

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

September 10, 2009

Lecture 4: Cell Lines, Chemo, and Repro

Microarrays let us simultaneously measure the mRNA expression levels of thousands of genes in a sample of interest.

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.

What do we know about drug effectiveness?

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>

Using the NCI60 to Predict Sensitivity

ature.com/naturemedicine

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 Chan⁴, Robyn Sayer⁴,
Janiel Cragun⁴, Hope Cottrill⁴, Michael J Kelley², Rebecca Petersen⁵, David Harpole⁵, Jeffrey Marks⁵,
Andrew Berchuck^{1,6}, Geoffrey S Ginsburg^{1,2}, Phillip Febbo¹⁻³, Johnathan Lancaster⁴ &
Joseph R Nevins¹⁻³

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

The main conclusion is that we can use microarray data from cell lines (the NCI60) to define drug response “signatures”, which can be used to predict whether patients will respond.

They provide examples using 7 commonly used agents.

This got people at MDA very excited.

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!

How it Works

1. Identify and **collect the data** sets.
2. Using the sensitivity information for a drug of interest (docetaxel) **select the extreme cell lines**.
3. Using the array profiles of these cell lines, **select the features** (genes) that best distinguish sensitive from resistant.
4. Use array values for the chosen features to **train** a binary model distinguishing sensitive from resistant cell lines.
5. **Test** the model for its ability to make accurate predictions using expression data sets from patient tumors.

Gathering Data

1. **Drug response:** assays on NCI60 from DTP at NCI (http://dtp.nci.nih.gov/docs/cancer/cancer_data.html)
2. **Training:** Affymetrix U95Av2 arrays on NCI60, performed in triplicate by Novartis (<http://dtp.nci.nih.gov/mtargets/download.html>)
3. **Testing:** 24 breast tumors on U95Av2; Chang et al (2003) Lancet, 362:362-9. GSE349, GSE350 from GEO. (GSM4913 is mislabeled; It should be “sensitive”. Pers comm.)

Identifying the Drugs

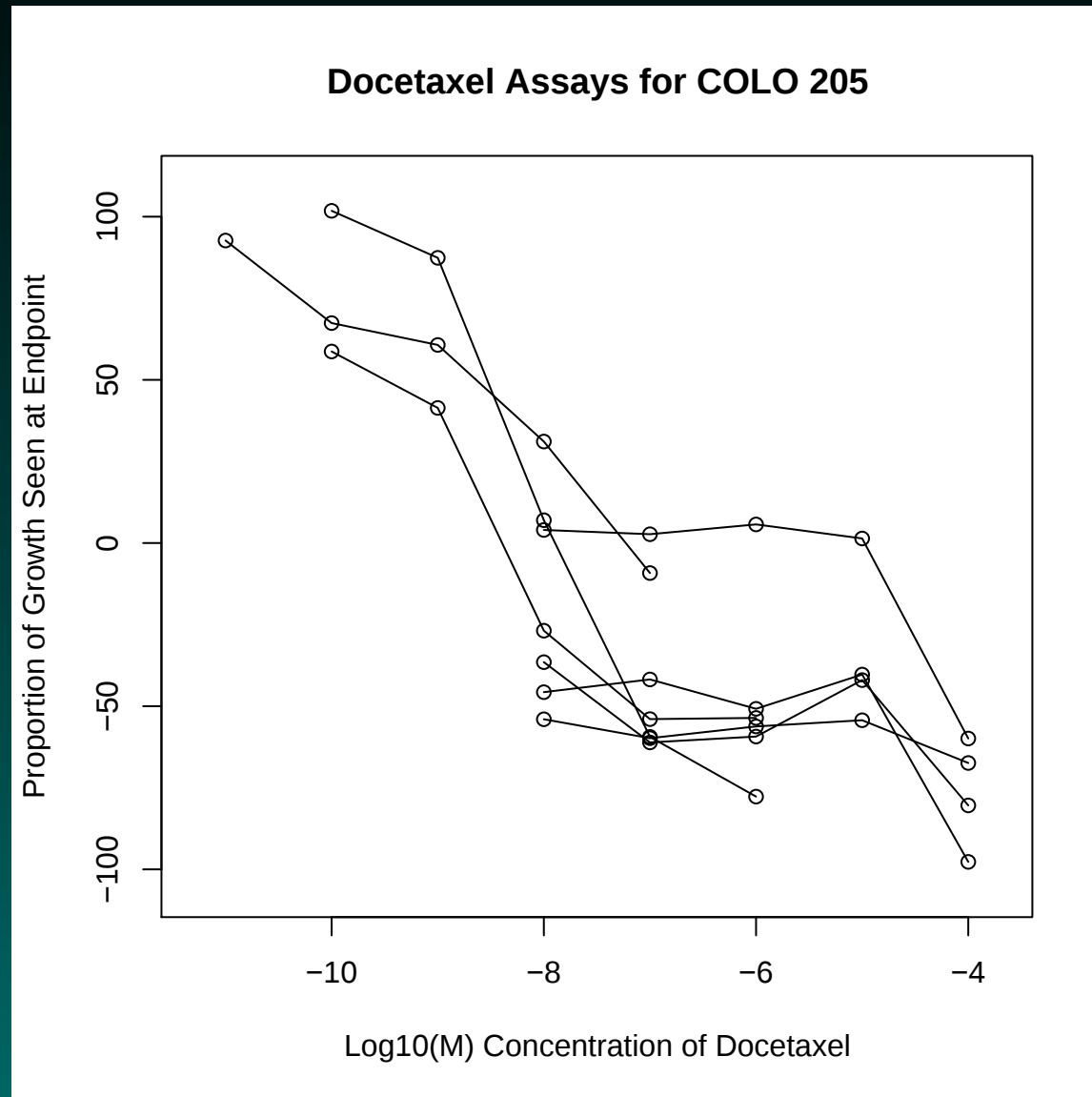
The paper gives the *names* of the drugs profiled, but response data is indexed by NSC number. How would you find these numbers?

