GS01 0163
Analysis of Microarray Data

Keith Baggerly and Kevin Coombes
Department of Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center

kabagg@mdanderson.org

kcoombes@mdanderson.org

September 11, 2007

Lecture 4: Cell Lines, Drugs, and Docs

Microarrays and Cancer:

What has the focus been?

What should the focus be?

Can we figure out what's going wrong in cancer?

Finding patterns of aberrant gene expression

Can we figure out who to treat?

Disease identification and Disease subtyping

Making Research "Translational"

Can we figure out how to treat them?

Long term: *How should I plan to treat patients 5 years from now?* Develop drugs targeting specific abnormalities.

Short term: *How should I treat the patient in my office today?* Figure out which types of available treatments (chemotherapeutic regimens) are likely to be effective.

We have some examples in the long term category (Gleevec, Herceptin), but we'd like to have more examples in the short term category. So, how can we get there?

What tools do we have?

Cancer, Chemo, and Cell Lines

1955 – Cancer Chemotherapy National Service Center (CCNSC) established. One goal: test drugs as anticancer agents. Candidate drugs, assigned an NSC number, were tested for efficacy in leukemic mice.

1976-82 – CCNSC incorporated into Developmental Therapeutics Program (DTP); Human tumors in mice.

1985-90 – Human tumor cell line panel (NCI60) established as first line test.

Today – Tens of thousands of cytotoxic agents have been evaluated for activity against the standard panel.

http://dtp.nci.nih.gov/timeline/noflash/index.htm

A Source of Excitement

Genomic signatures to guide the use of chemotherapeutics

Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1–3], Johnathan Lancaster[4] & Joseph R Nevins[1–3]

Potti et al (2006), Nature Medicine, 12:1294-1300.

The main conclusion of this paper is that we can use microarray data to define "signatures" that will suggest whether a patient will respond to a given agent, and that these signatures can be defined using data on cell lines (the NCI60). They provide examples using 7 commonly used agents.

The Other Exciting Bit...

All the analyses were performed on publicly available data sets.

The drug response data is publicly available, and maintained by the NCI.

The microarray profiles of the cell lines are publicly available, and at least some are available from the NCI.

Microarray profiles of patient and cell line samples that did and did not respond to various drugs are available from public repositories (esp GEO).

We should be able to do it ourselves!

Analysis Plan

1. Identify and collect the data sets.

2. Using the sensitivity information for a drug of interest (docetaxel) select cell lines from the extremes.

3. Using the array profiles of the chosen cell lines, select features (genes) that best distinguish sensitive from resistant.

4. Use the array values for the chosen features to train a binary model to distinguish sensitive from resistant cell lines.

5. Test the model for its ability to make accurate predictions on expression data sets from patient tumors.

Data Sources

1. Drug response: assays on NCI60 cell lines from DTP at
   NCI (`http://dtp.nci.nih.gov/docs/cancer/`
   `cancer_data.html`)

2. Training: Affymetrix U95Av2 arrays on NCI60 cell lines,
   experiments performed in triplicate at Novartis (`http:`
   `//dtp.nci.nih.gov/mtargets/download.html`)

3. Testing: 24 breast tumors on U95Av2; experiments at
   Baylor reported in Chang et al (2003) Lancet, 362:362-9.
   Available at GEO as GSE349, GSE350, and GDS360.
   (Sample 377, GSM4913, is mislabled at GEO. It should be
   "sensitive". Personal communication from the Lancet
   authors.)

Identifying the Drugs

The paper gives the *names* of the drugs profiled, but response data is indexed by NSC number. How would you find these numbers?

Identifying the Drugs

The paper gives the *names* of the drugs profiled, but response data is indexed by NSC number. How would you find these numbers? Google!

NSC Number	Drug
628503	Docetaxel (Taxotere)
123127	Adriamycin (Doxorubicin)
26271	Cytoxan (Cyclophosphamide)
141540	Etoposide
125973	Paclitaxel (Taxol)
19893	5-Fluorouracil
609699	Topotecan

So, What Does the NCI/DTP Report?

For every drug/cell line combination, the DTP estimates 3 measures of drug sensitivity: GI50, TGI, and LC50.

Summary tables are available from `http://dtp.nci.nih.gov/docs/cancer/cancer_data.html`
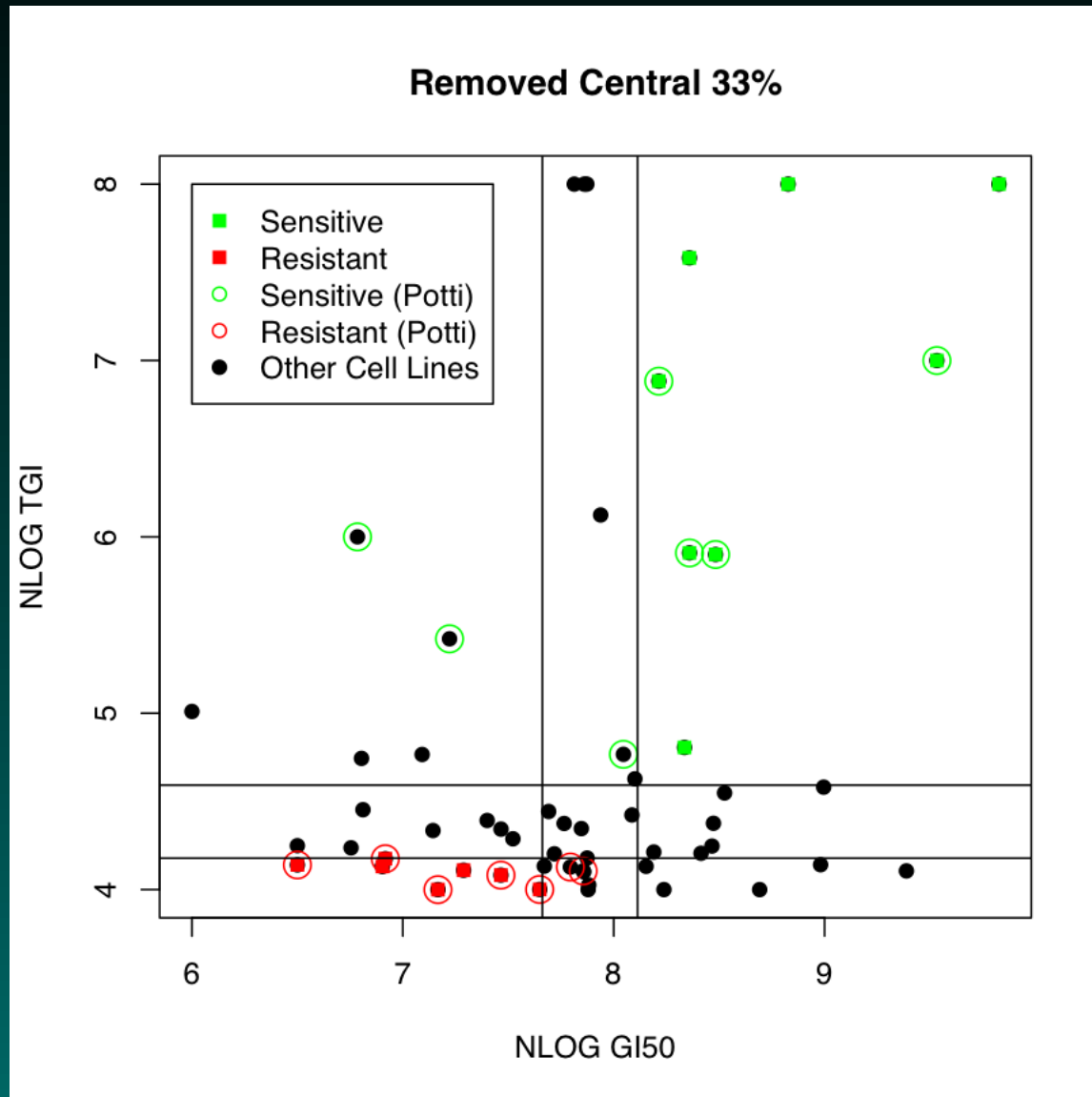
Raw data is available from the bulk download site, `http://dtp.nci.nih.gov/dtpstandard/subsets/dose.jsp`

So, what are these numbers?

What the Data Looks Like

Docetaxel Assays for COLO 205
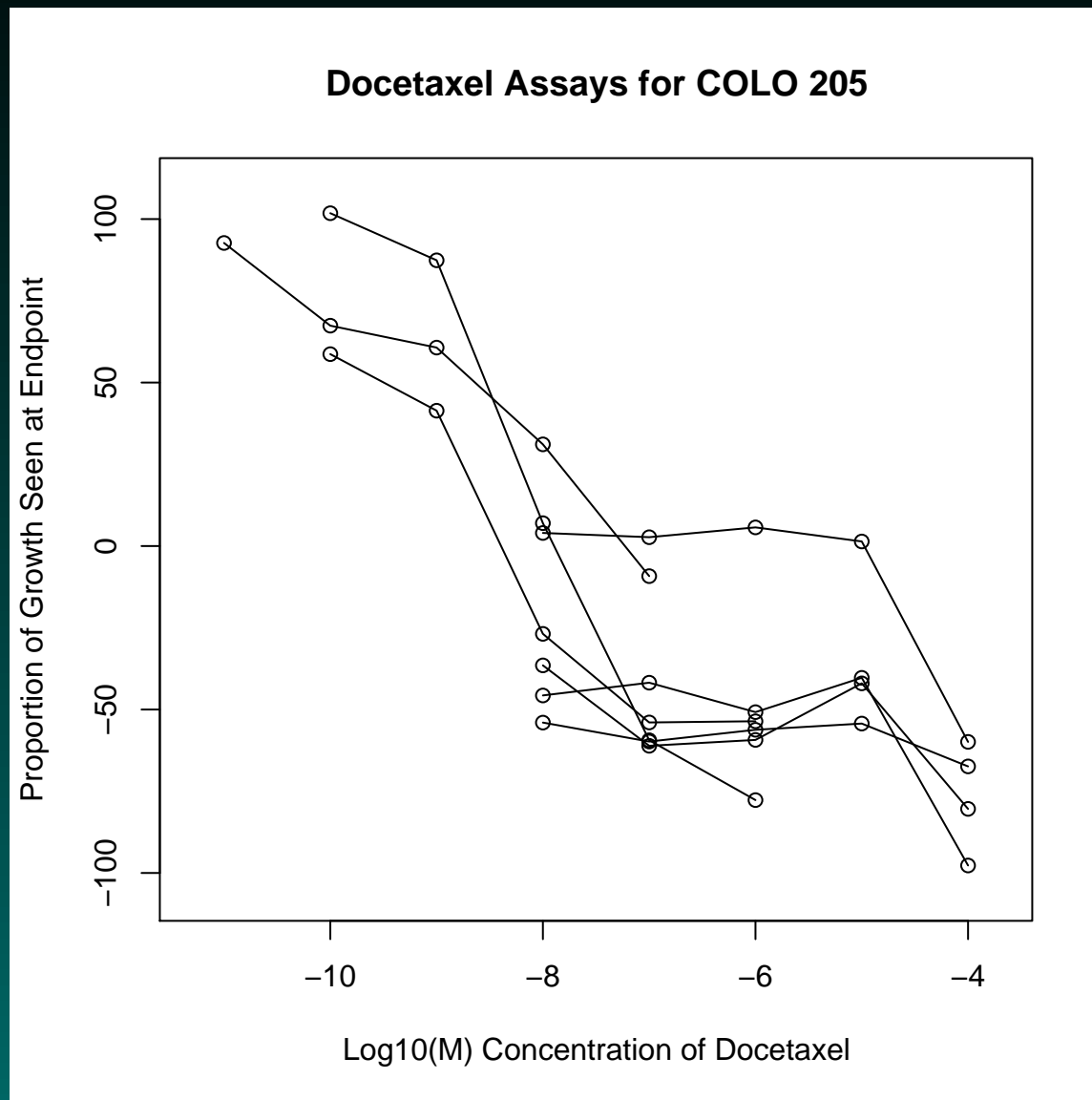
Adding Cutoffs
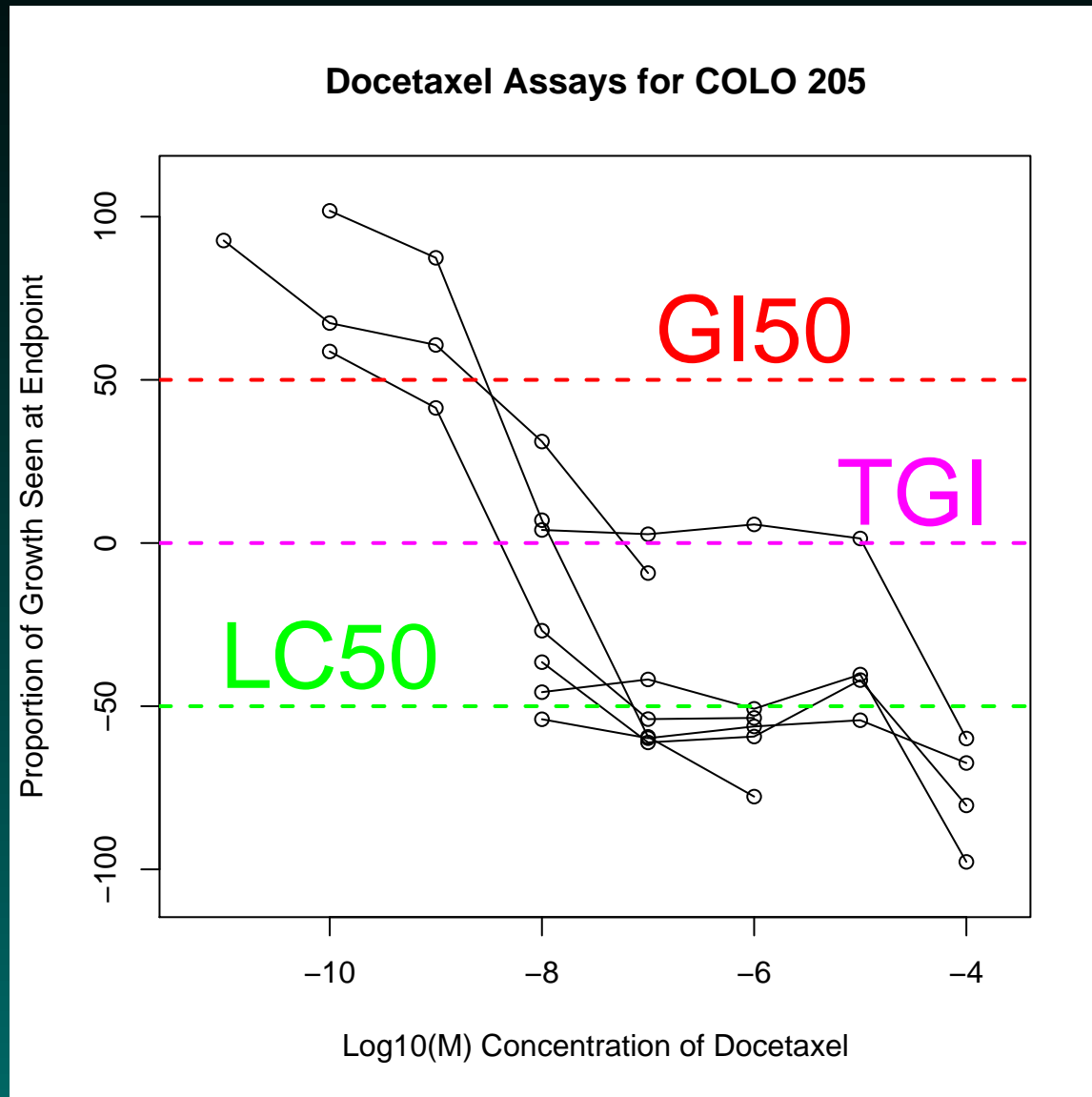
Selecting Cell Lines for Docetaxel

The Cell Lines Selected

- **Sensitive** (8 lines)

  - COLO 205, HCC-2998, HL-60(TB), HT29, MDA-MB-435, NCI-H522, RPMI-8226, SF-539

- **Resistant** (7 lines)

  - 786-0, ACHN, CAKI-1, EKVX, IGROV1, OVCAR-4, SF-268

Note: the paper does not say which cell lines were used.

