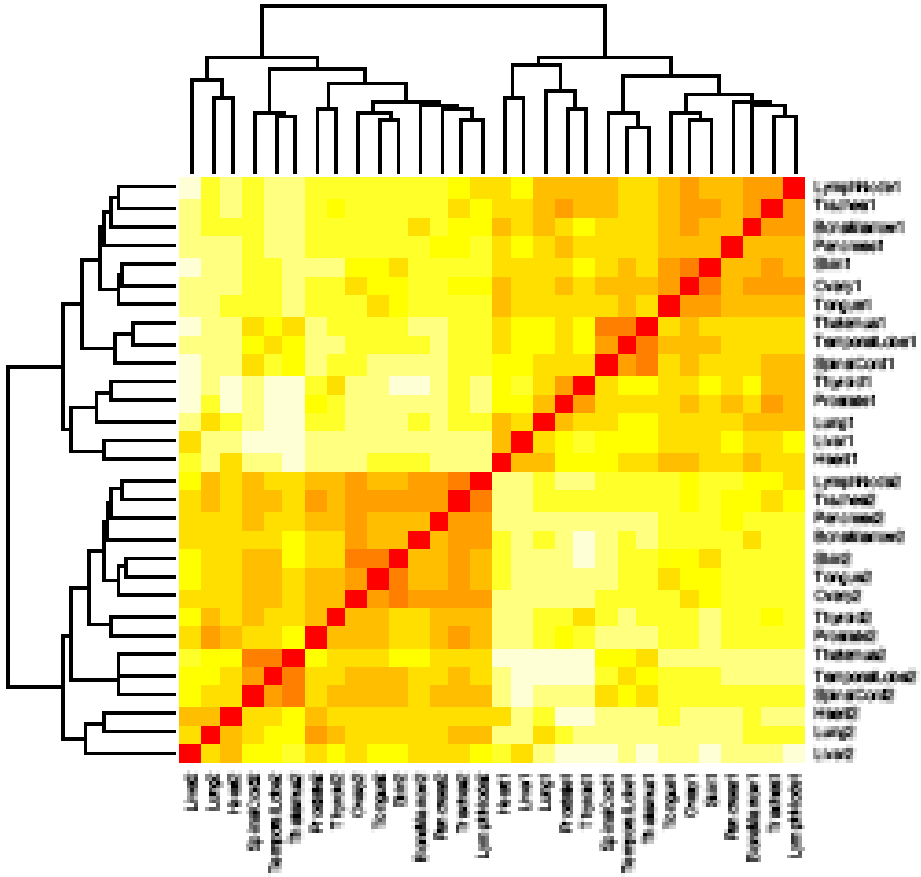
For a given new array, report the median of

$$y_{ij} - \hat{\phi}$$

as the sample estimate.

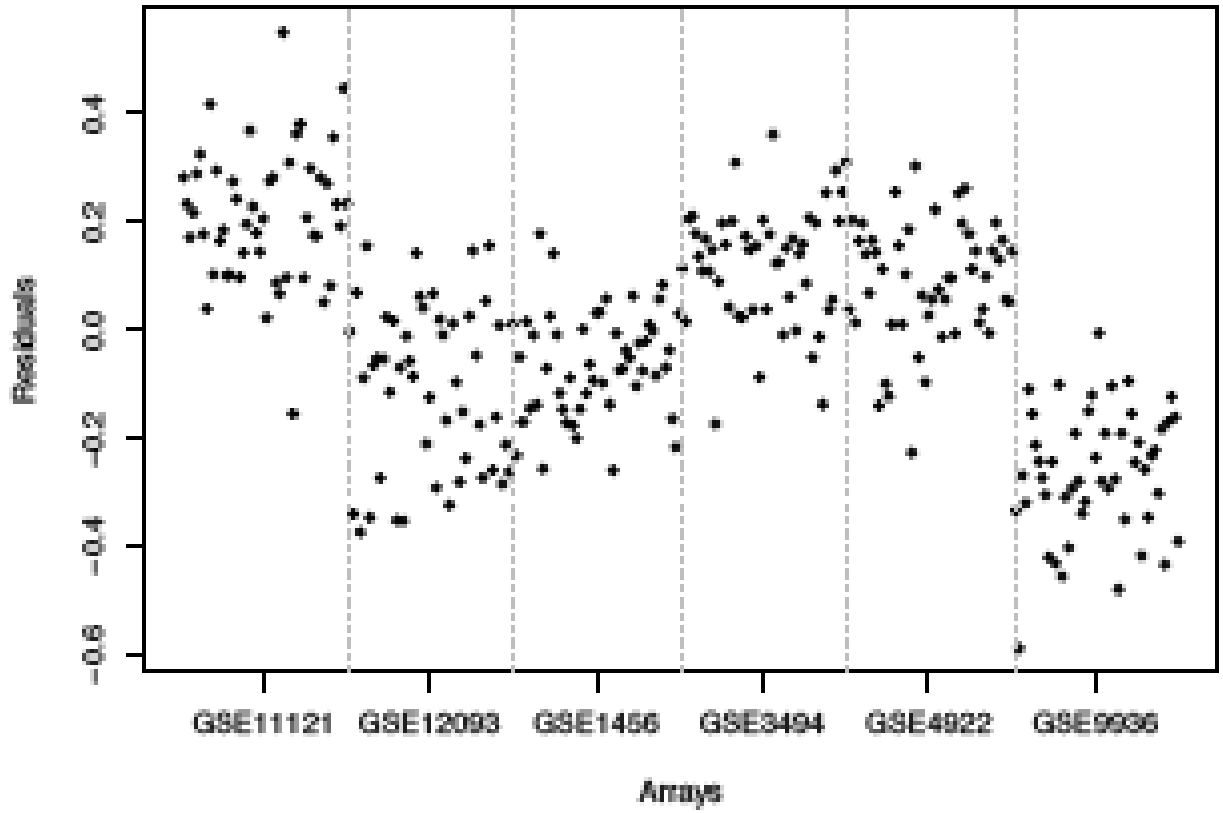
Does it work?

Why Might it Fail?