Identifying the Drugs

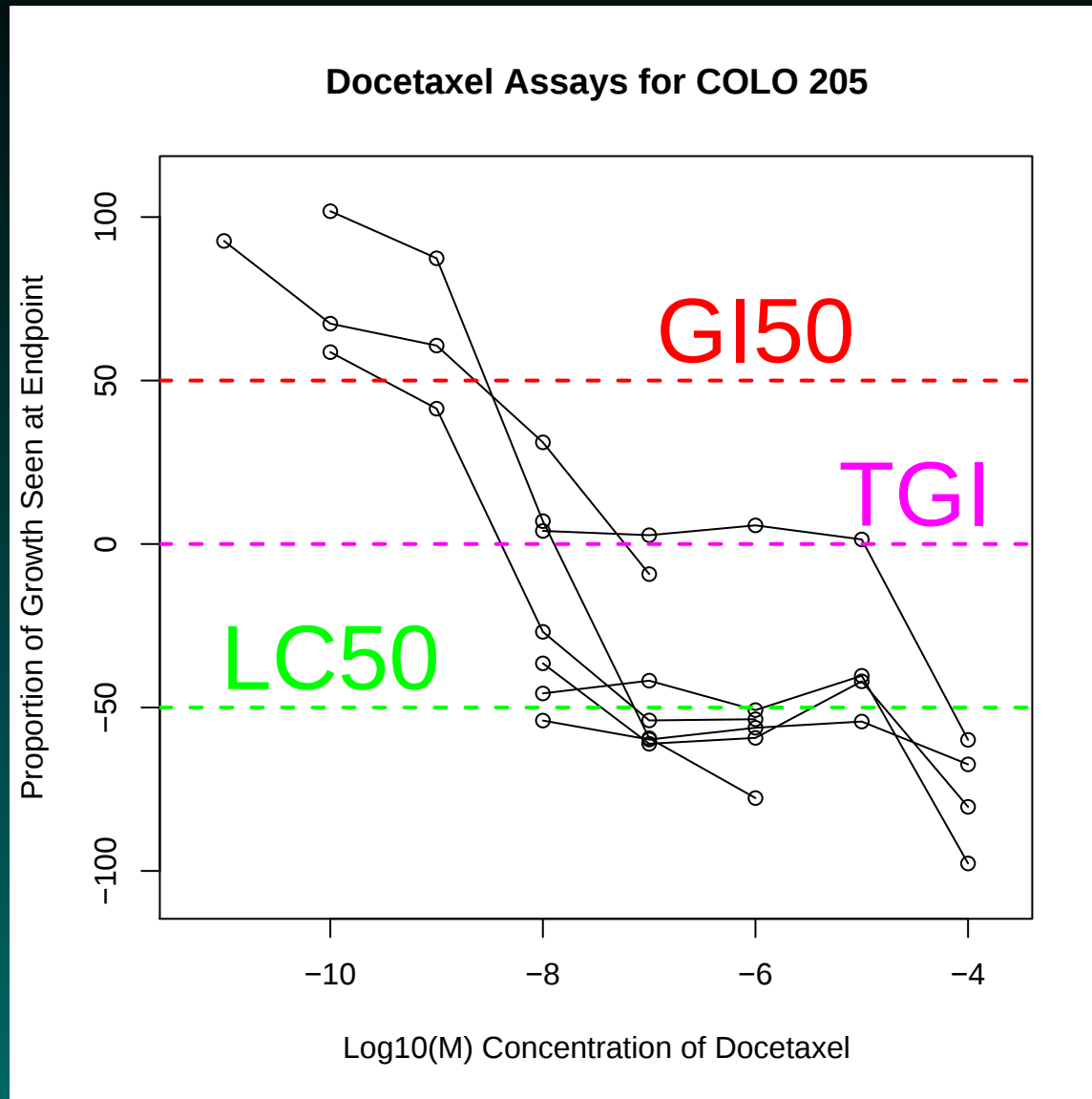
The paper gives the *names* of the drugs profiled, but response data is indexed by NSC number. How would you find these numbers? **Google!** (NCI Drug Dictionary)