On to the training data!

Processing the Microarray Data

The data posted at the DTP web site that we'll use comes from experiments on U95Av2 arrays performed by Novartis, and apparently quantified using MAS5.0. The experiments were performed in triplicate (referred to later as Series "A", "B", and "C"); 180 arrays in all.

We performed quantile normalization using the `normalize.quantiles` function in version 1.10.0 of the `affy` package from BioConductor. We also log-transformed all of the data (base 2). All of our computations were initially performed using version 2.3.1 of the `R` statistical programming environment.

Feature Selection

Following the paper, we selected the top 50 genes based on a two-sample t-test between sensitive and resistant cell lines. The set of selected features varies based on which set of replicate experiments was used. The overlaps:

	Avg	Joint	A	B	C
Avg	50	21	12	17	10
Joint	21	50	12	15	21
A	12	12	50	7	4
B	17	15	7	50	7
C	10	21	4	7	50

We'll use the average for now.

Training a Model: Metagenes are PCs

The paper uses "metagenes" to construct a predictive model. Metagenes summarize the information present in a chosen set of genes by taking weighted averages. *Mathematically, they're simply the principal components of the chosen matrix.*

For the model, they use probit regression with the metagene scores as inputs, choosing coefficients to separate the sensitive and resistant groups.
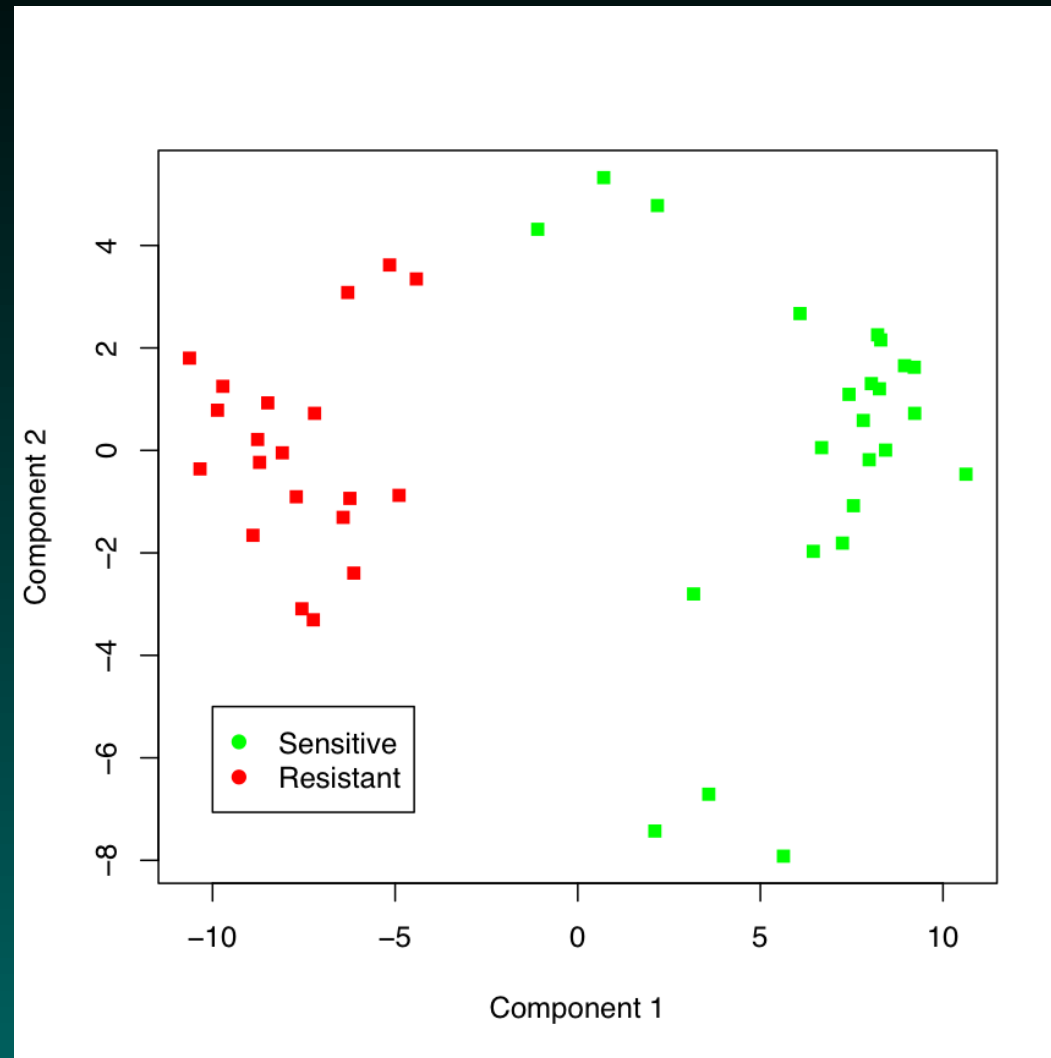
Given how we selected the genes, which PC would we expect to drive the prediction?

Only the First Component Contains Information

See!
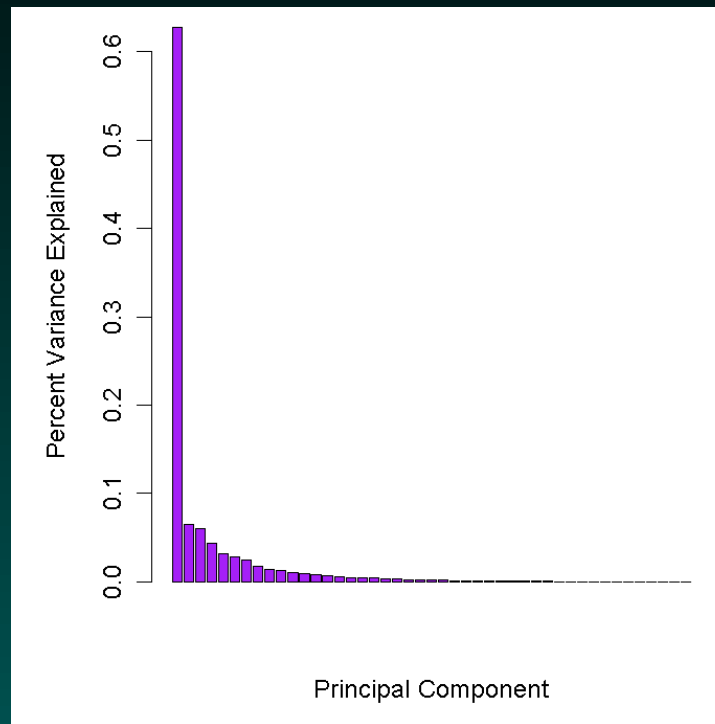
We performed PCA on the set of 45 arrays (15 cell lines in triplicate) that we chose, using the 50 selected genes. We used probit regression to predict sensitivity using the PCs.

Principal Components: NCI60 Training Data

We want test data to split like this. What data do we have?

The Test Data

Chang et al (2003), in a paper in *the Lancet*, tried to measure the response of breast tumors in vivo to docetaxel chemotherapy. Tumor size was assessed both before initiation of chemo and after a fixed number of treatments. Those that shrank the most were deemed "sensitive".

Affymetrix U95Av2 profiles were derived using biopsy samples taken before the initiation of chemo.

Following Chang et al, we quantified the data using the software program *dChip* (both PM-only and PM-MM models were tried). Then we mapped the values found to the quantiles defined by the training data.

Did it work?

Principal Components: Chang Test Data
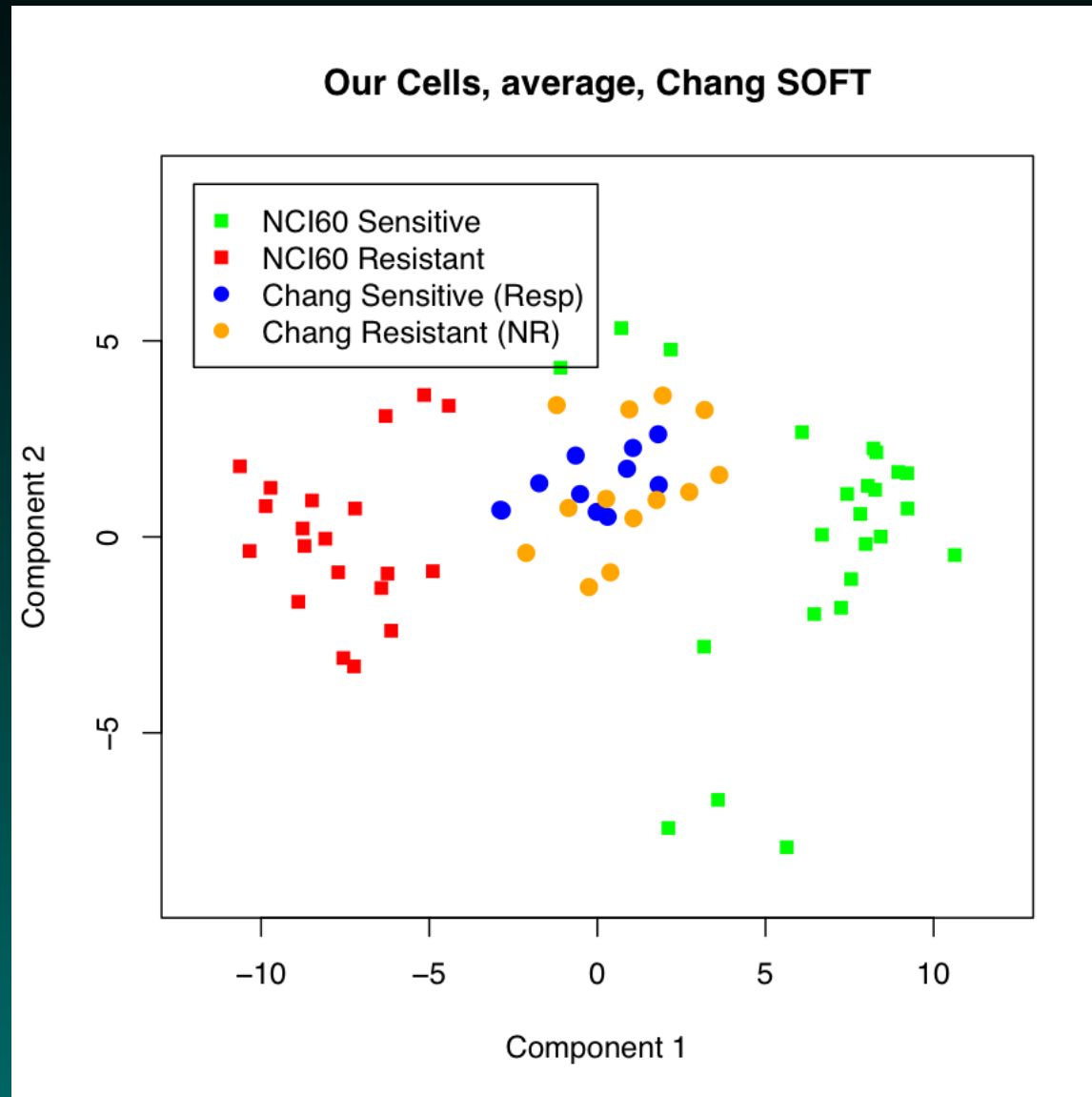
Prediction Accuracy on Chang Test Data

Only PC1 was significant in univariate analyses.

PC1		Truth	
		Resistant	Sensitive
Predicted	Resistant	0	2
	Sensitive	13	9

The best multivariate model (stepwise addition, using AIC):

AIC		Truth	
		Resistant	Sensitive
Predicted	Resistant	0	2
	Sensitive	13	9

Prediction Accuracy on Chang Test Data

Only PC1 was significant in univariate analyses.

PC1		Truth	
		Resistant	Sensitive
Predicted	Resistant	0	2
	Sensitive	13	9

The best multivariate model (stepwise addition, using AIC):

AIC		Truth	
		Resistant	Sensitive
Predicted	Resistant	0	2
	Sensitive	13	9

Did we do something wrong?

Reconstructive Bioinformatics

In the outline above, we tried to follow their qualitative approach. Now, we're going to try to figure out *exactly* how this worked.

- What cell lines were used?

- What training data was used? How processed?

- What features were selected?

- What were the metagenes?

- What were the predictions?

What cell lines were used?

We asked about this, to be sure we were working with the right data.

They responded, but not with precisely what we asked for.

They sent us a giant Excel table.

The First 2 Rows...

```
probe_set Adria0 0 0 0 0 0 0 0 0 0 0 1 1 1
 1 1 1 1 1 1 1 1 Adria1 Doce0 0 0 0 0 0
 0 0 0 0 1 1 1 1 1 1 1 1 1 Doce1 Etopo0
 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 Etopo1
 5-FU0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 5-FU1
 Cytox0  0 0 0 0 0 0 0 0 0 1 1 1 1 1 1
 1 1 1 Cytox1  Topo0 0 0 0 0 0 0 0 0 0
 0 0 0 1 1 1 1 1 1 1 1 1 Topo1  Taxol0
 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1  Taxol1
36460_at          41.671947          21.820335
 125.794838          93.459251          79.06321
```

Does this answer the question?

The Novartis Data

The first 3 rows of the Novartis "individual" file at the NCI:

```
Probe Set Name,ID,Gene,cellname,pname,
panelnbr,cellnbr,Signal,Detection,P Value
36460_at,GC26855_B,POLR1C,K-562,Leukemia,
   7,5,120.478737,P,0.017001
36460_at,GC26855_B,POLR1C,MOLT-4,Leukemia,
   7,6,113.542229,P,0.001892
```

The first probeset id is the same. The numbers look similar.

What happens if we try to match the Novartis data just using the values for the first probeset?

The Lines We See...

We find matches!

Tabulating the arrays in each group that were used:

Batch	Not Used	Used
A	6	53
B	59	0
C	58	0
D	4	0

They used the A set! (The values match for all the other probesets too.) Further, given the matches, we know which cell lines were used. This works for 6 drugs (not cytoxan).

Side note: yes, it is the NCI59, not NCI60...

How Were Cell Lines Chosen?

We want to understand this, so that we can do it.

Supplementary Methods: *"[W]e chose cell lines ... that would represent the extremes of sensitivity to a given chemotherapeutic agent (mean GI50 $\pm$ 1 SD). ... [T]he log transformed TGI and LC50 dose ... was then correlated with the respective GI50 data. ... Cell lines with low GI50 ... also needed to have a low LC50 and TGI. ..."*

Ok, start with GI50, pick cell lines more than 1sd away from the mean, and make sure that the TGI and LC50 scores are at least on the right side of their median values.

How Were Cell Lines Chosen?

We want to understand this, so that we can do it.