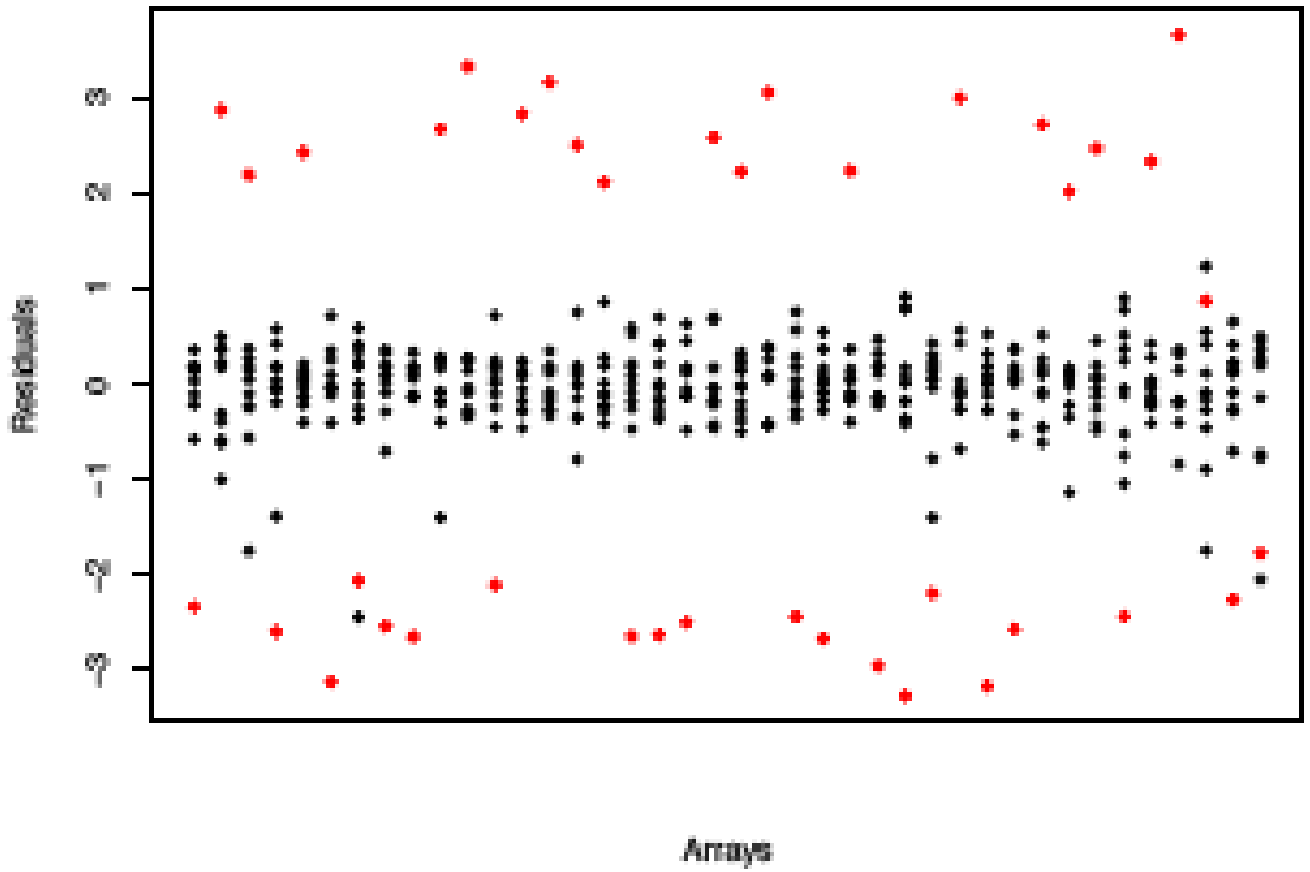
McCall et al, Fig 4A. What's driving the clustering?

Residuals for One Probe



McCall et al, Fig 2B. Batch effects are large, and can corrupt further estimates.

and On a Related Note



McCall et al, Fig 3B. Some probes are simply far more variable than others.

Revising the Model

$$Y_{ijkn} = \theta_{in} + \phi_{jn} + \gamma_{jkn} + \epsilon_{ijkn}$$

where k denotes batch, γ is a random effect with variance $Var(\gamma_{jkn}) = \tau^2_{jn}$, and the random error is probe-specific: $Var(\epsilon_{ijkn}) = \sigma^2_{jn}$.

So, how do we estimate all of the terms?

Start with the Familiar

Begin with standard RMA to get starting estimates of $\hat{\theta}_i$ and $\hat{\phi}_j$. Then, use the residuals to estimate the between and within batch variation terms.

The reference set was chosen to balance tissue and experiment, with 5 arrays from 170 distinct combinations.

Estimating One New Array

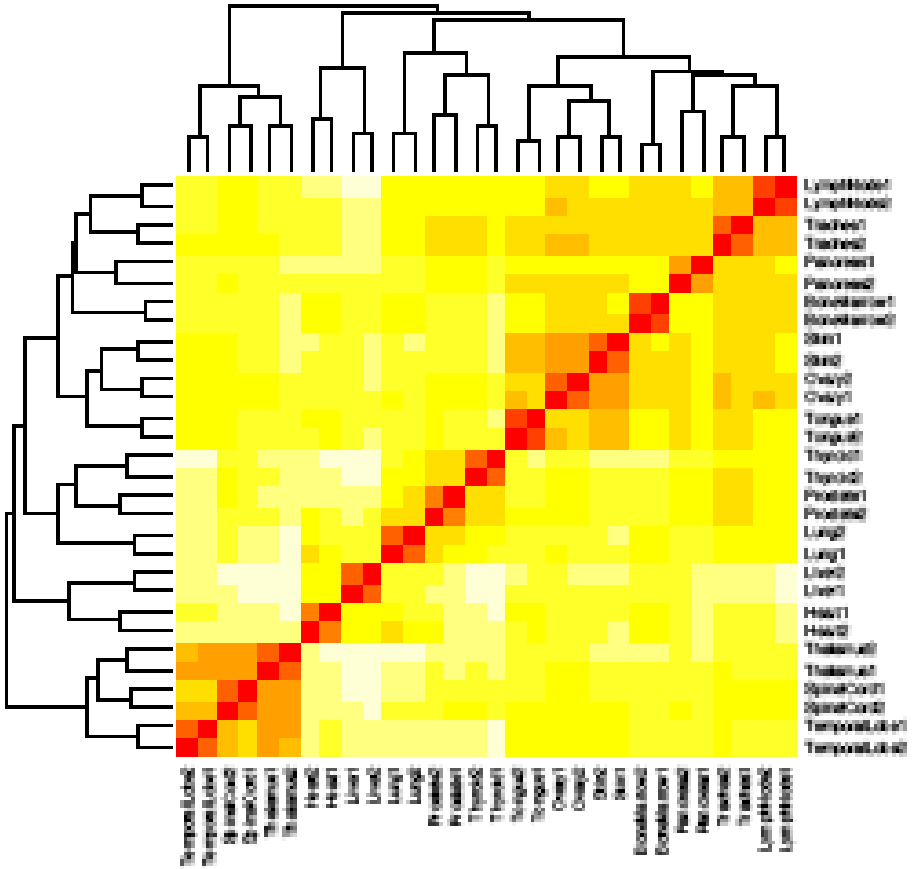
Given that we have probe-specific batch and variance terms,

$$\mathrm{Var}(\hat{Y}_{jn}) = \tau_{jn}^2 + \sigma_{jn}^2$$

use these to produce a weighted mean from the individual probe values (with weights inversely proportional to the variance).

Does it work?

Clustering Redux



McCall et al, Fig 4B. What's driving clustering now?

Can We Do Better With More Arrays?

Sure. More information is (almost) always good.