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

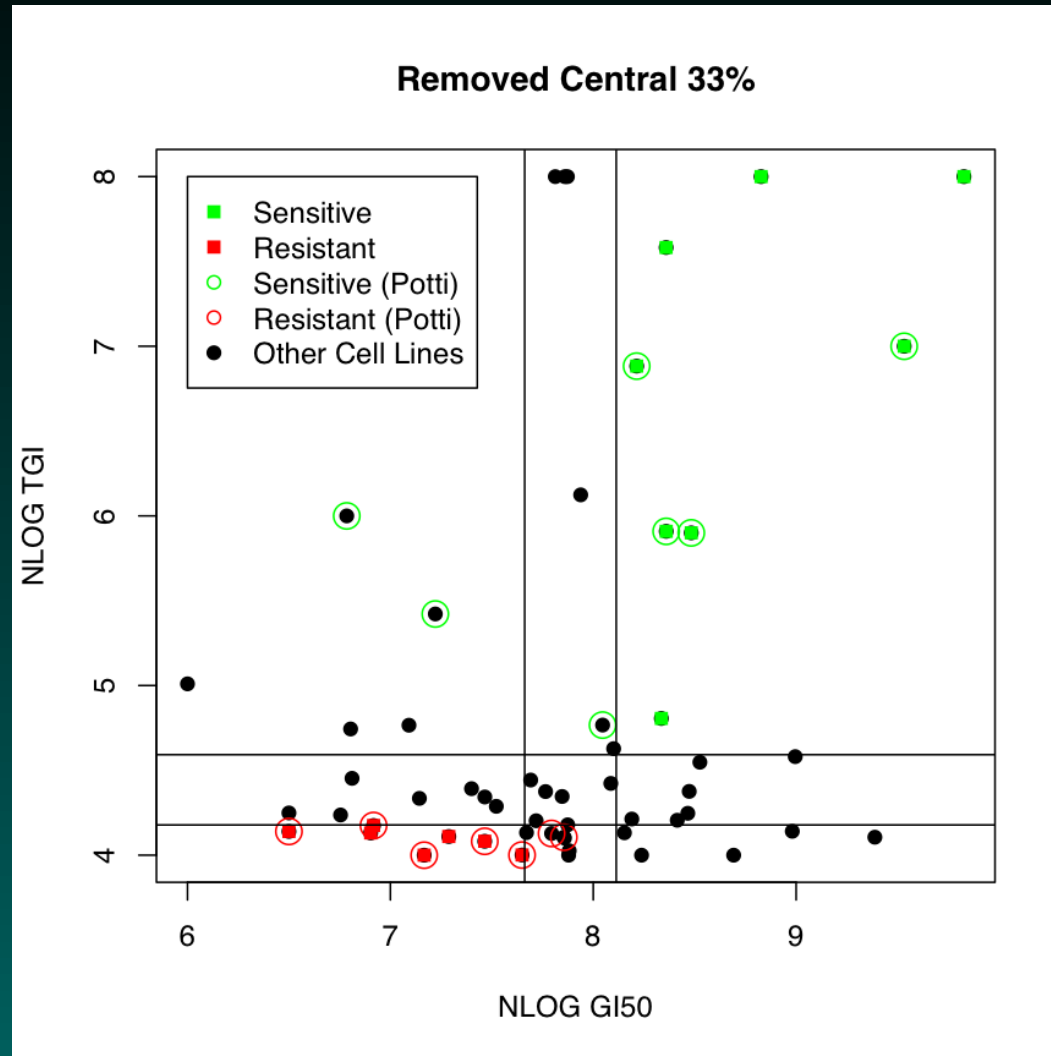
What Drug Sensitivity Data Looks Like



Adding Cutoffs



Selecting Cell Lines for Docetaxel



Use GI50 and TGI; LC50 shows little change.