Supplementary Methods: *"[W]e chose cell lines . . . that would represent the extremes of sensitivity to a given chemotherapeutic agent (mean GI50 ± 1 SD). . . . [T]he log transformed TGI and LC50 dose . . . was then correlated with the respective GI50 data. . . . Cell lines with low GI50 . . . also needed to have a low LC50 and TGI. . . ."*

Ok, start with GI50, pick cell lines more than 1sd away from the mean, and make sure that the TGI and LC50 scores are at least on the right side of their median values.

If we do this, we get very few cell lines resistant or sensitive to docetaxel.

What Are GI50s for the Docetaxel Lines?

This is actually not as simple a question as you might think.

Docetaxel has been assayed many times, and some of the assays didn't cover the same concentration ranges. Sometimes the maximum concentration is 4 (-log10 scale), sometimes 6, and sometimes 7. So we looked at each.

Docetaxel, Max Conc 4
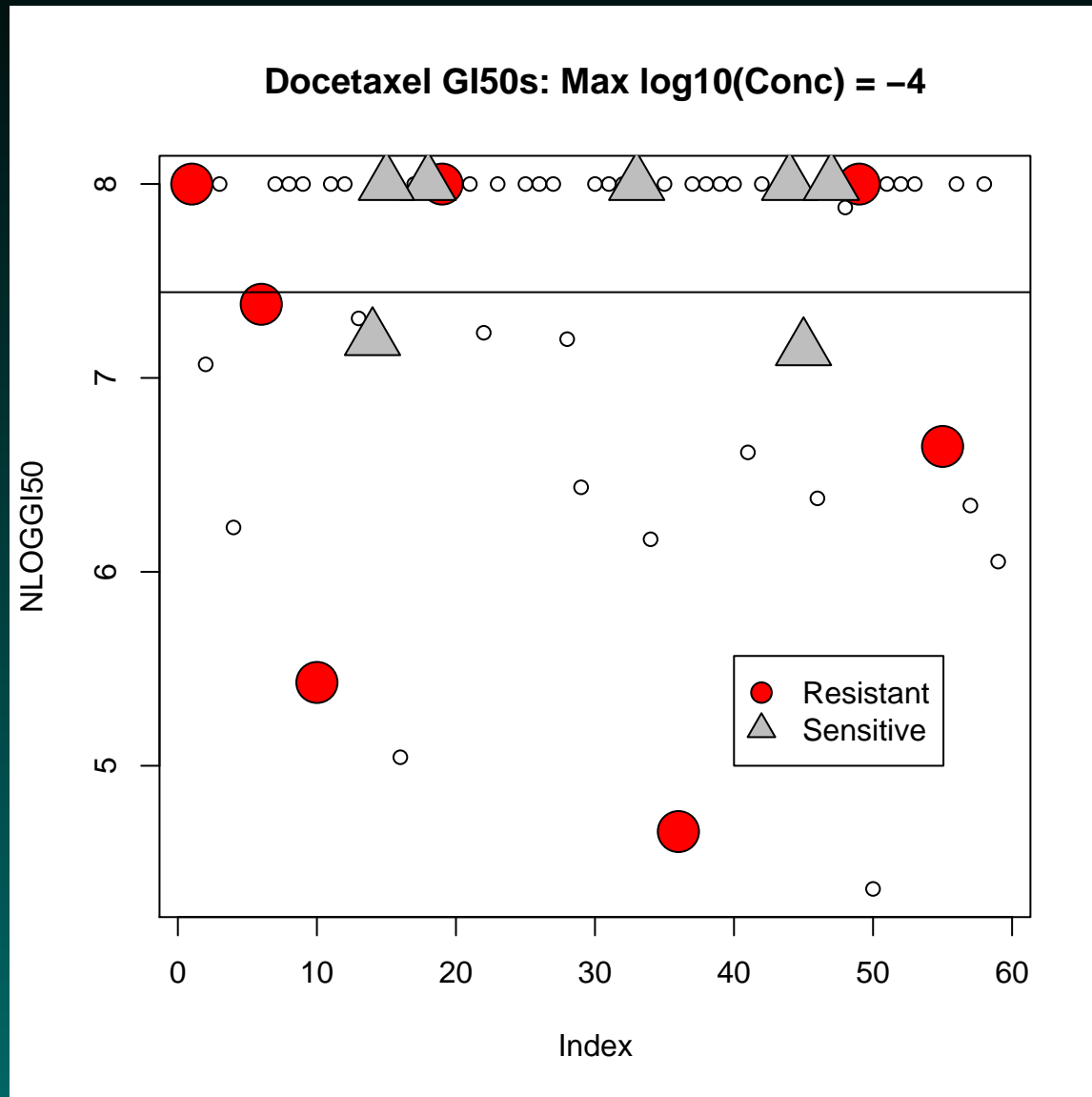
Docetaxel, Max Conc 6
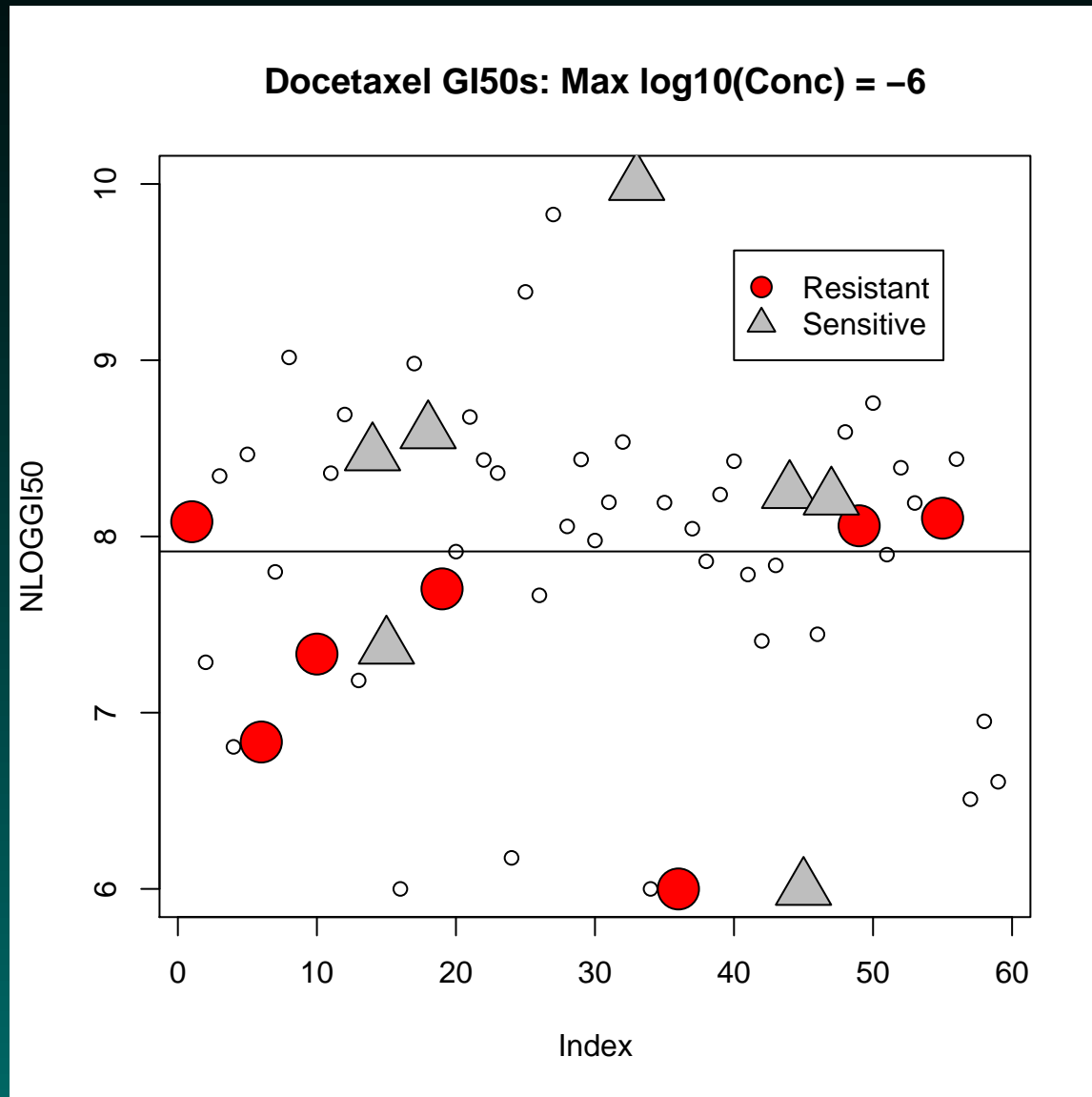
Docetaxel, Max Conc 7
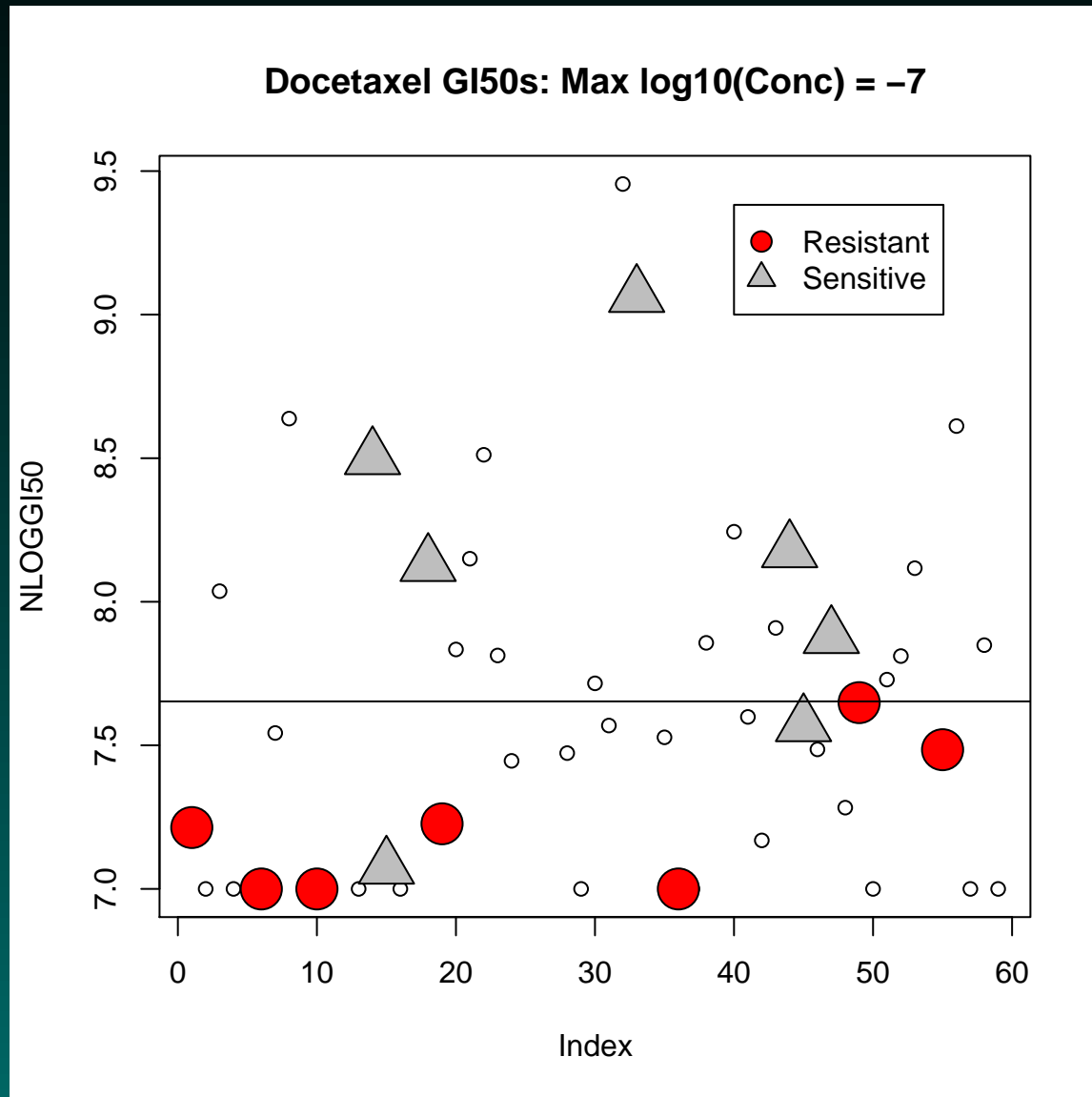
Observations

There's something odd about these plots.

The GI50 scores for the resistant and sensitive sets overlap. There shouldn't be any overlap.

Observations

There's something odd about these plots.

The GI50 scores for the resistant and sensitive sets overlap. There shouldn't be any overlap.

This is true for *every* drug.

We've asked them about this. They gave us a more precise rule, but it doesn't give these lines either. (Their chosen lines for docetaxel do separate by TGI.)

Given their cell lines, can we find the same features?

They Reported the Features!

We actually have the lists of probesets used in the predictors. These were supplied in supplementary table 1, and as separate files at the website they named in the supplementary methods document: `http://data.cgt.duke.edu/NatureMedicine.php`*. Interpretations are given in the paper for why many of these make sense.

How were these found? According to the supplementary methods: *"a variance fixed t-test was used to calculate significance"*.

Ok, let's try this. (Log transform and quantile normalize the data first!) We'll use 5-FU to give docetaxel a break. Let's build a heatmap.