If we have a new batch of arrays (and recognize the batch boundaries), then we can estimate the batch effect for all of the arrays and render our estimates more precise (effectively removing the τ^2 term.

Our estimates are correlated, but again it boils down to a weighted mean.

Are These Easy to Use?

Good question. The packages are still in early release form, but they are R packages nonetheless.

Since the barcode weights were initially assembled using refRMA as opposed to fRMA, it would probably be good to go back and reweight things (which I suspect they're doing or have done).

Revisiting the Barcode

It's been two years since the initial barcode paper. Since then, fRMA has been better locked down, and more extensive reference databases have come online.

Can we come up with a better definition?

Can we keep more genes?

Can we use barcodes for more than tissue type?

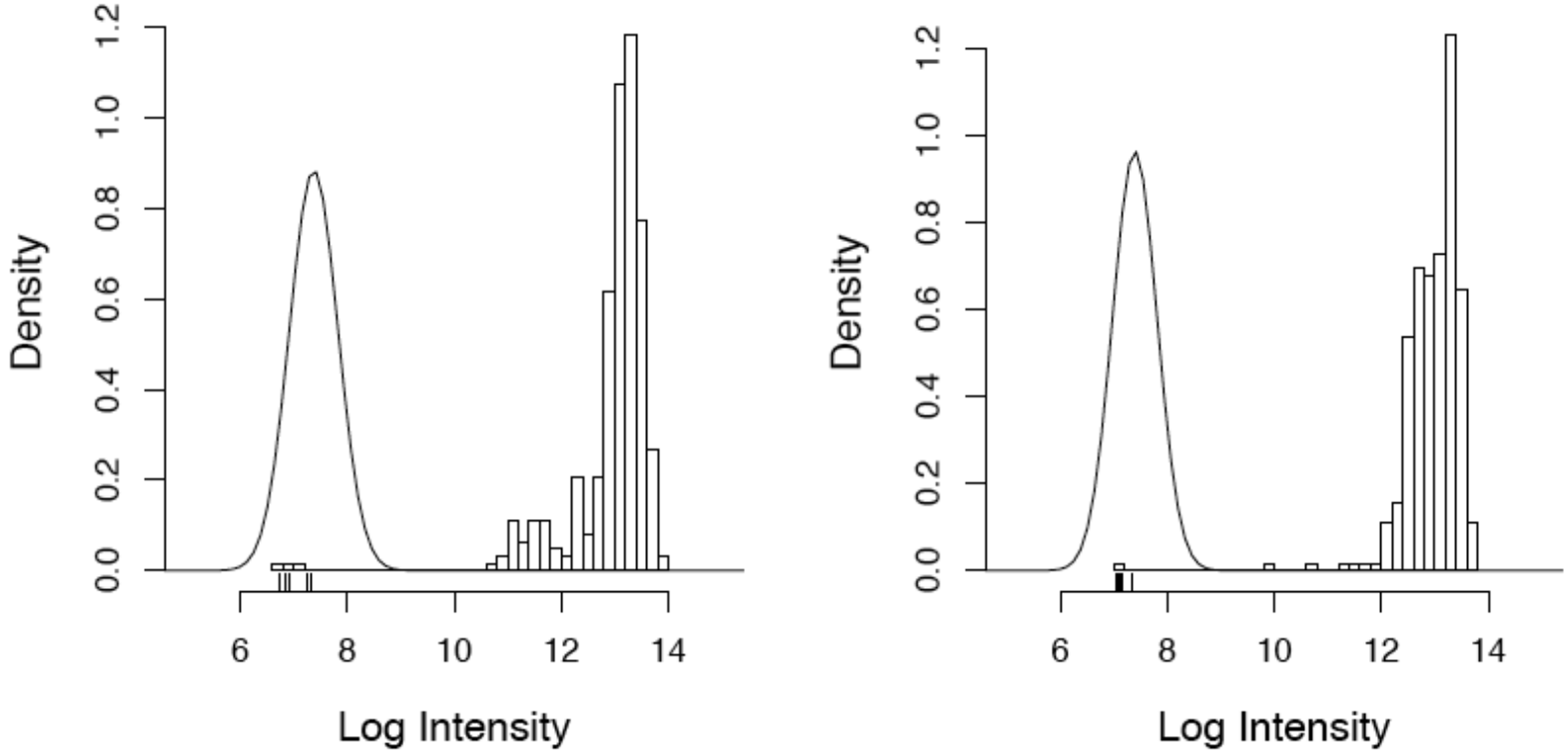
New Data

316 tissue types profiled at least 5 times in GEO; 8277 U133A arrays in all.

All arrays were processed with fRMA to give expression values.

Also ran 5 yeast samples on human arrays to get a better idea of “null” expression.

The Vital Yeast



McCall et al (WP 200), Fig 3. Yeast levels clearly identify null expression.