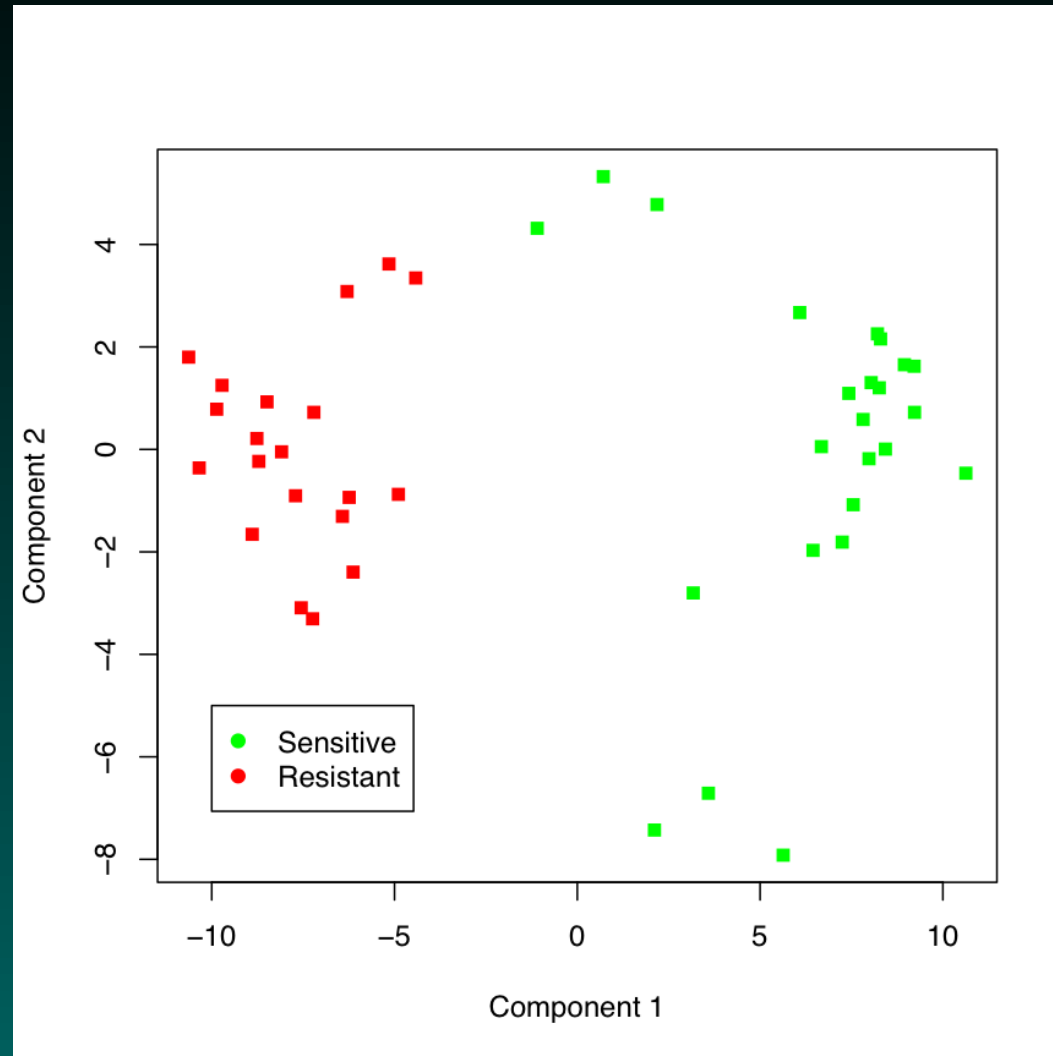
Building a Model

Following Potti et al., we selected the top 50 genes based on a two-sample t-test between sensitive and resistant cell lines.

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.*

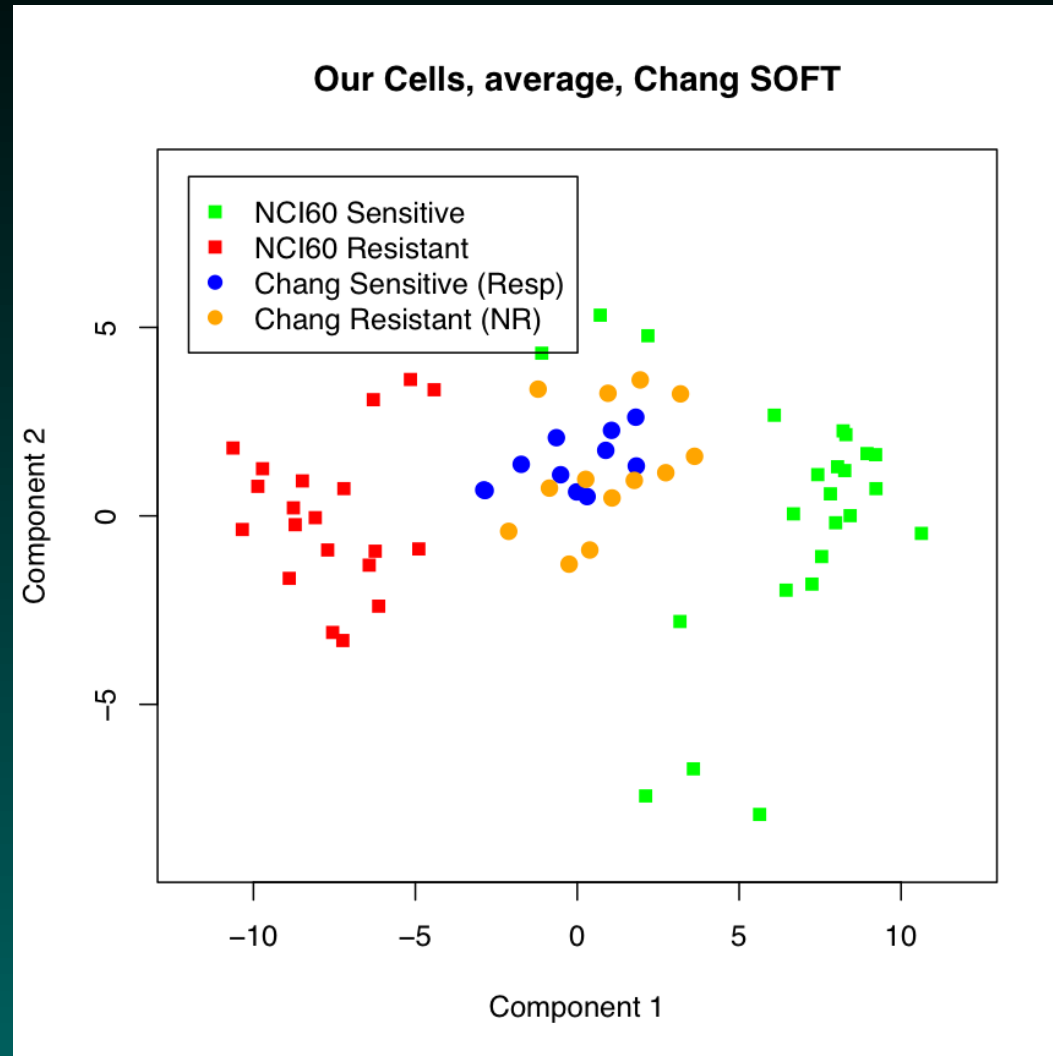
For the model, they use probit regression with the metagene scores to separate sensitive and resistant groups.