5-FU, Their Cell Lines and Genes
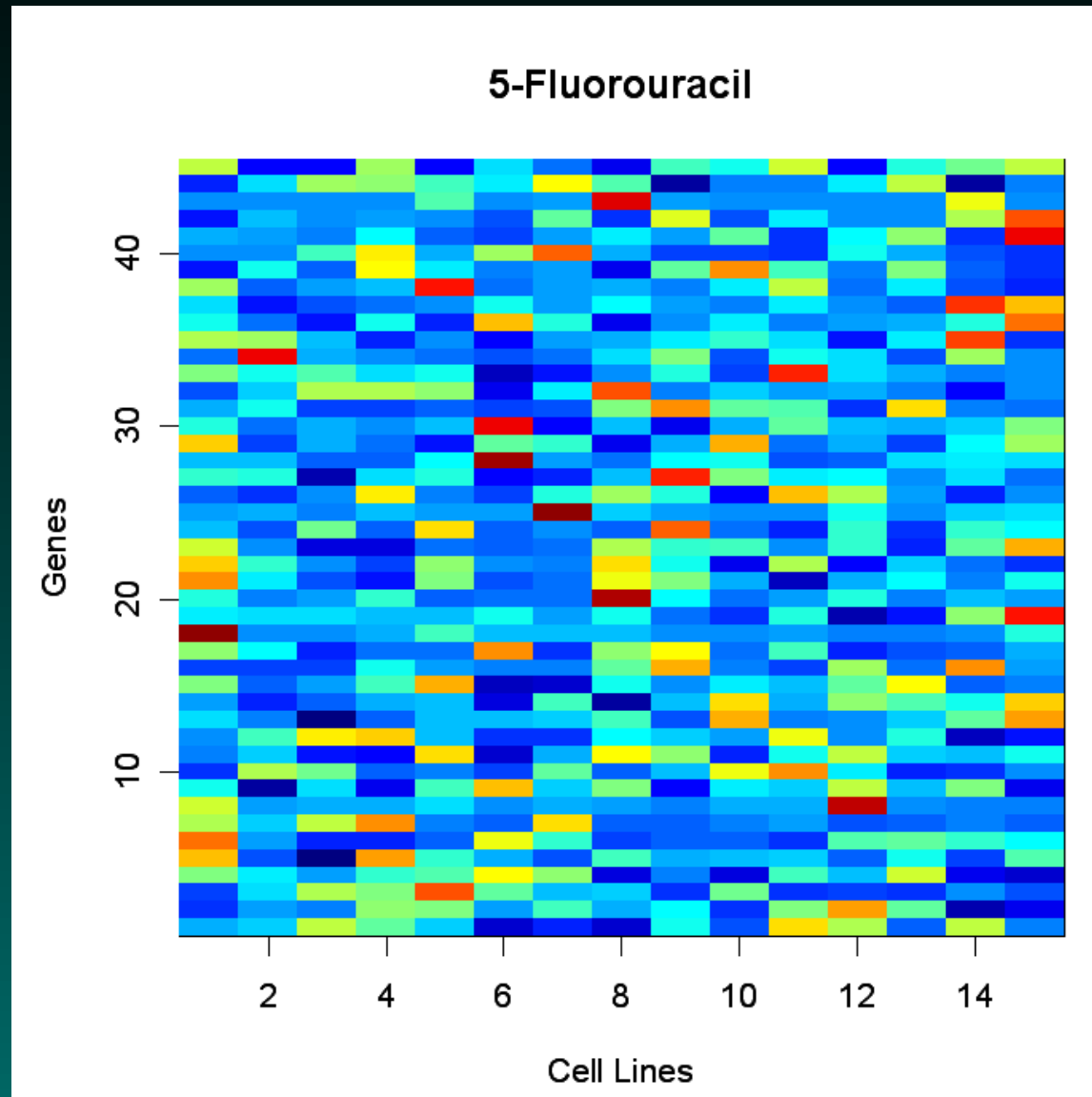
5-FU, Genes From T-test
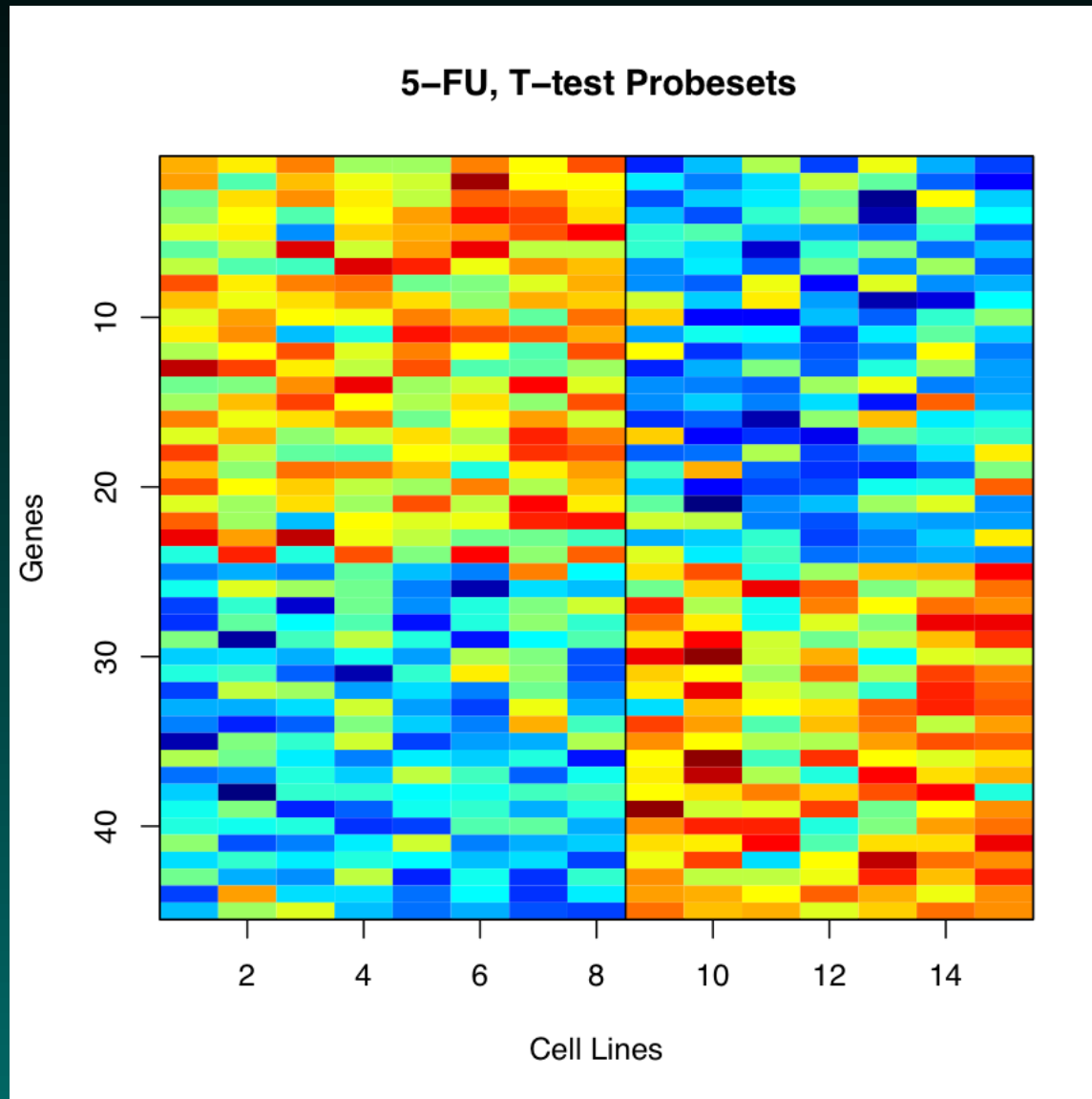
The Figure in the Paper
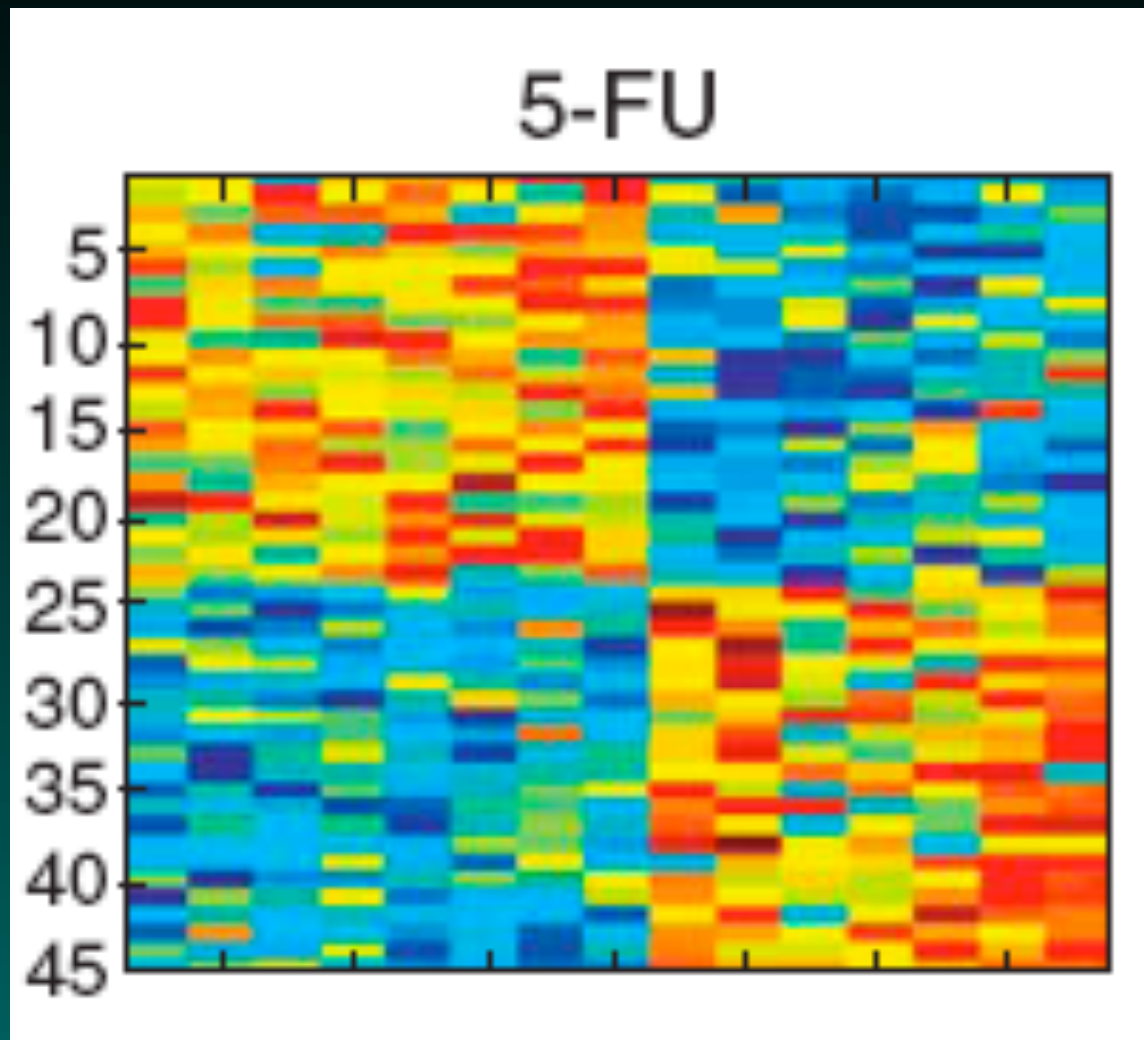
How are the Lists Different?

```
> temp <- cbind(
    sort(rownames(pottiUpdated)[fuRows]),
    sort(rownames(pottiUpdated)[
        fuTQNorm@p.values <= fuCut]);
> colnames(temp) <- c("Theirs", "Ours");
> temp
      Theirs          Ours
[1,] "1519_at"       "151_s_at"
[2,] "1711_at"       "1713_s_at"
[3,] "1881_at"       "1882_g_at"
[4,] "31321_at"      "31322_at"
[5,] "31725_s_at"    "31726_at"
[6,] "32307_r_at"    "32308_r_at"
...
```

Uh Oh...

For almost every item on their list, the very next item in lexicographical order is on ours.

This suggests that there was an off-by-one indexing error in mapping between numbers and probeset ids. (This has some negative implications for the interpretations.)

If we just take their list "plus one", what do the p-values look like?

P-values for 5FU, and 5FU+1
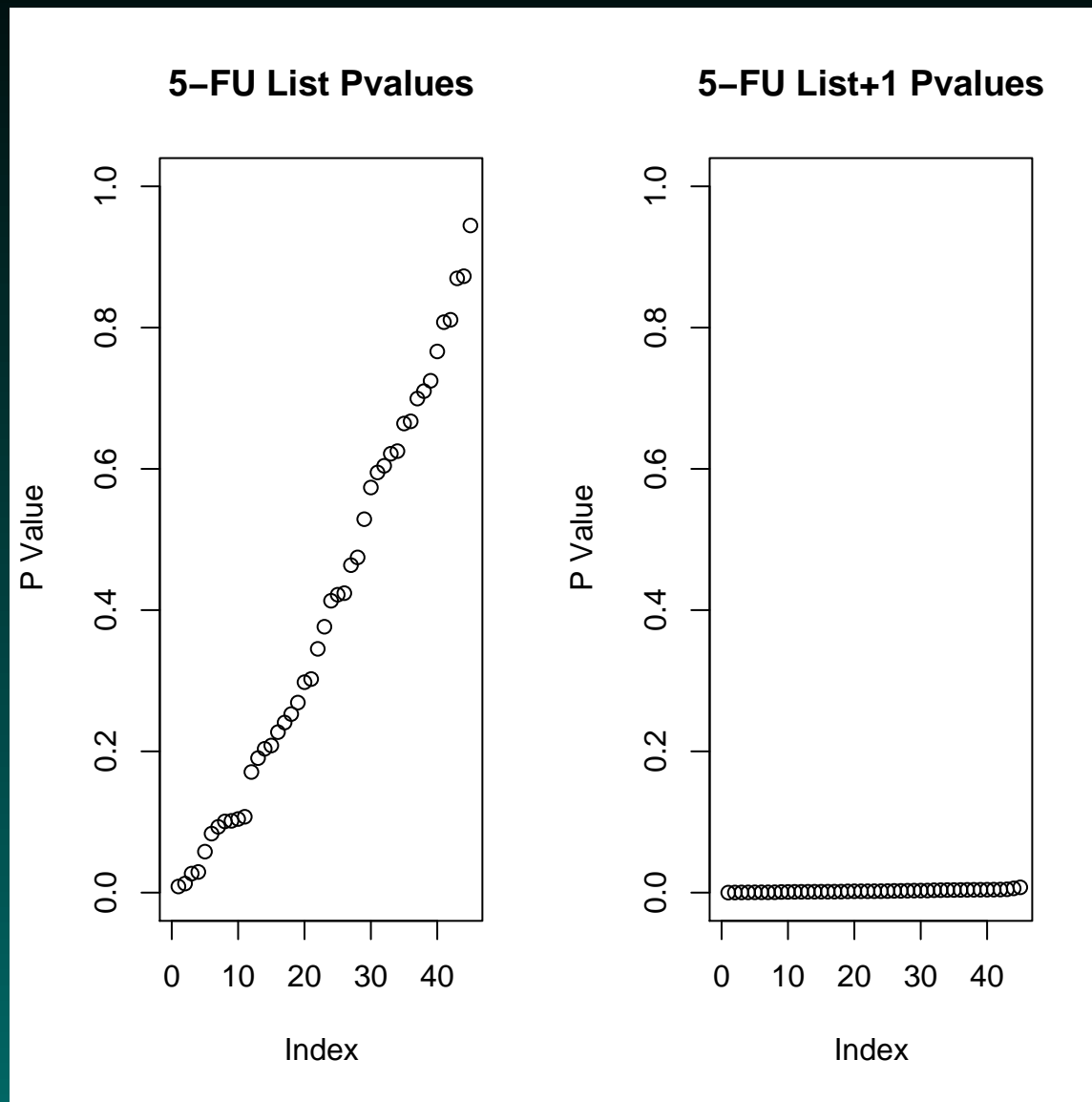
P-values for Drugs+1
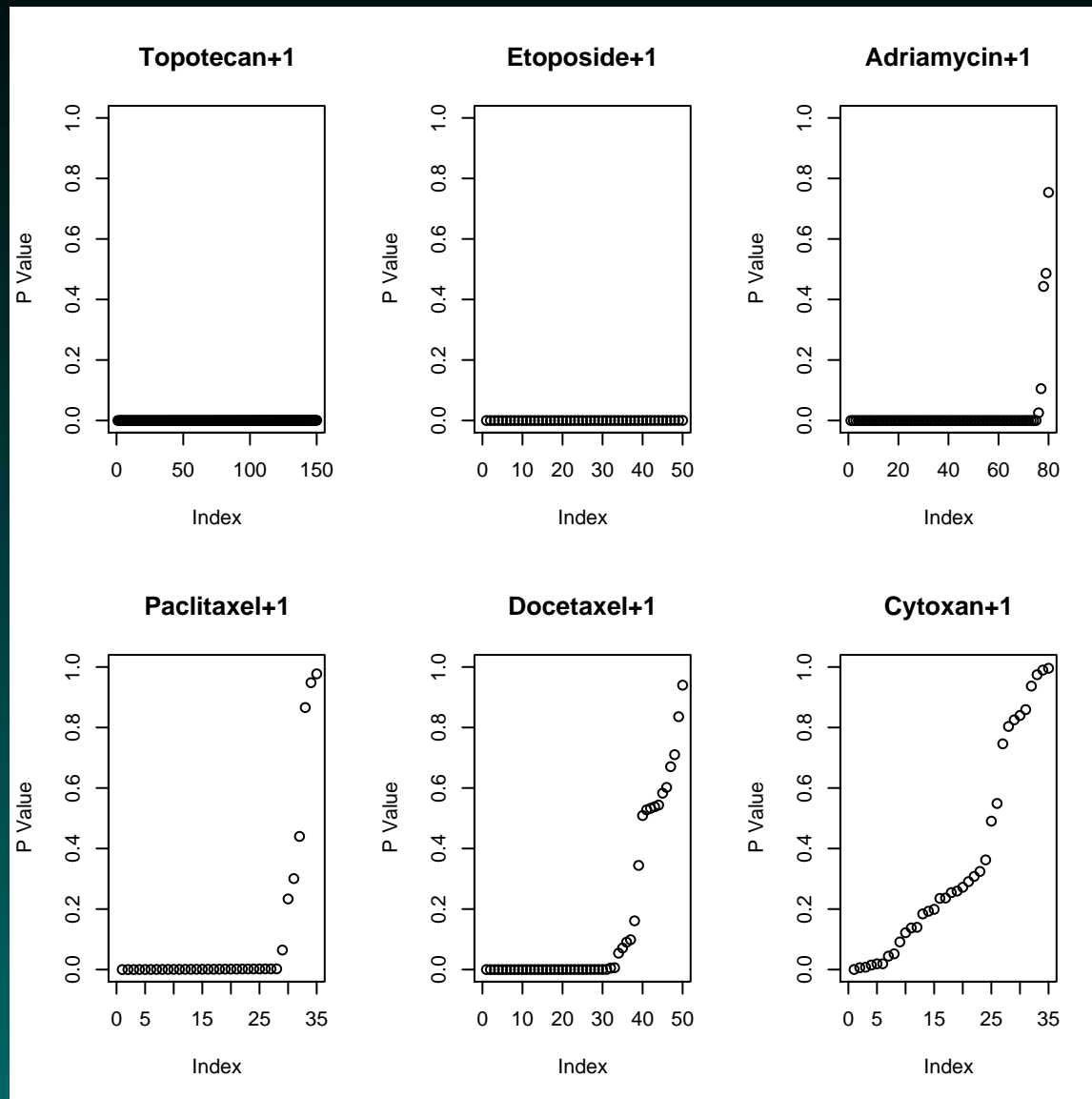
How Their Software Works