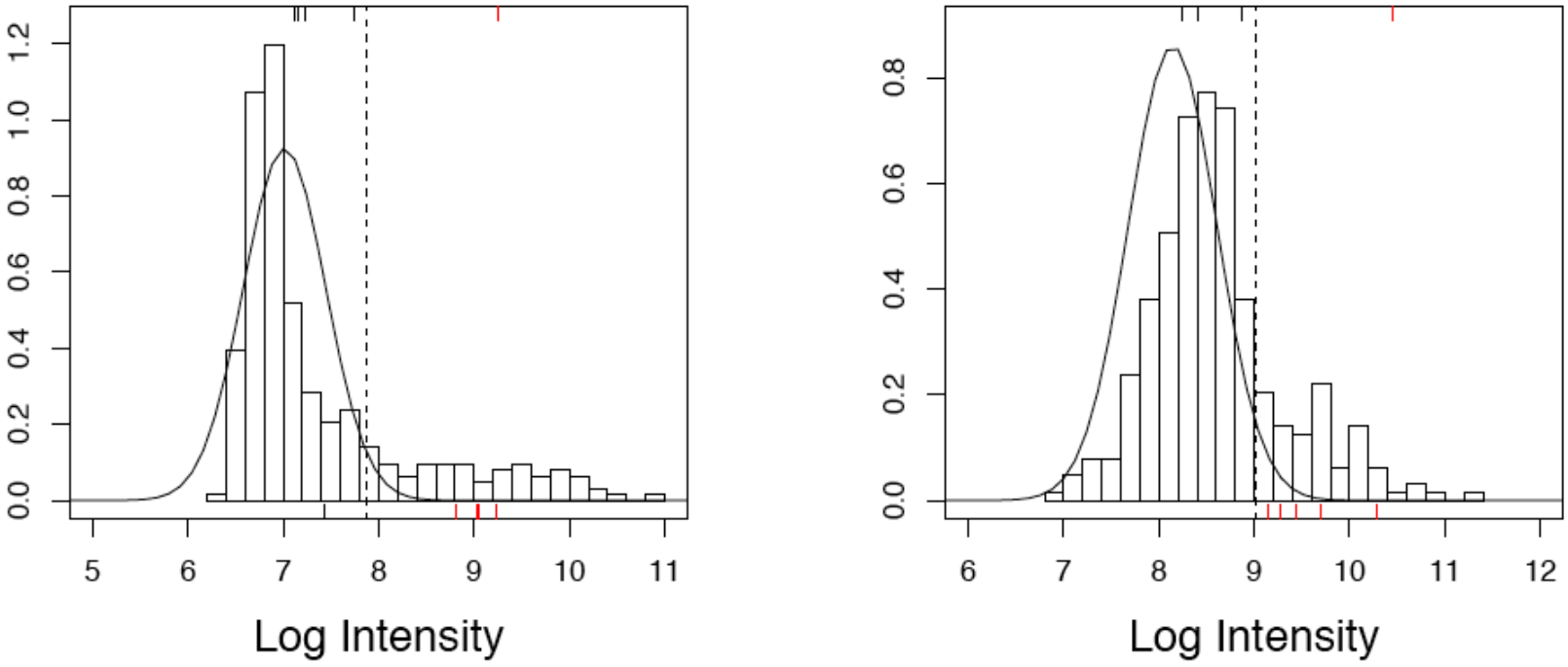
Keeping More Genes

The new model:

$$\begin{aligned}(y_{ijg}|\theta_{jg}) &\sim N(\theta_{jg}, \sigma_g^2) \\ \theta_{jg}|\mu_g &\sim (1-p_g) * N(\mu_g, \tau_g^2) + p_g * U(\mu_g, S) \\ \mu_g &\sim N(\epsilon, \lambda^2)\end{aligned}$$

For gene g in sample i from tissue j. Values are assumed to come from a mixture of the unexpressed and expressed intensity distributions.

Revisiting Cancer Prediction



McCall et al (WP 200), Fig 6. Clear split between expressed and not, but not by 6 SD.

What We Want to Know

Can we relate miR expression to mRNA regulation?

Is this regulation altered in cancer?

Can TCGA data let us address these questions?

The Data At Hand (miRNA)

8 data folders (4 per tissue) acquired as of May 7 2010:

Level 1 Data (Agilent output)
Level 2 Data (Probe-level data, 2421/1510 obs)
Level 3 Data (miR-level data, 799/534 obs)
MageTab (SDRF) File

The first three all contain files for 529 (ovary) or 386 (GBM) samples at various stages of processing.

The last contains mappings of names used at various levels.

The Data At Hand (Affy mRNA)

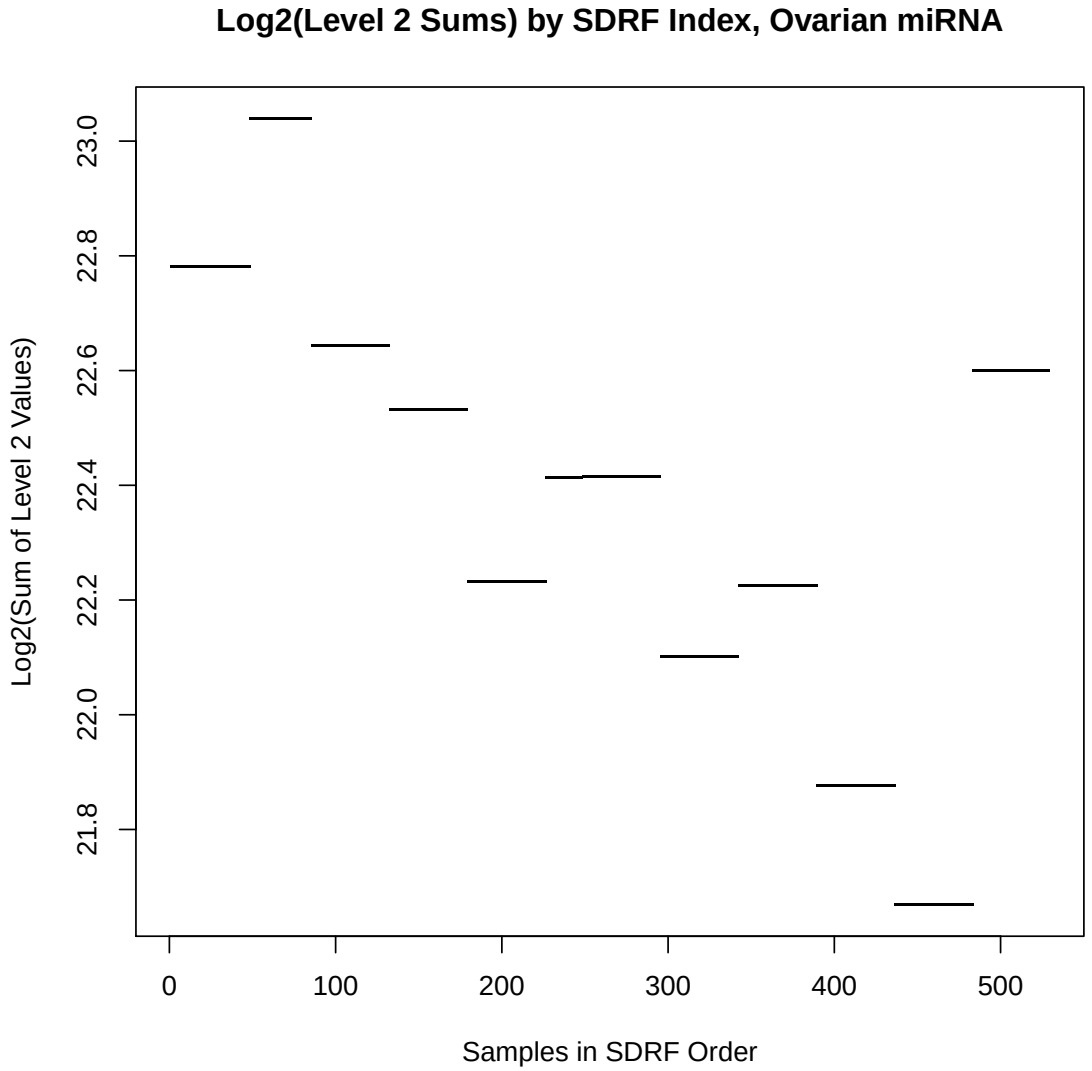
71 data folders (3 each for 23 batches, and 1 mapping per tumor type) acquired as of Jul 3 2010:

Level 1 Data (HT_HGU133A CEL files)
Level 2 Data (Probe-level data, 22277 obs)
Level 3 Data (Gene-level data, 12042 obs)
MageTab (SDRF) File