Fit Training Data



We want the test data to split like this...

Fit Testing Data



But it *doesn't*. Did we do something wrong?

Forensic 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 features were selected?
- How were the models built?
- 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 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
Cyttox0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1
1 1 1 Cyttox1 Topo0 0 0 0 0 0 0 0 0 0 0
0 0 0 1 1 1 1 1 1 1 1 1 1 Topo1 Taxo10
0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 Taxo11
36460_at 41.671947 21.820335
125.794838 93.459251 79.06321

```

Does this answer the question?

The First 2 Rows...

```

probe_set Adria0 0 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
Cyttox0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1
1 1 1 Cyttox1 Topo0 0 0 0 0 0 0 0 0 0 0
0 0 0 1 1 1 1 1 1 1 1 1 1 Topo1 Taxo10
0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 Taxo11
36460_at 41.671947 21.820335
125.794838 93.459251 79.06321

```

Does this answer the question?

The Novartis Data

Some of the first few 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_A, POLR1C, SF-539, CNS, 12,  
16, 41.671947, A, 0.189687  
...  
36460_at, GC26855_A, POLR1C, SNB-75, CNS, 12,  
5, 21.820335, A, 0.438361
```

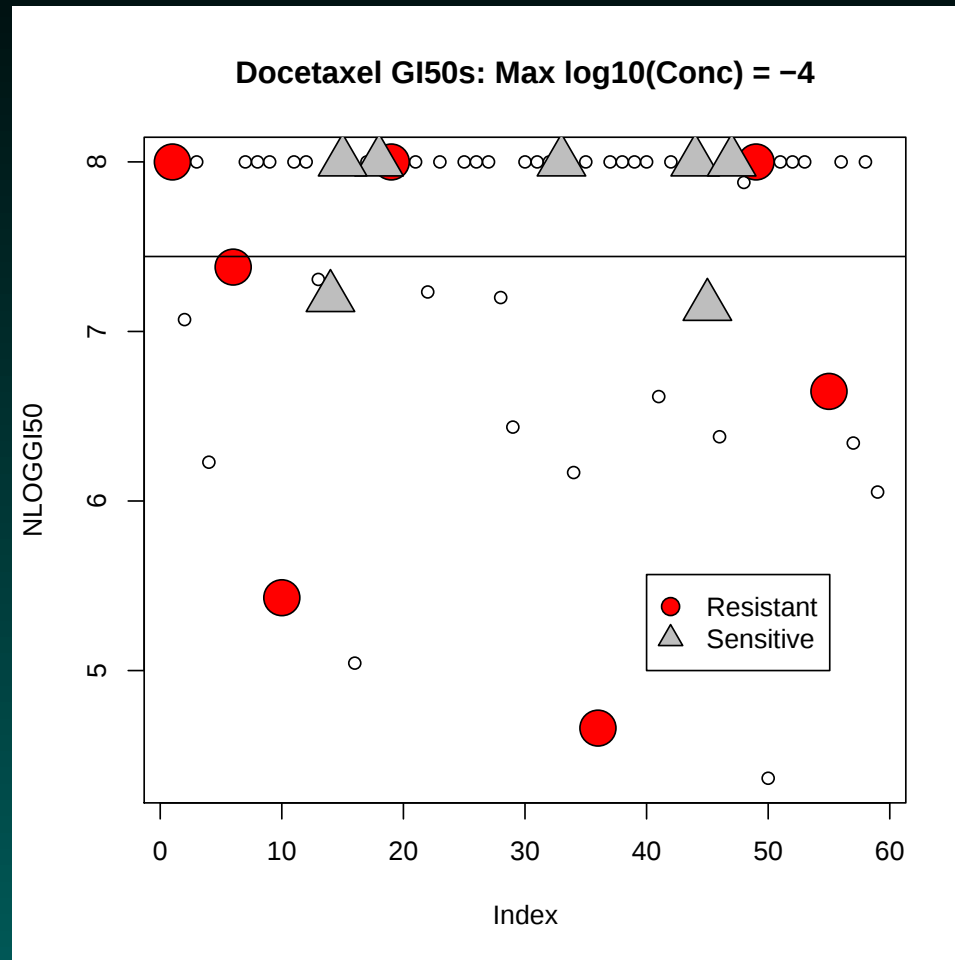
Using this approach, we can match the cell lines used for 6 of the drugs examined (not cytoxan). The “A” series replicates were used throughout.

How Are Cell Lines Chosen?

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. . . .”