We mentioned that the tables of features were available at `http://data.cgt.duke.edu/NatureMedicine.php`; it turns out that there was *software* there as well.

The software requires two files (and some parameter values):

1. a quantification matrix, genes by samples, with a header line giving the classification (0 = Resistant, 1 = Sensitive, 2 = Test)

2. a list of probeset ids in the same order as the quantification matrix. *The list of probeset ids should not have a header line.*

What do we get using their software?

Heatmaps Match Exactly for Docetaxel!
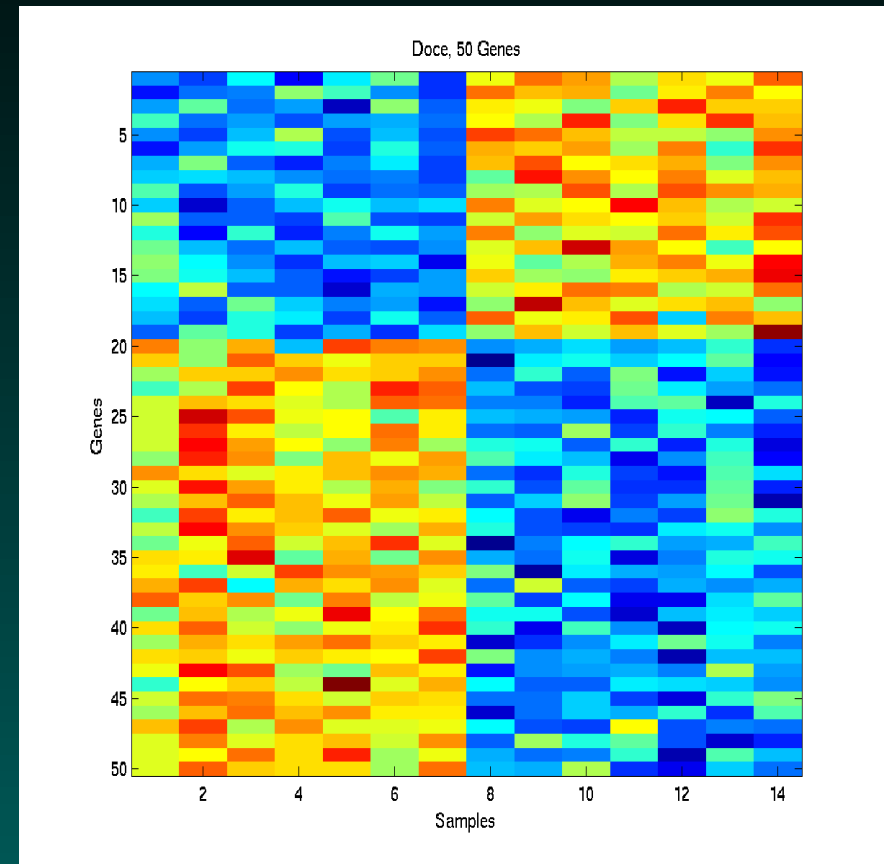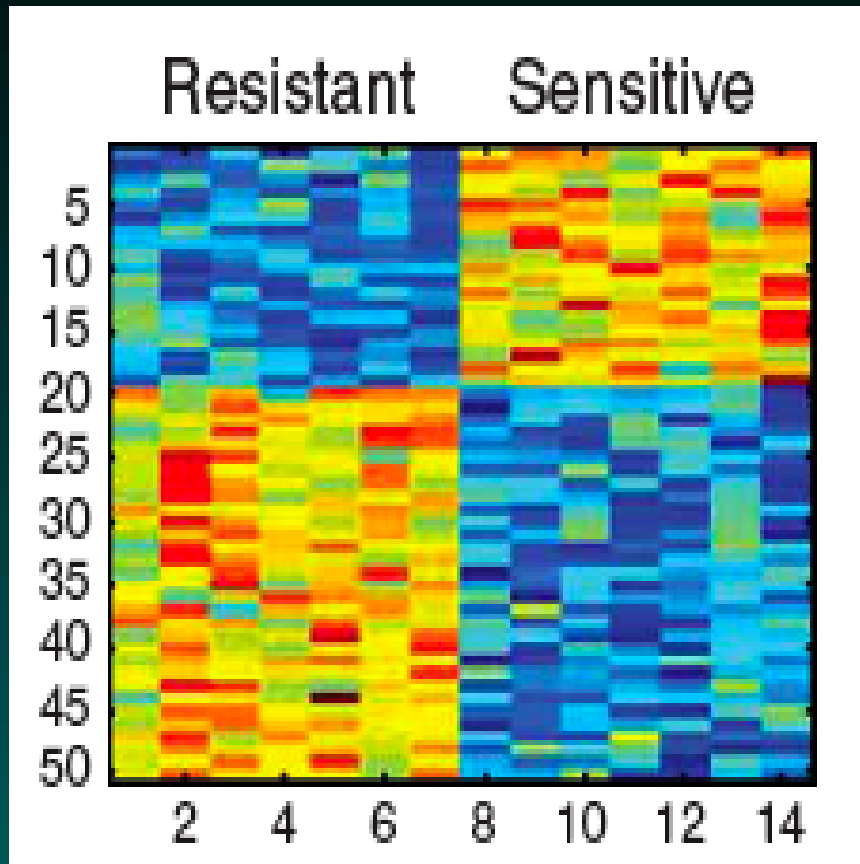
From Potti et al, Figure 1

From the software

Heatmaps Match Exactly for 5 Others!

From the paper:
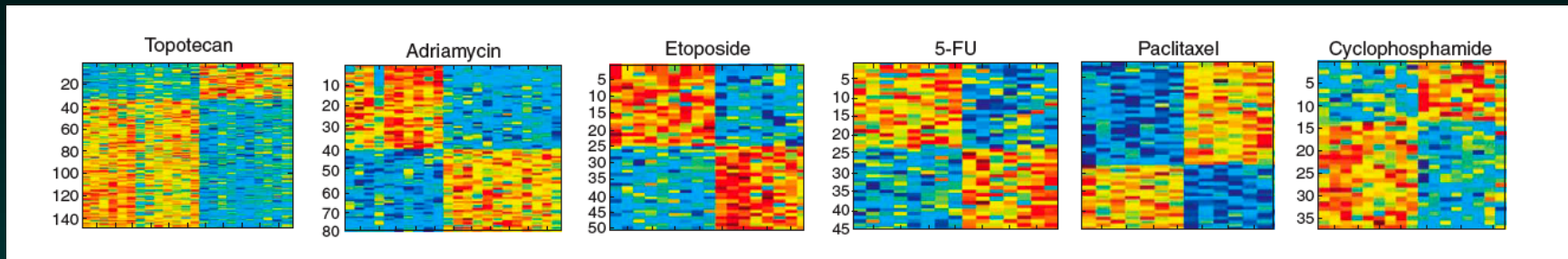
From the software:
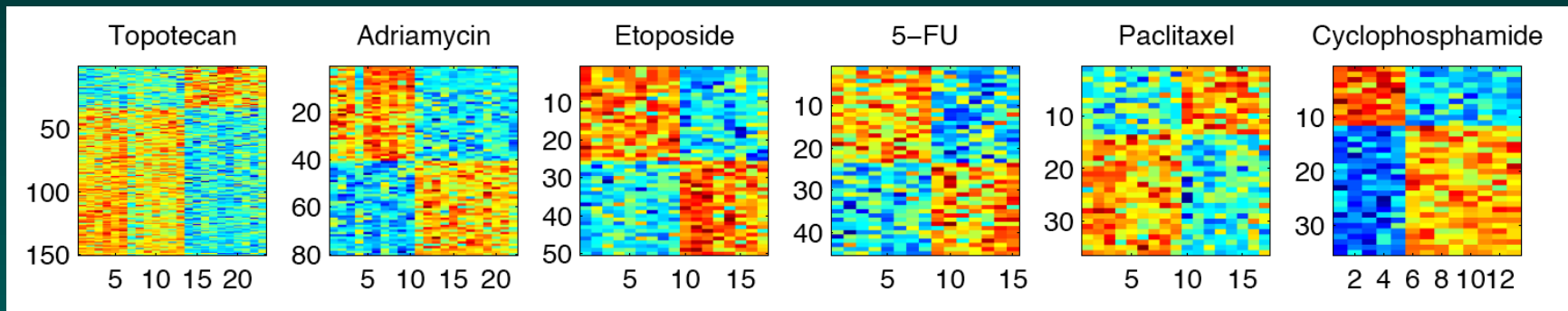
Paclitaxel is mislabeled as Cyclophosphamide in the paper, and we don't have anything that works for Cytoxan.

Observations

- This outcome strongly suggests that we're working with the same cell lines that they used.

- If we supply an id list with a header, we get the same off-by-one error we saw in the reported lists. Note that this doesn't have to be off-by-one alphabetically, but rather with respect to the list supplied.

- Even though the heatmaps match exactly for 6 drugs, the gene lists reported by the software match those reported for only 3. With the other 3, the problems are with the same outliers we found trying to do it ourselves.

Predicting Patient Response to Chemo

Ok, given the cell lines and the features, we want to make *predictions*.

Of course, in making these predictions, we want to know what we have to shoot for.
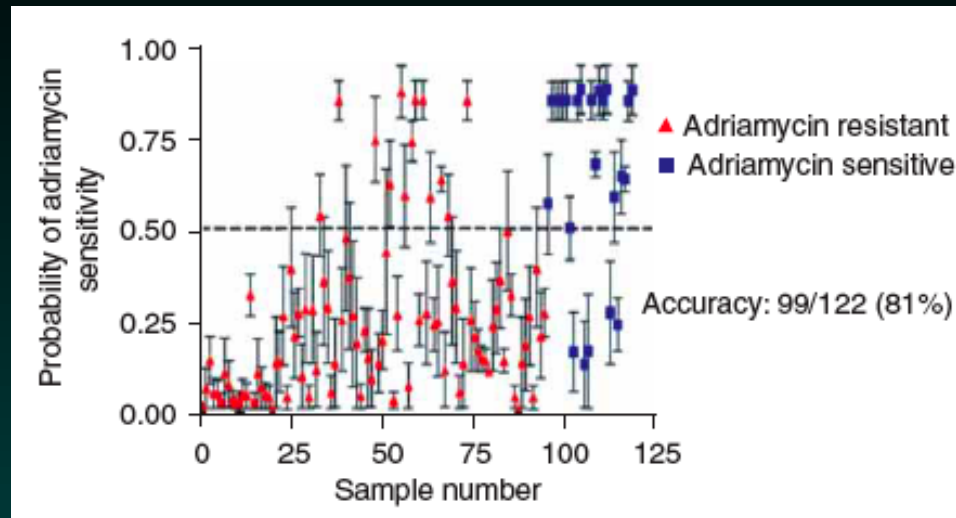
How good were their reported predictions?

Let's look at some pictures...