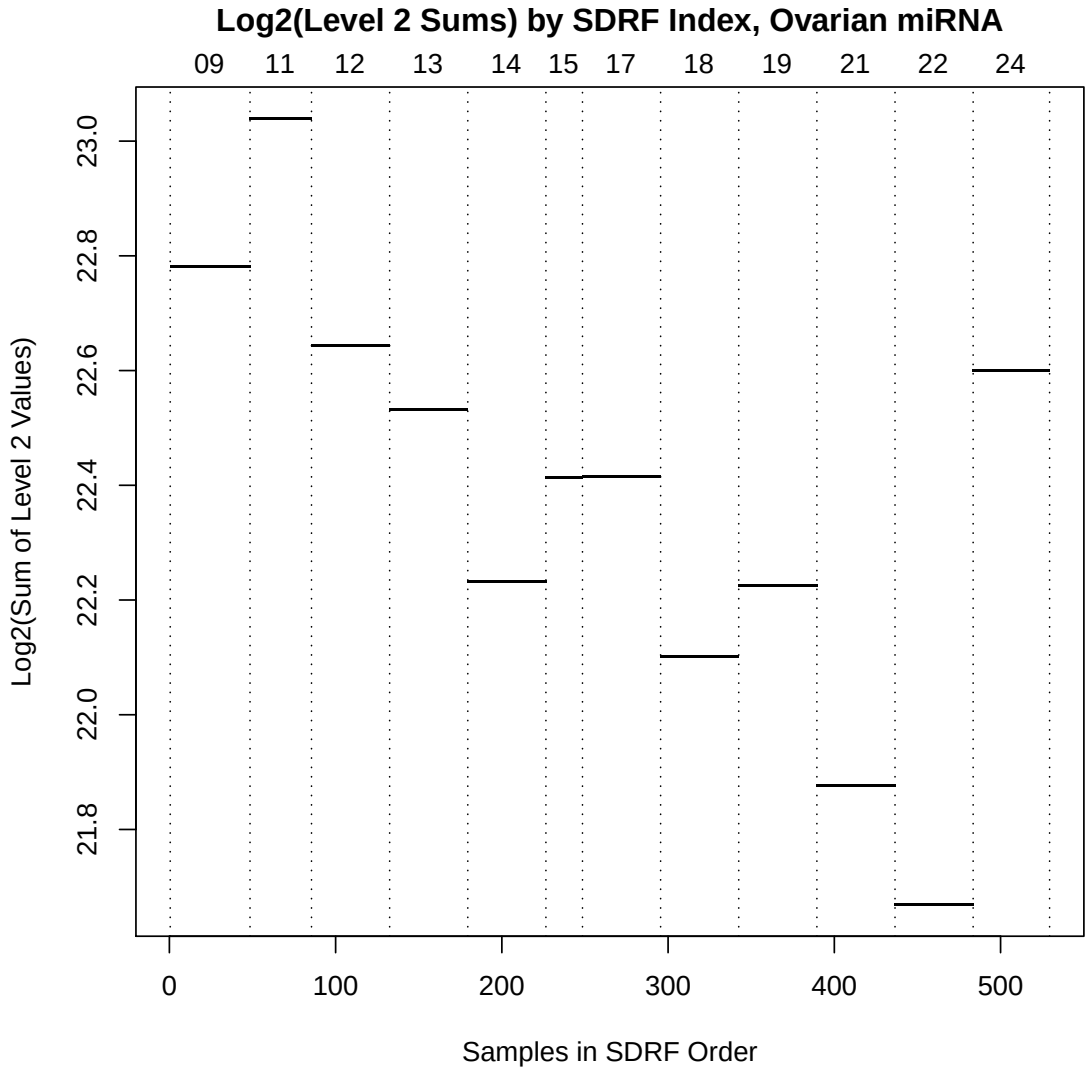
Arrays were reportedly processed with RMA.

Can we simply look for high correlations between Level 3 measurements across assay types?

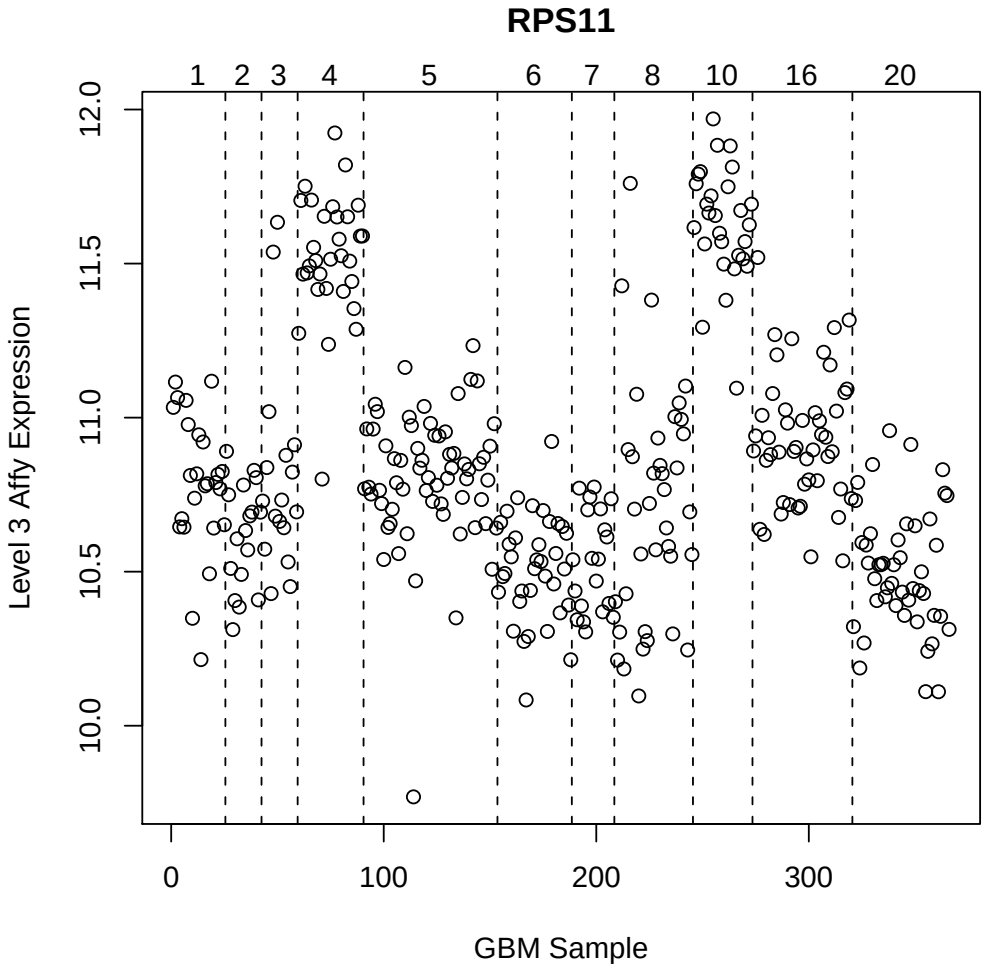
Level 2 Sums: Do They Look Constant?



Level 2 Sums: A Pattern?