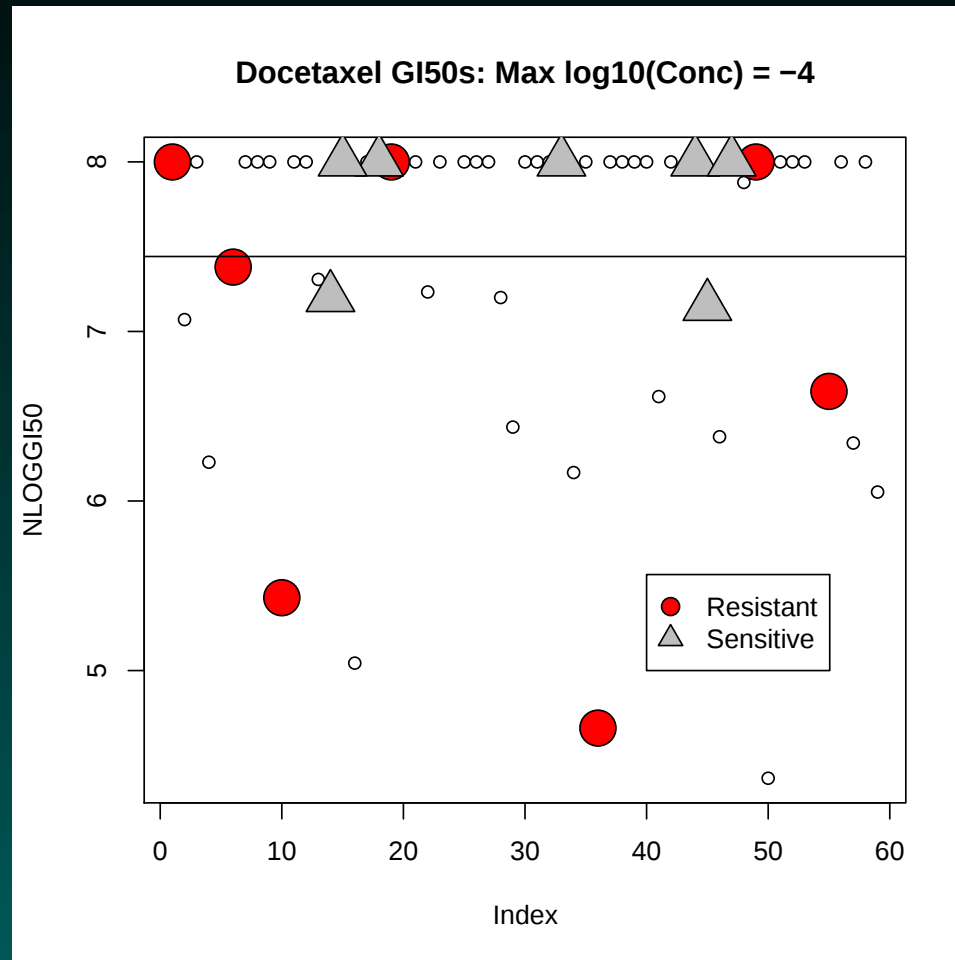
So, how clearly separated are the lines used?

Docetaxel GI50s



Values *overlap*?

Docetaxel GI50s



Values *overlap*? This holds for all drugs tested.

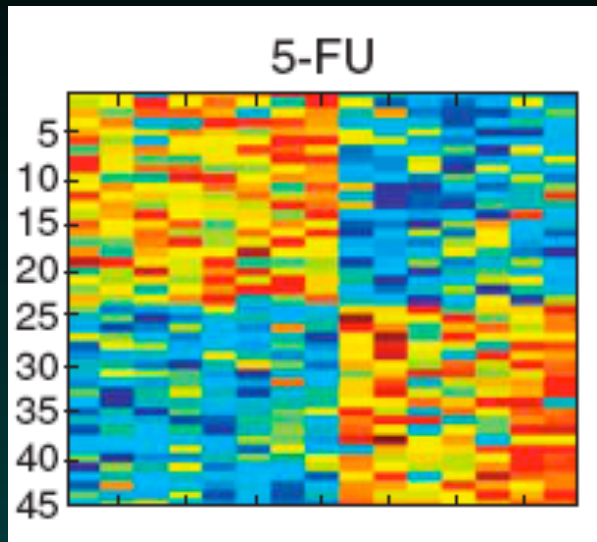
The cell lines don't make sense. What about the features?

They Reported the Features!

Lists of the probesets used were supplied in supplementary table 1, and at the website named in the supplementary methods document: <http://data.cgt.duke.edu/Combo1.php> (now NatureMedicine.php). The paper explains why many of these genes make sense.

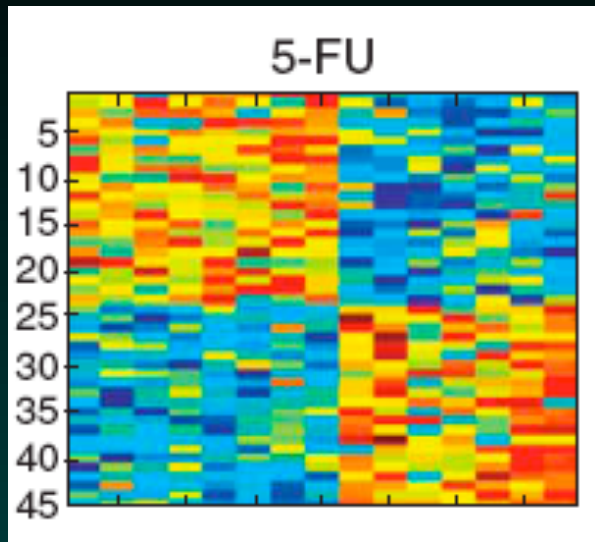
How were these found? According to the supplementary methods: *“a variance fixed t-test was used to calculate significance”*.