Predicting Docetaxel (Chang 03)
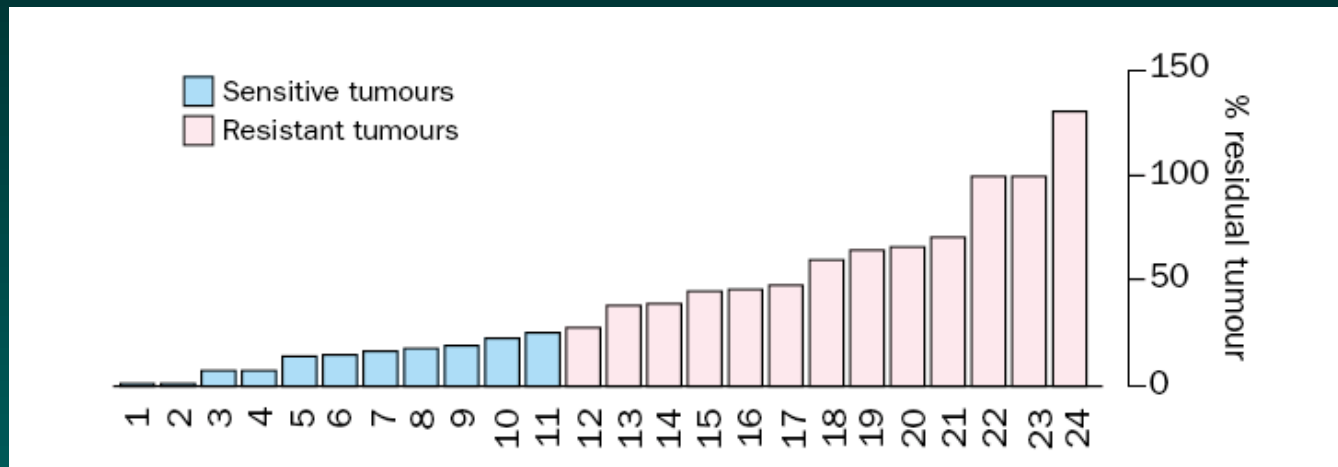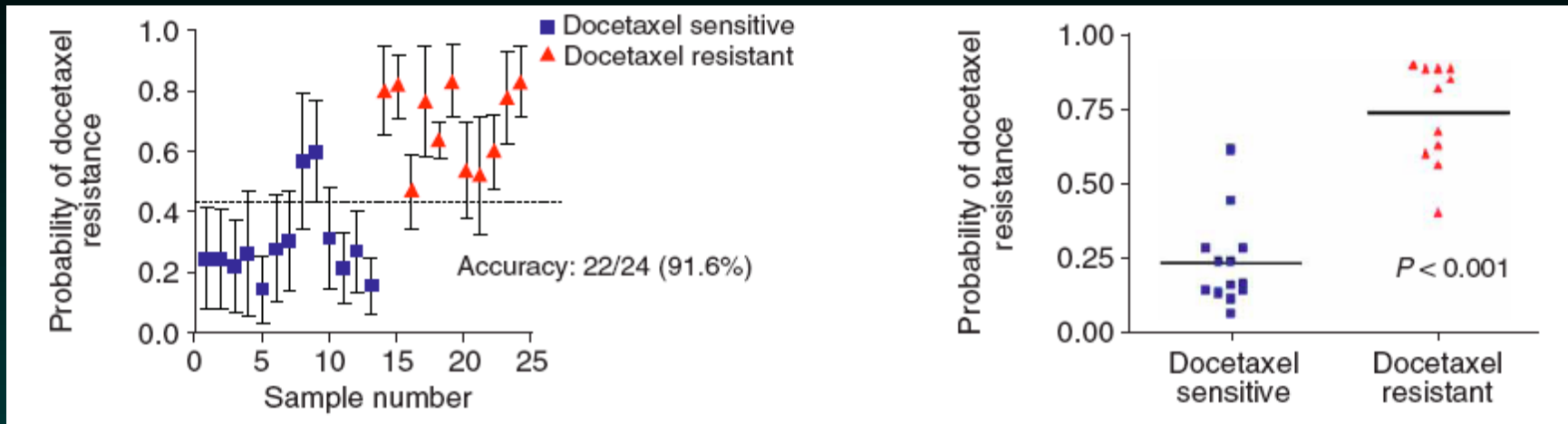
Predicting Adriamycin (Holleman 04)
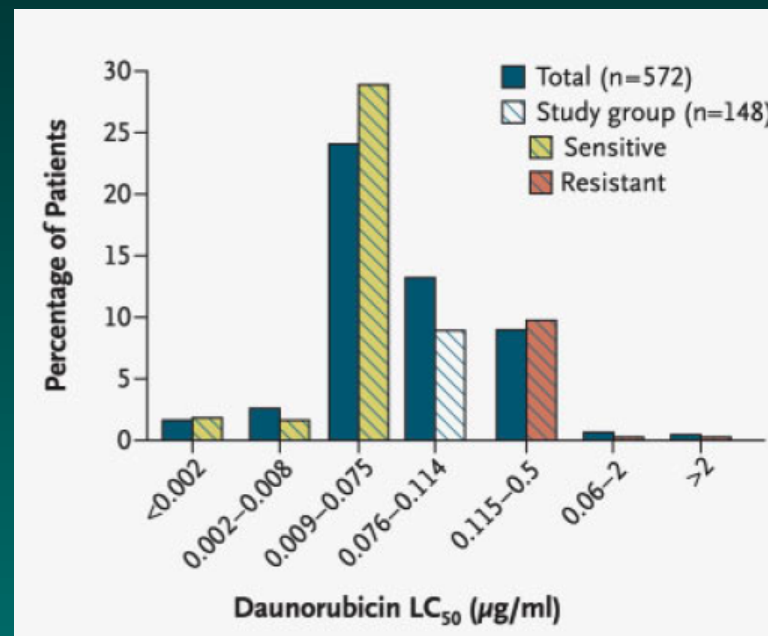
What's Going On?

First, their software seems to be finding structure that we're missing.

Second, this structure can go *the wrong way*.

One area for potential mixup is in labeling samples as "0" or "1" instead of "Sensitive" and "Resistant", but that's not the case for these two drugs.

Another possibility is that the structure present is *unrelated* to the drug response.

Time to muck about in the code...

Mbinregsvd.m (Version 1)

```
% data &  parameters
 [N,n]=size(x); nvalid=length(ivalid); itrain=1:n
 k=2; nit=nmc(1); nbi=nmc(2); nskip=nmc(3);

% reduce to top most correlated ngene  ...
 z=Z; z(ivalid)=[]; X=x; X(:,ivalid)=[];
 ind=select_genes_cor(X,z,ngene); li=ind(1:ngene)

'running optimised training set analysis '
X=x(li,:)-repmat(mean(x(li,:),2),1,n);
A,D,F=svd_mw(X); XX=X;
```

Mostly gobbledegook, but there's something in that last
block...

Mbinregsvd.m (Version 2)

```
'running optimised training set analysis '
%% perform SVD on all samples ...  change
X=x(li,itrain)-
    repmat(mean(x(li,itrain),2),1,ntrain);
%X=x(li,:)-repmat(mean(x(li,:),2),1,n);
```

In Version 1, the SVD identifying the metagenes is applied to an expression matrix involving *both* training and test data.
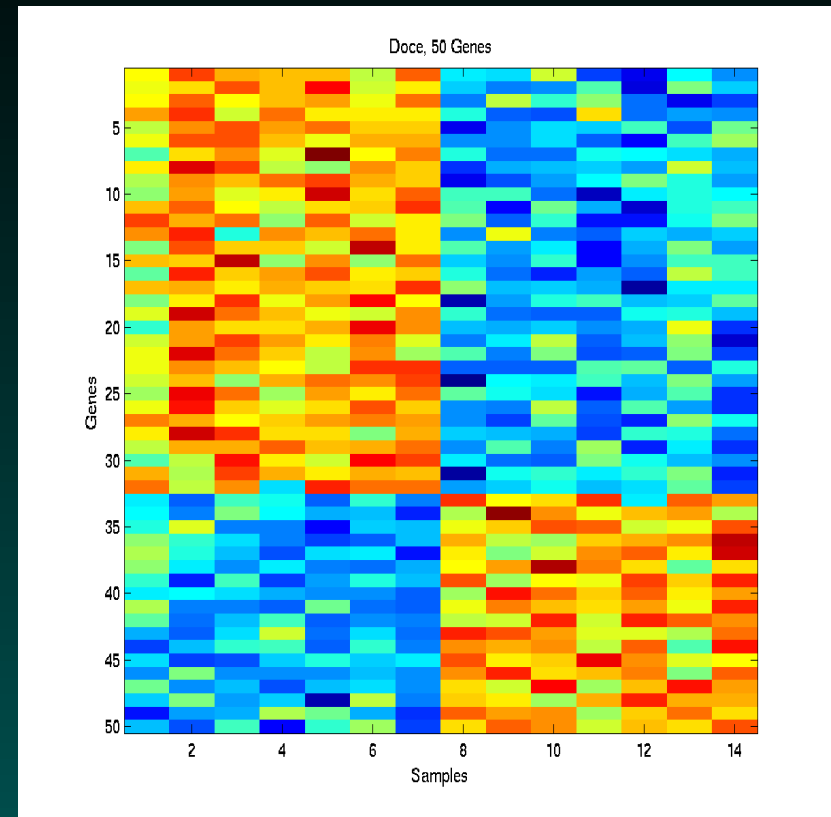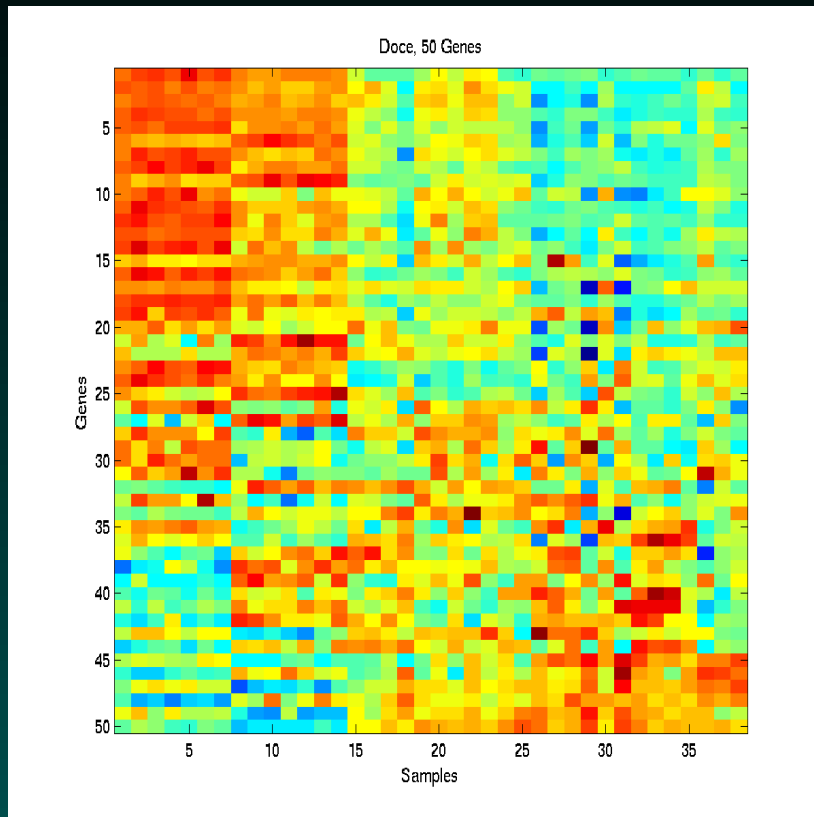
Does this matter?

The Metagene (PCA) plots
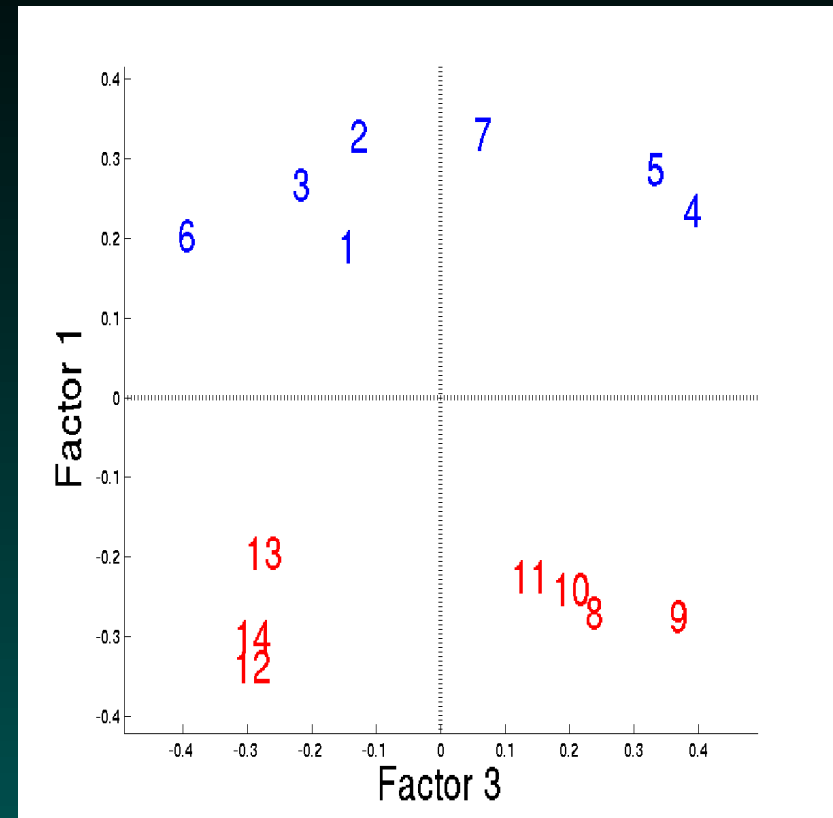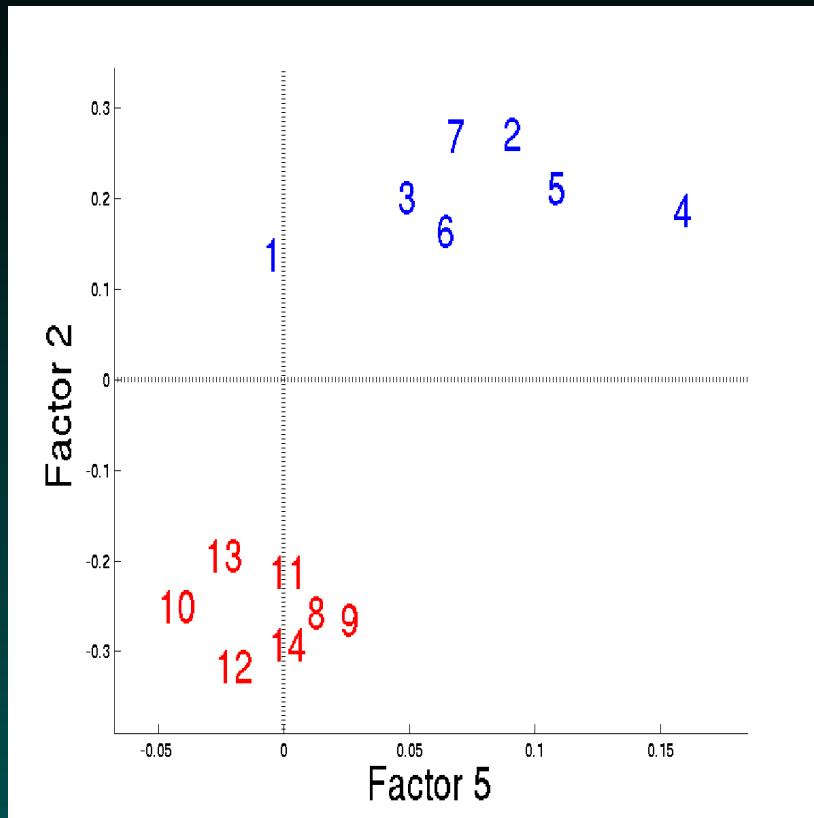
- Version 1