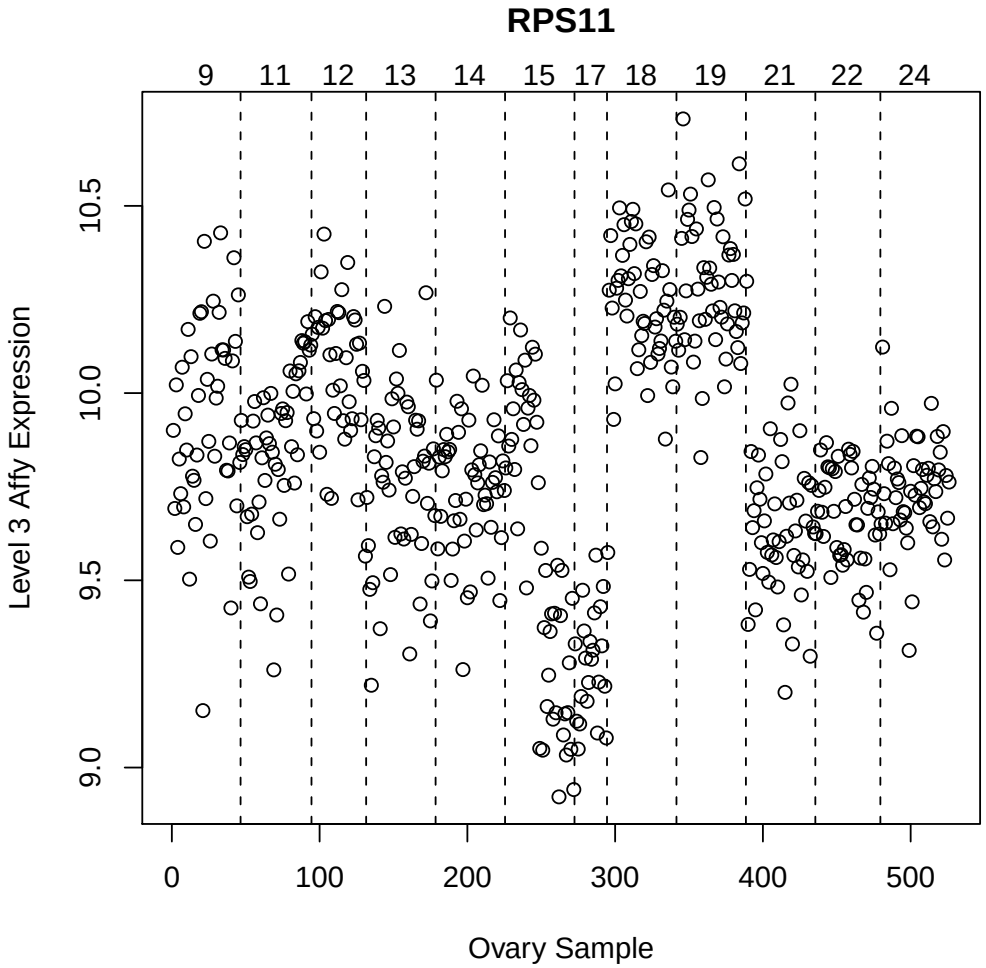


What About GBM Level 3 mRNA Data?



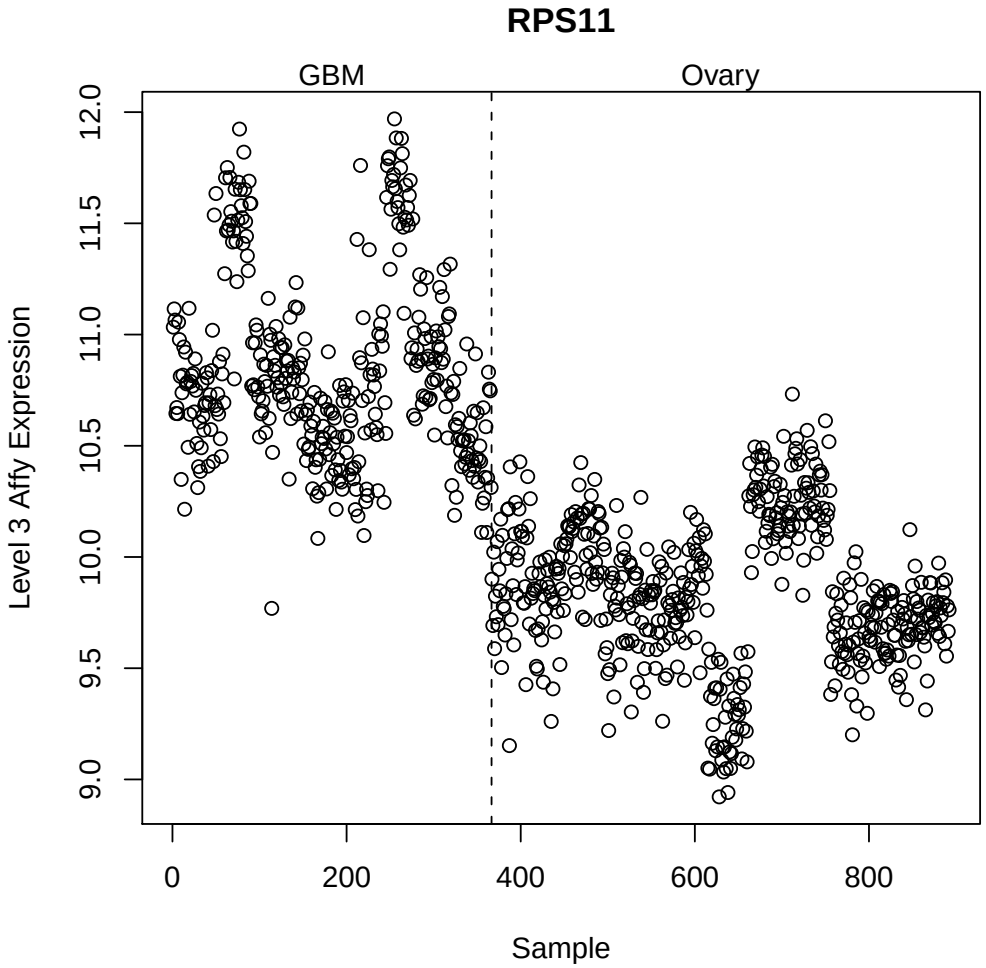
Some shifts here

What About Ovary Level 3 mRNA Data?



and here

What About Combined Level 3 mRNA Data?



and here

Summary and Options

Batch effects are clearly visible in the L3 data for both types of assays.

It's not clear the data is always on comparable scales.

Suggestion:

For each miR (mRNA), take the L3 data and rank and center the values, *batch by batch*.

Then check all the resultant miR/mRNA Pearson correlations.

Of course, before we do that, we've got to combine the data...

Mapping the Data

Using the SDRF files, we attempted to replace all L2 and L3 “array names” with their TCGA sample ids for consistent mapping.

Problem 1: some GBM samples were run multiple times (almost all in batches 3 and 4). In some cases Affy replicates occur in more than one batch.

Solution 1: keep the first rep of each sample, such that the batch information agrees with that posted on the TCGA data matrix.

Mapping the Data (2)

Problem 2: The ordering of samples is different for the different assays, and different batch boundaries are used.