5-FU Heatmaps

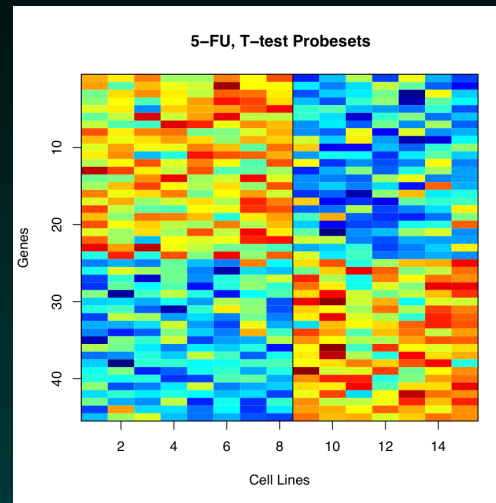


Nat Med Paper

5-FU Heatmaps

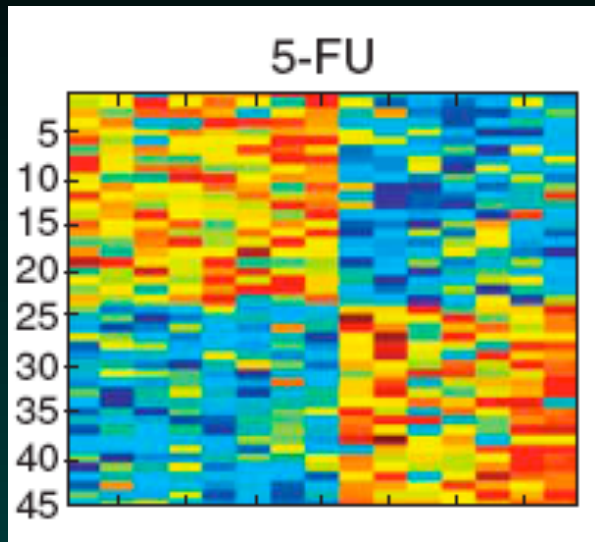


Nat Med Paper

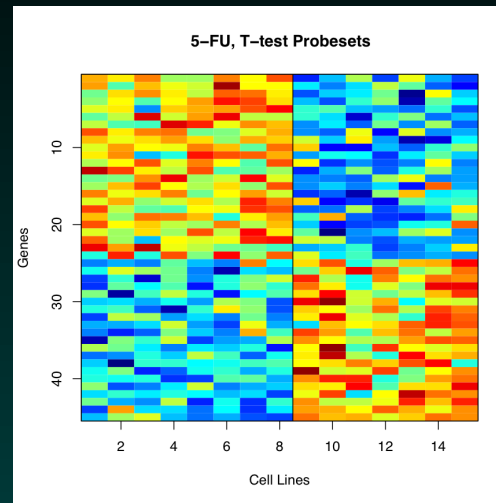


Our t-tests

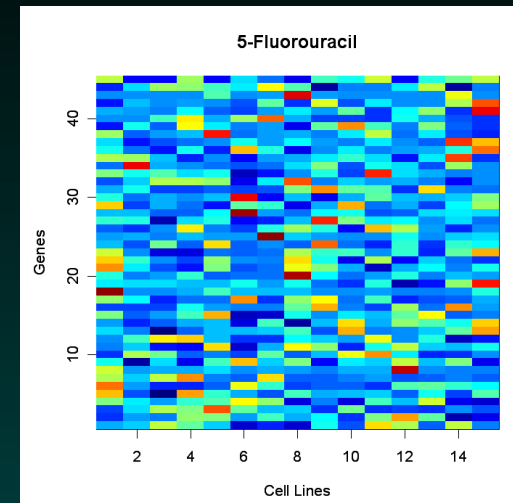
5-FU Heatmaps



Nat Med Paper