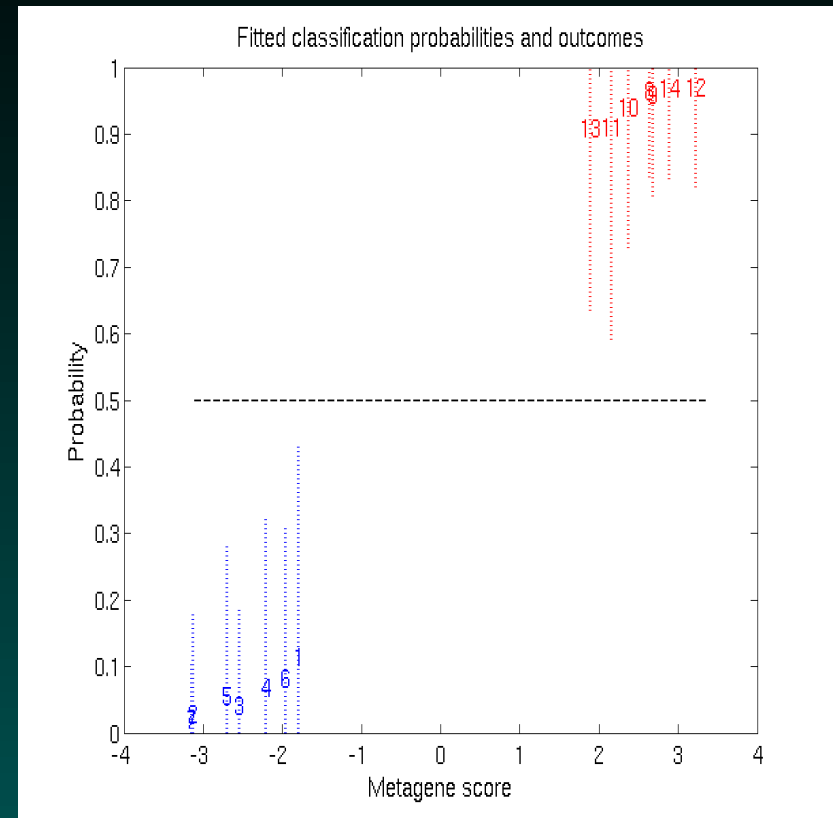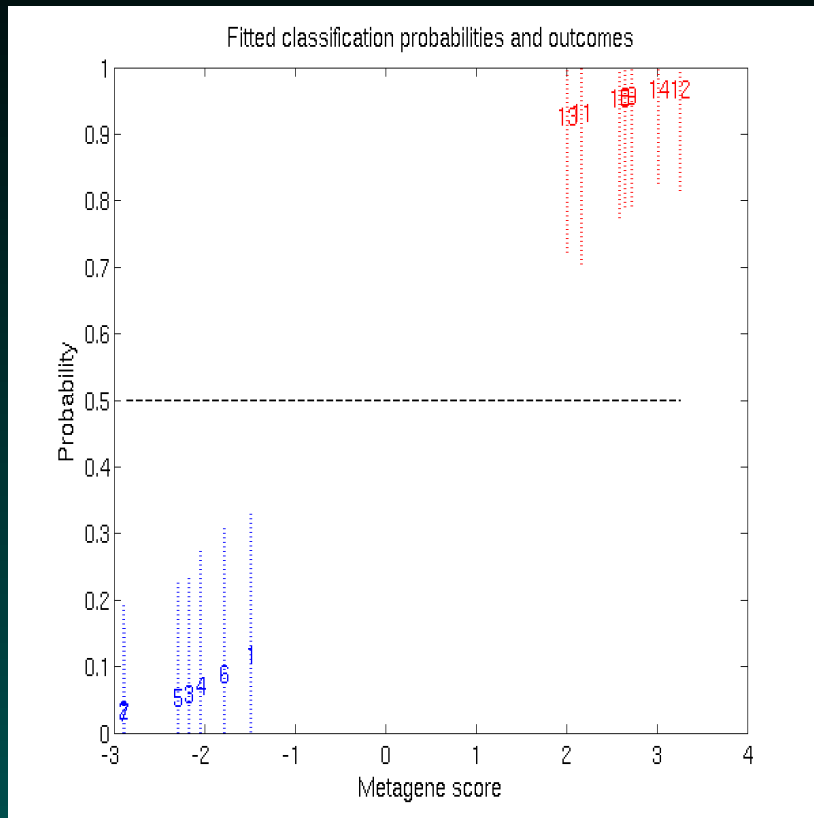- Main data split is along second PC (y-axis)

- Version 2

- Main data split is along first PC (y-axis)

The Heatmaps

- V1; 3 PCs are important. • V2; only uses training data. Between train and test, within training, within test.
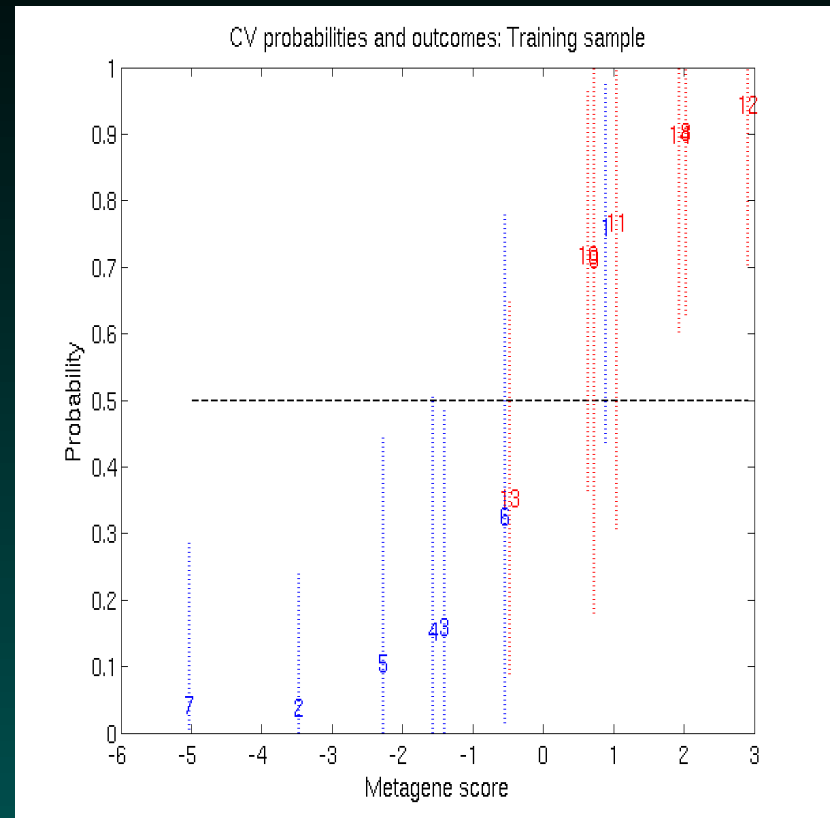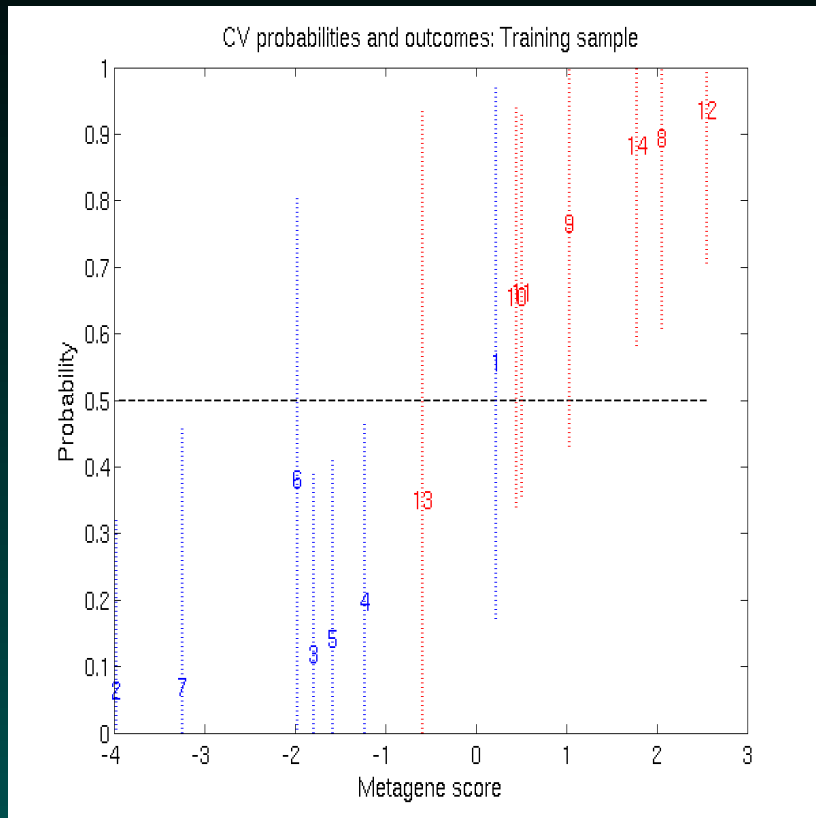
The Training Set Predictions

- V1; We do a good job.

- V2; We do a good job.
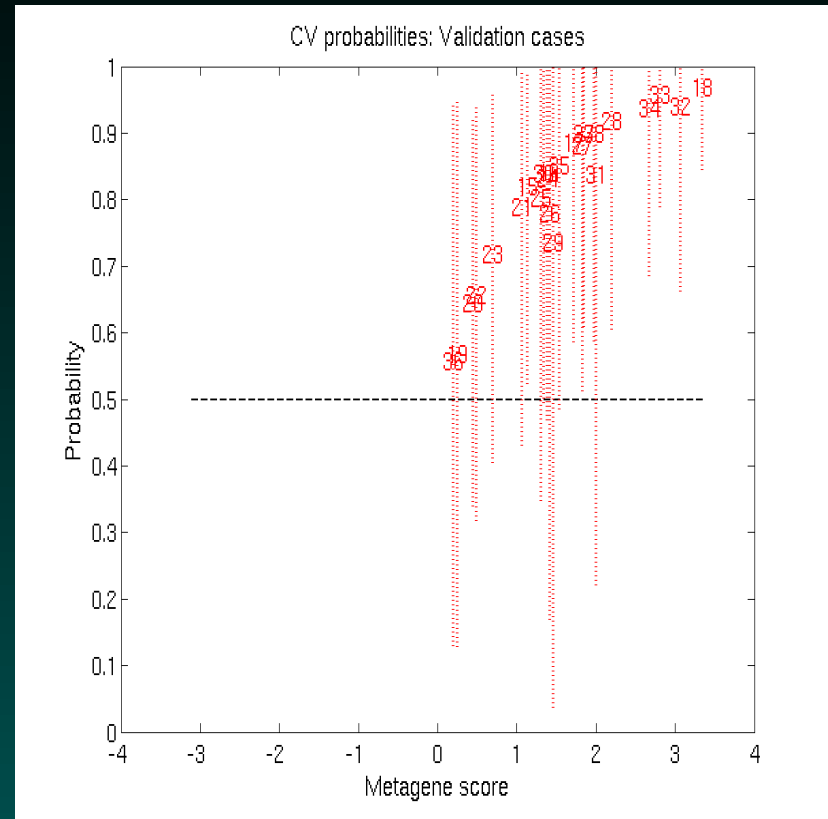
The Training Set Predictions (LOOCV)

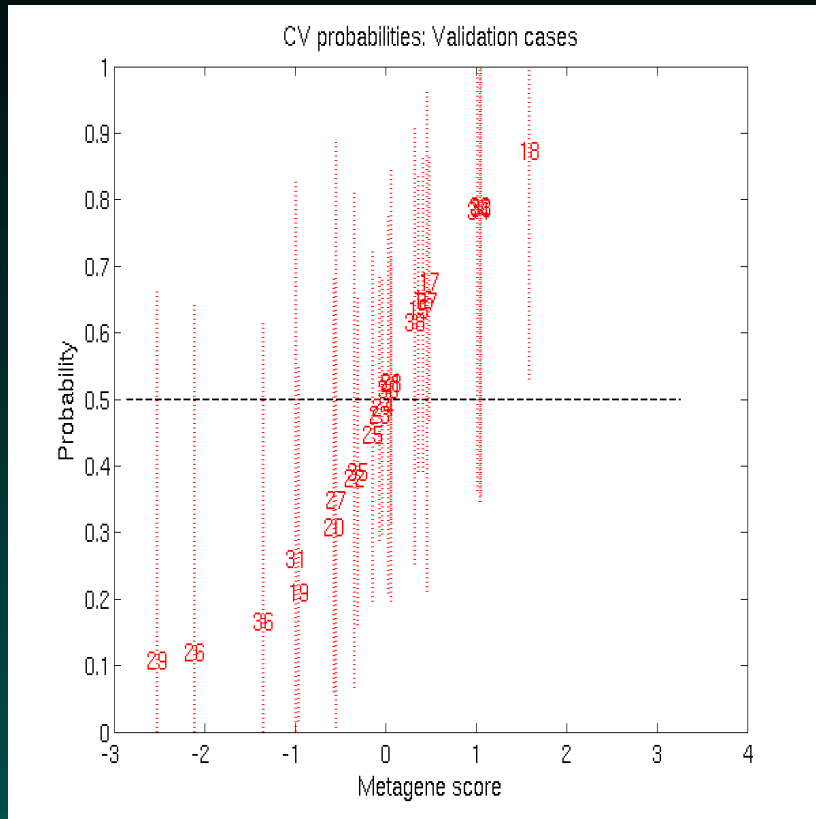- V1; We do a good job.

- V2; We do a good job.

Now, there's just one more...

The Test Set Predictions

- V1; We call some both ways.

- V2; We don't.

Of course, this doesn't mean they're right. If we choose genes using training data, this is mostly standardizing. But.
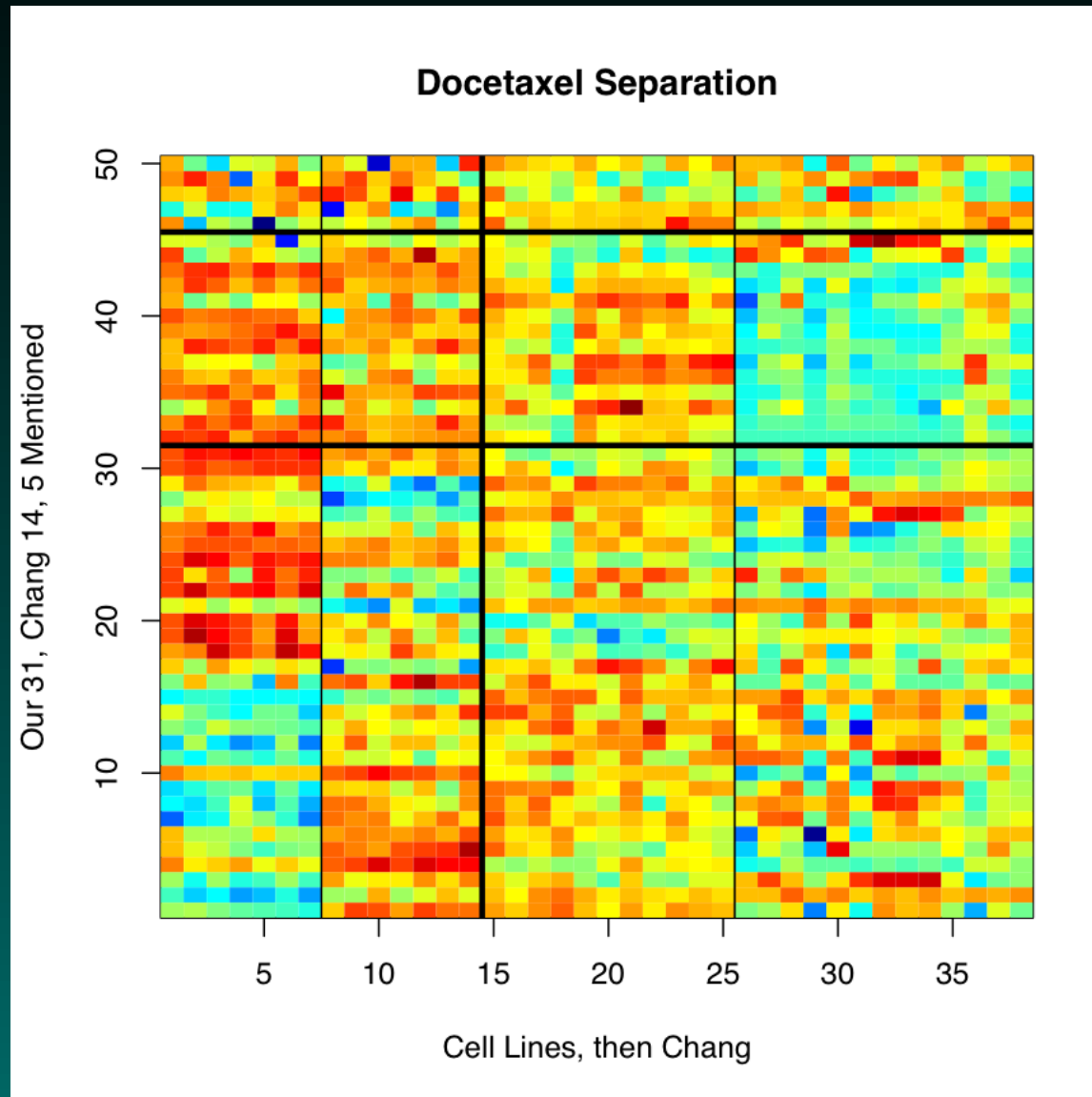
Remember the Outliers!