Solution 2: We use the ordering from the MiRNA SDRF files, which more closely tracks with the ordering on the TCGA data matrix. Batch boundaries were determined from the matrix.

Problem 3: Some samples were run on only one assay.

Solution 3: Drop these. Restricting attention to the common samples reduces the GBM miR data from 371 samples to 366, and Ovary from 529 samples to 526.

Mapping the Data (3)

Problem 4: The ordering of probesets and gene names is not necessarily the same across tissues.

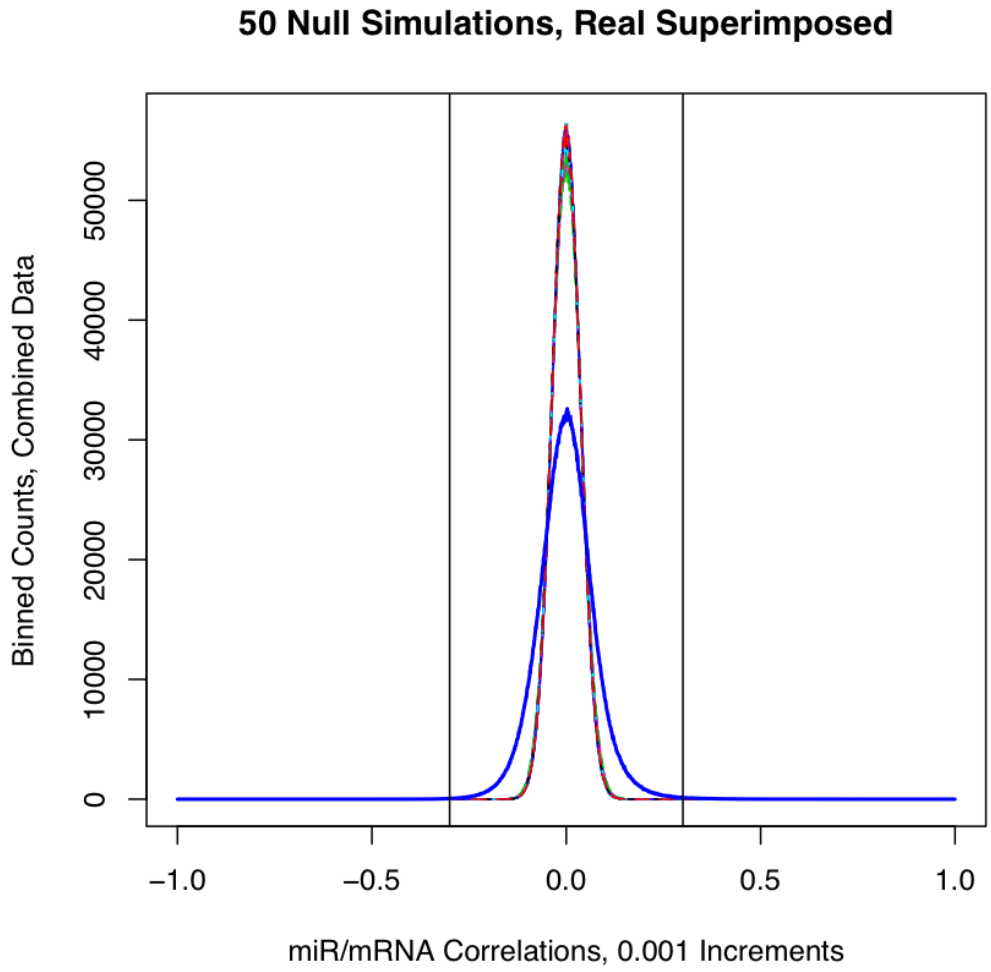
Solution 4: Explicitly enforce this matching in R.

Problem 5: Different versions of the Agilent miRNA array were used for GBM and Ovary.

Solution 5: Keep only the common probes (Level 2) and common miRs (Level 3).

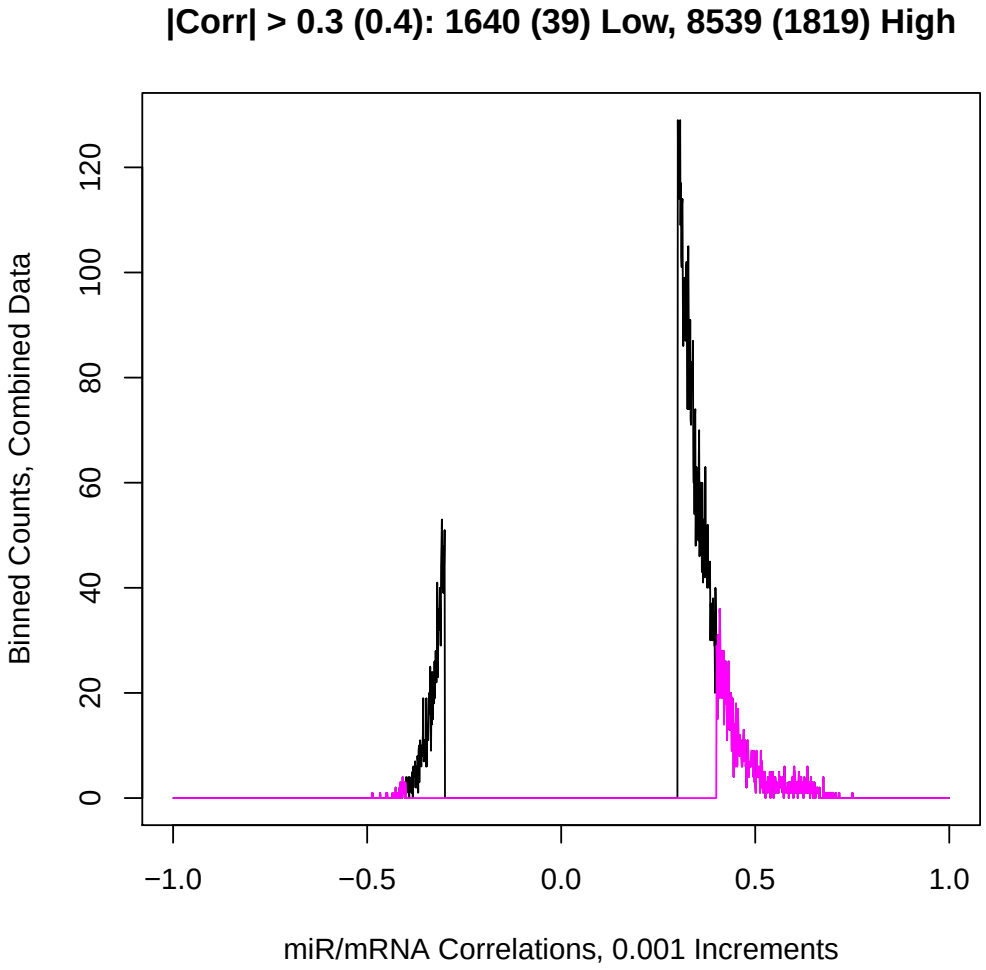
After all this, are the Level 3 correlations big? We checked
(1) the combined data,
(2) just the GBM data, and
(3) just the Ovary data.

Level 3 Corrs, Combined Data



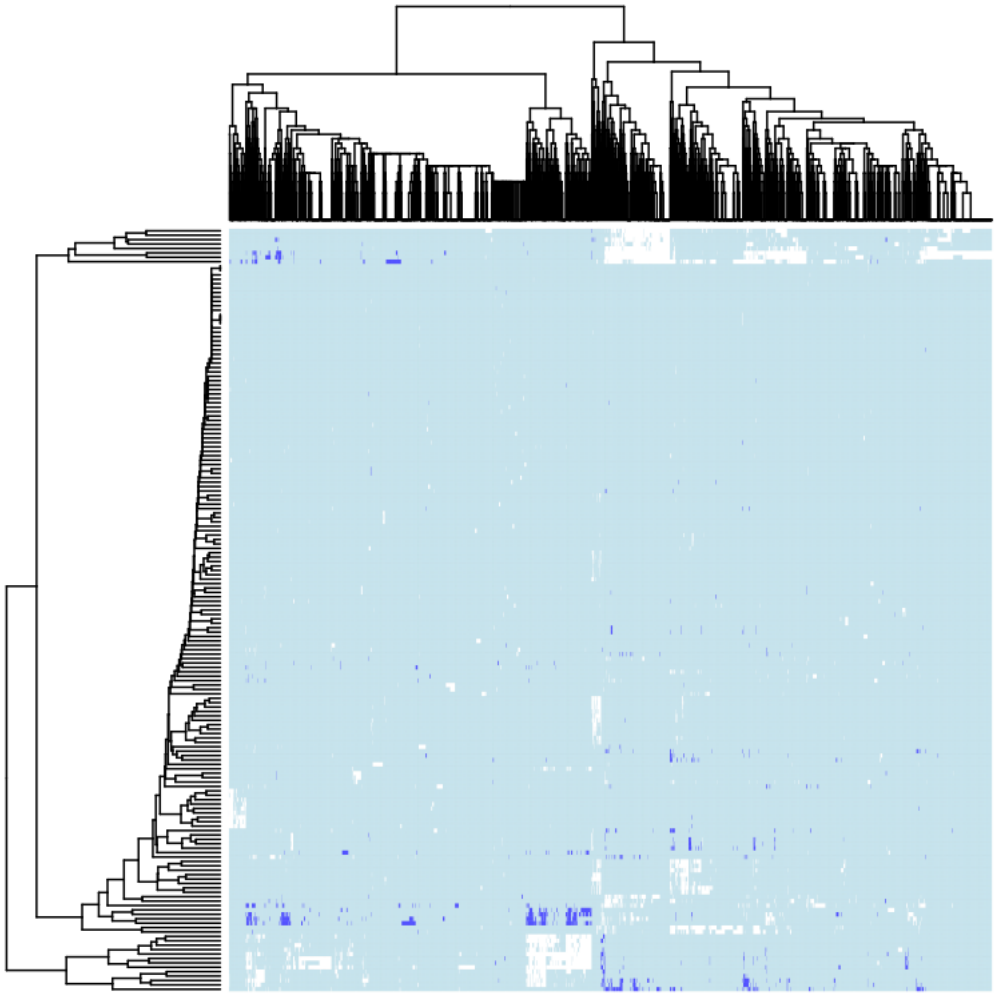
Base takes 10 sec to compute. Set cutoffs at 0.3.

Level 3 High Corrs, Combined Data



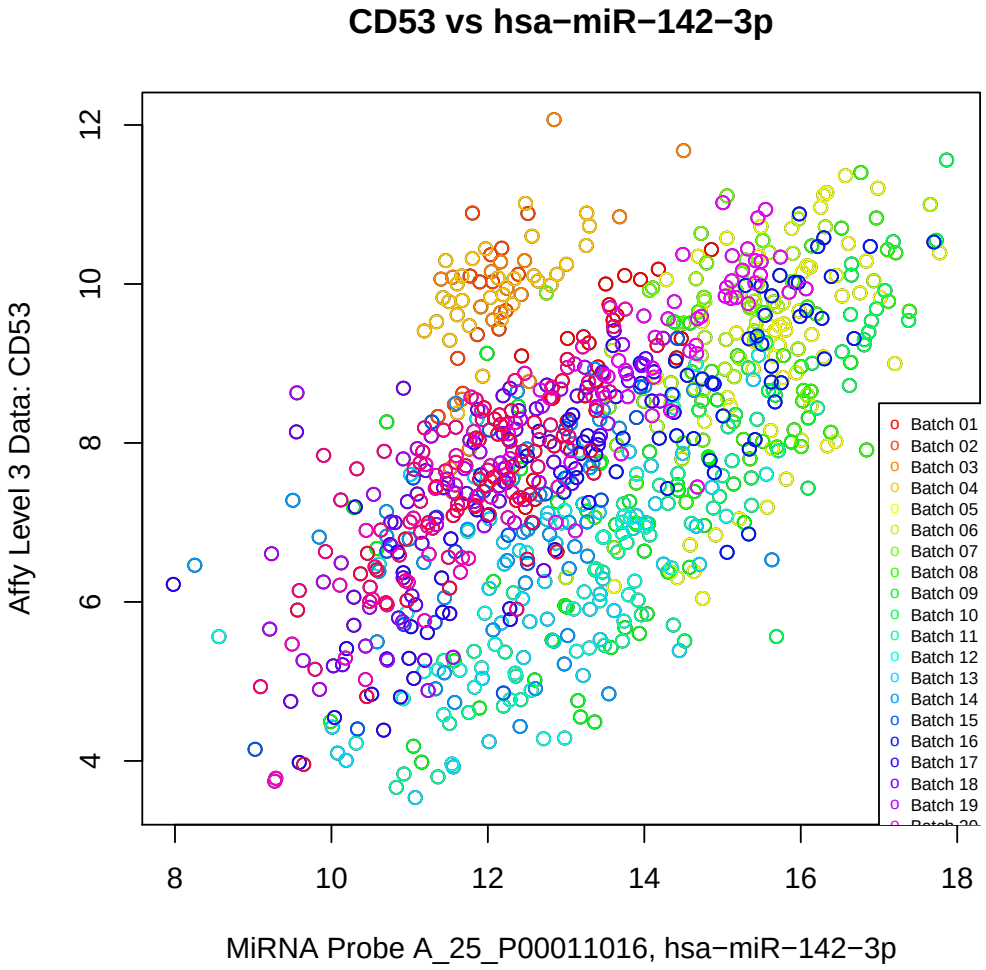
Note: more high corrs than low!

Not All MiRs are Equal!



|Corl > 0.3; Blue < –0.3, White > 0.3. 2348 Genes, 173 miRs

Do Batches Really Matter?



again, using L2 miR data (log2!)