When we ran their software on the docetaxel training data, we matched 31 of the items in their list after accounting for the off-by-one error. There were 19 outliers.

We stared at these 19. Then we stared at them again, this time without offsetting. Looking at the unadjusted list, 14 of these probesets were on the list of 92 probesets identified by Chang et al as doing a good job separating their data (the test data).

That leaves 5. These 5 we simply can't explain, either in terms of the training or the testing data.

So, What About This Heatmap?

Docetaxel Separation

Our 31, Chang 14, 5 Mentioned

Cell Lines, then Chang

So, What About These P-Values?

Probesets: Software (white), Chang (red), and Paper (green)

So, What About These P-Values? pt.2

Probesets: Software (white), Chang (red), and Paper (green)
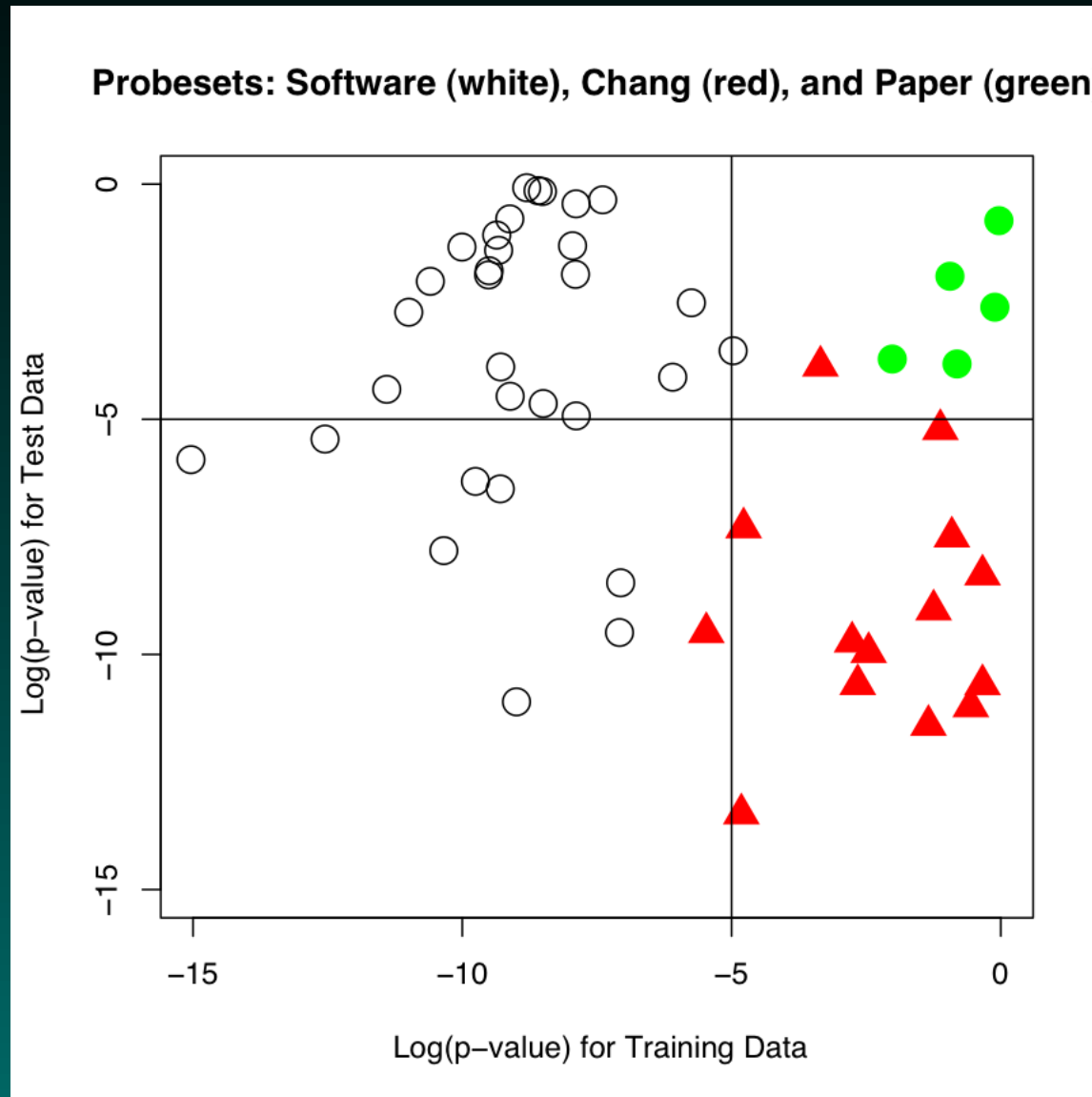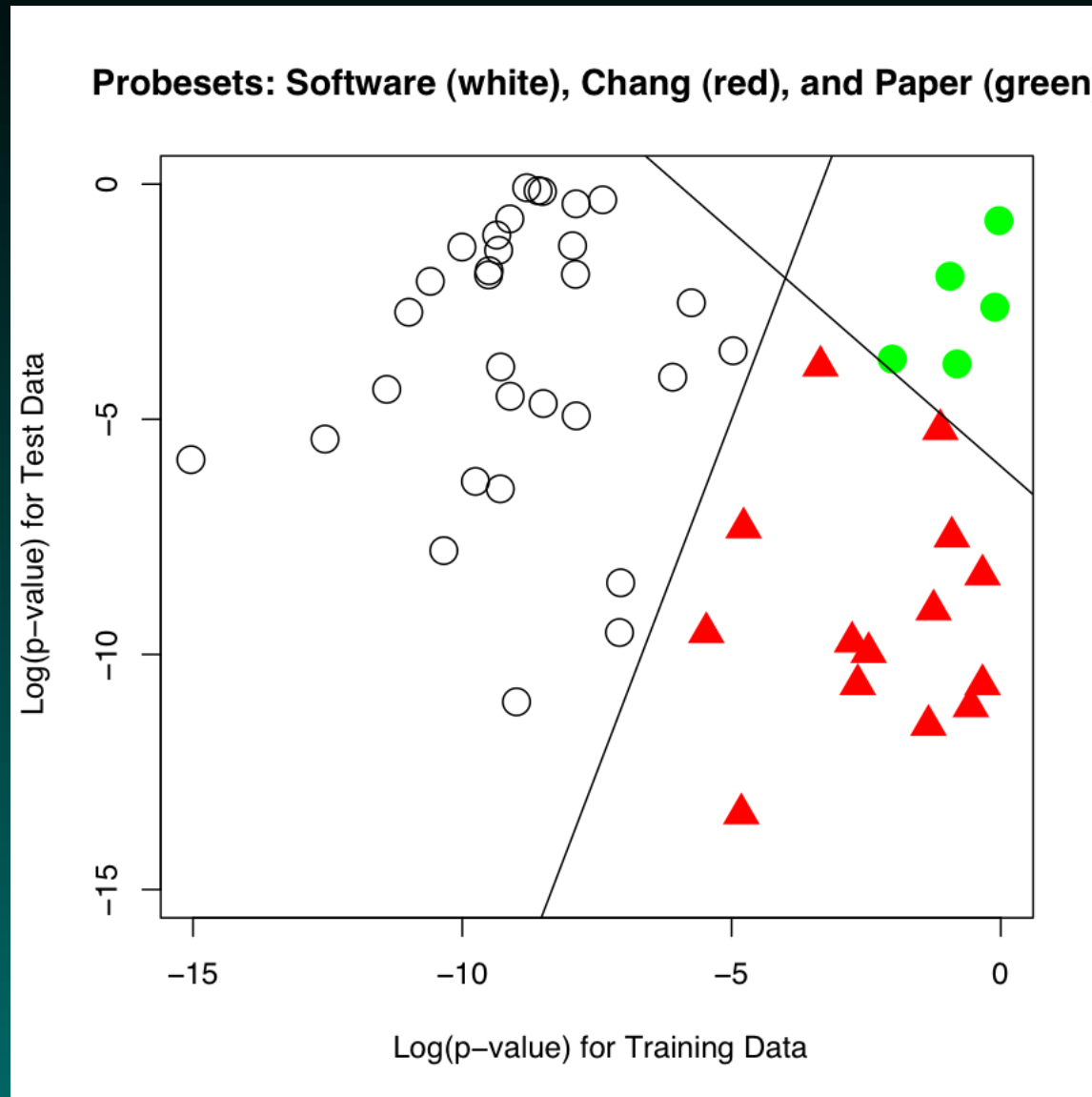
Do the NCI60 Cell Lines Work?

We tried one other trick. Working with docetaxel, we picked *random* cell lines from the A set and arbitrarily assigned them to "Sensitive" and "Resistant" groups. We then selected the best 50 genes, fit PCs, built a model, and made predictions.

We did this a few hundred times, every time keeping track of (a) the fraction of "responsive" patient samples that were classed as "sensitive", and (b) the fraction of "nonresponsive" patient samples that were classed as "resistant".

How well did the real cell lines do?

Predictions With Random Lines
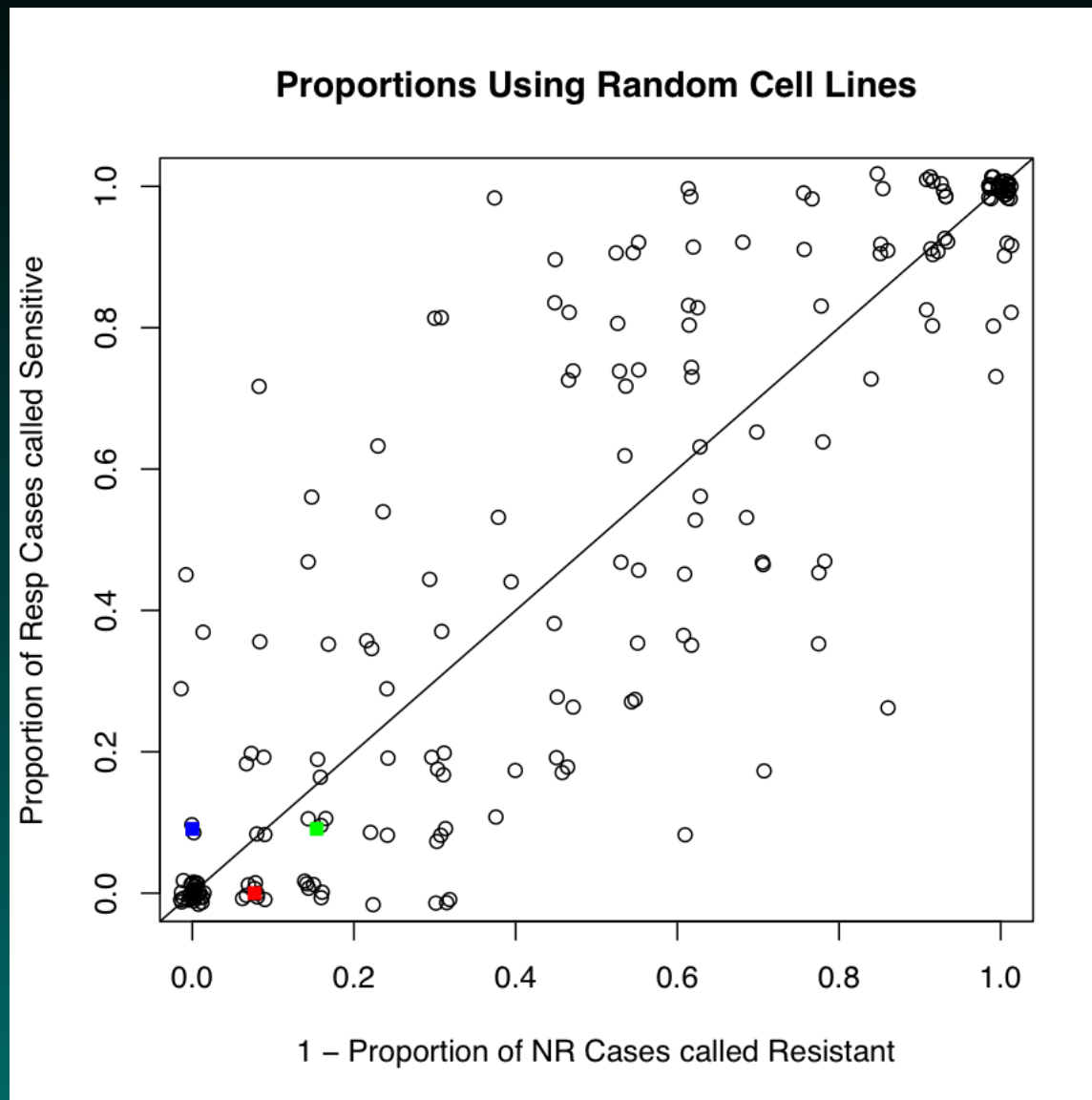
Do We Think It Works?

Brief pause for dramatic tension...

Do We Think It Works?

Brief pause for dramatic tension...

No.

Actually, we might be more surprised if it *did*.

- the cell lines are from a variety of different tumor types with known differences in reponsiveness.

- the training and test sets were run at very different times and under different conditions.

- different array platforms and definitions of sensitivity were used.

What About This is Hard?

Metagene MCMC details aside, *it's not really the math.*

With complex high-throughput analyses, the most difficult step in many cases is simply documenting the analysis clearly and in sufficient detail for it to be reproduced.

Note: this is not just so that *others* can reproduce it. It's so that *your own lab* can reproduce it months later.

Is This an Isolated Case?

Some figures from Economics (Dewald, Thursby and Anderson (1986), Amer Econ Rev, 76:587ff):

Acquiring data: positive response rate when data was requested for published papers: 22/62. This fraction supplied data and/or code.

Positive response rate when data was requested before the paper had appeared: (75 or 68)/95. This fraction either (1) replied, or (2) supplied data and/or code.

Checking datasets: 54 datasets checked. Number with sufficient detail to attempt reproduction: 8.

Number where reproduction was achieved: 2.

What We're Doing At MDA

Do our own analyses clear the reproducibility bar?

Not always painlessly. We're working on developing more template analyses using *Sweave*, where documentation has been interweaved with the code used to produce the analysis.

Given the Sweave source and the data files, anyone else can run our analysis and get the same results. We're assembling the web site showing the results I discussed here even now, so you can see how well we're doing.

What We're Doing At MDA

Do our own analyses clear the reproducibility bar?

Not always painlessly. We're working on developing more template analyses using *Sweave*, where documentation has been interweaved with the code used to produce the analysis.

Given the Sweave source and the data files, anyone else can run our analysis and get the same results. We're assembling the web site showing the results I discussed here even now, so you can see how well we're doing.

Only when the analysis is reproducible can we talk about whether or not it's "right".

An Observation

Drs Nevins and Potti made the data available, and they tried to answer the questions that we put to them. This was not always easy to do.

They believe that their method works, and they have made a good faith effort to show us how.

We disagree, but we may have gotten something wrong.

Our documentation is available on the web:
`http://bioinformatics.mdanderson.org/`
`Supplements/ReproRsch-Chemo/`.

Look it over, and decide what you think.