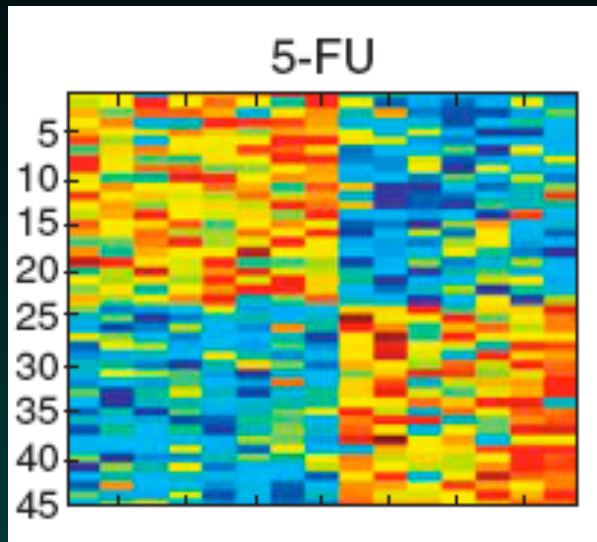


Our t-tests

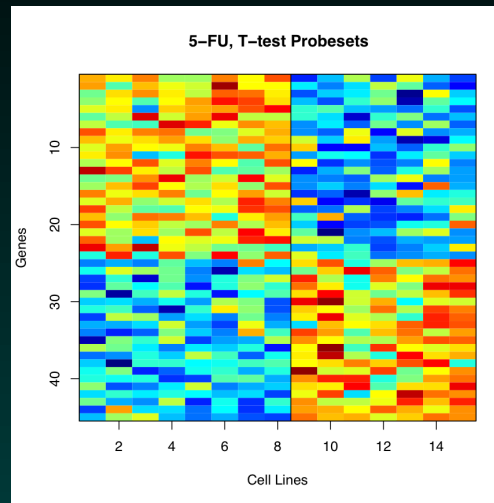


Reported Genes

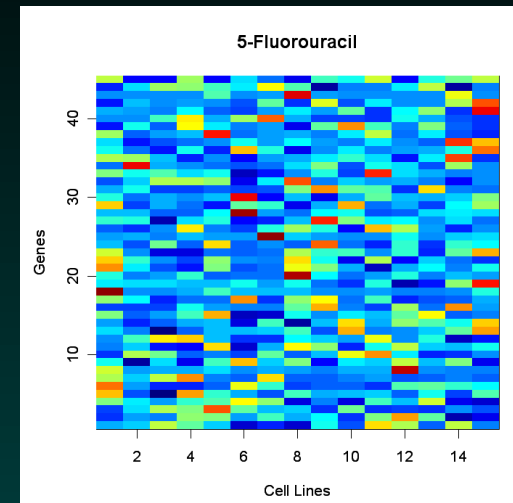
5-FU Heatmaps



Nat Med Paper



Our t-tests



Reported Genes

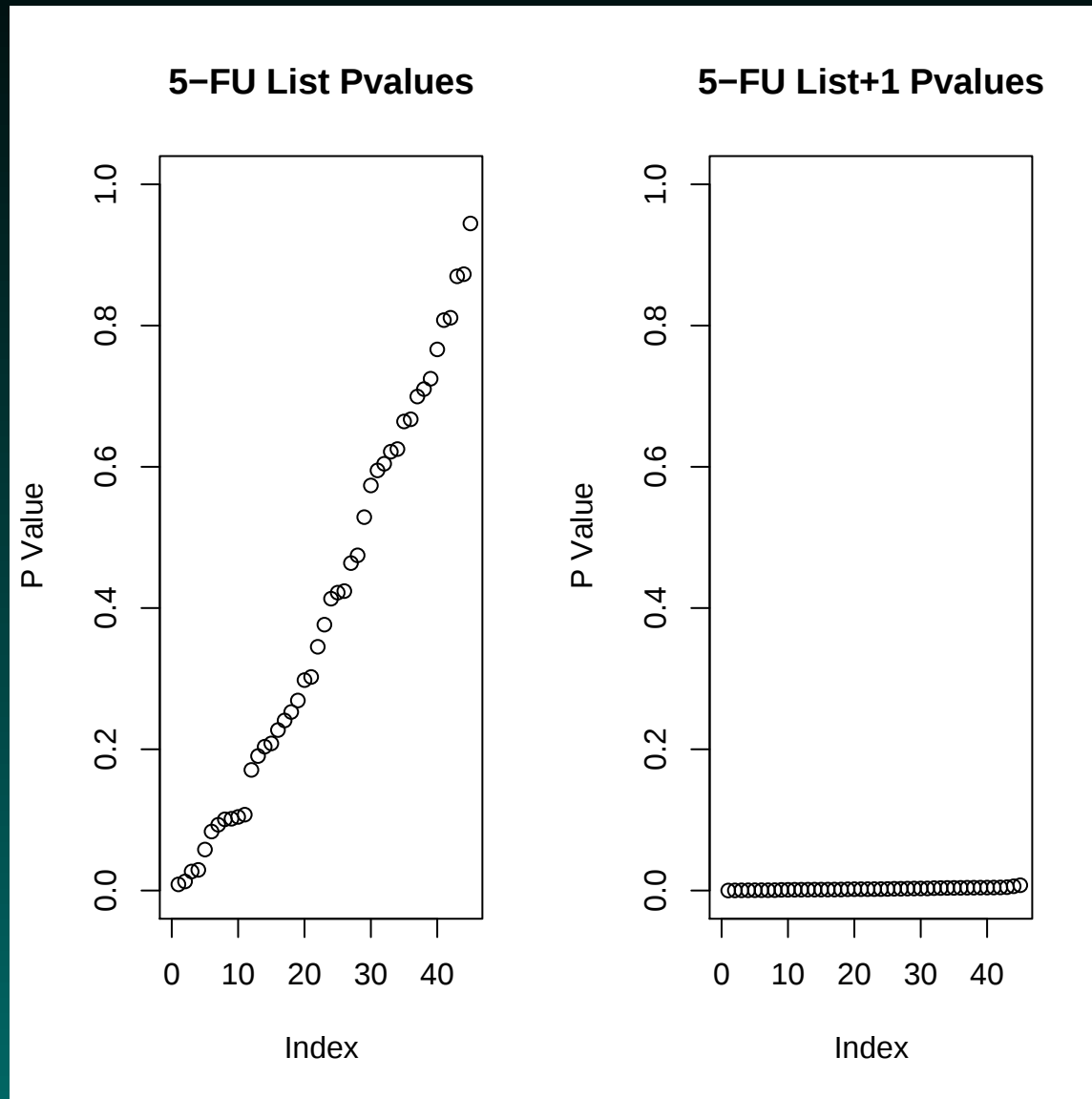
Something isn't quite right...

Their List and Ours

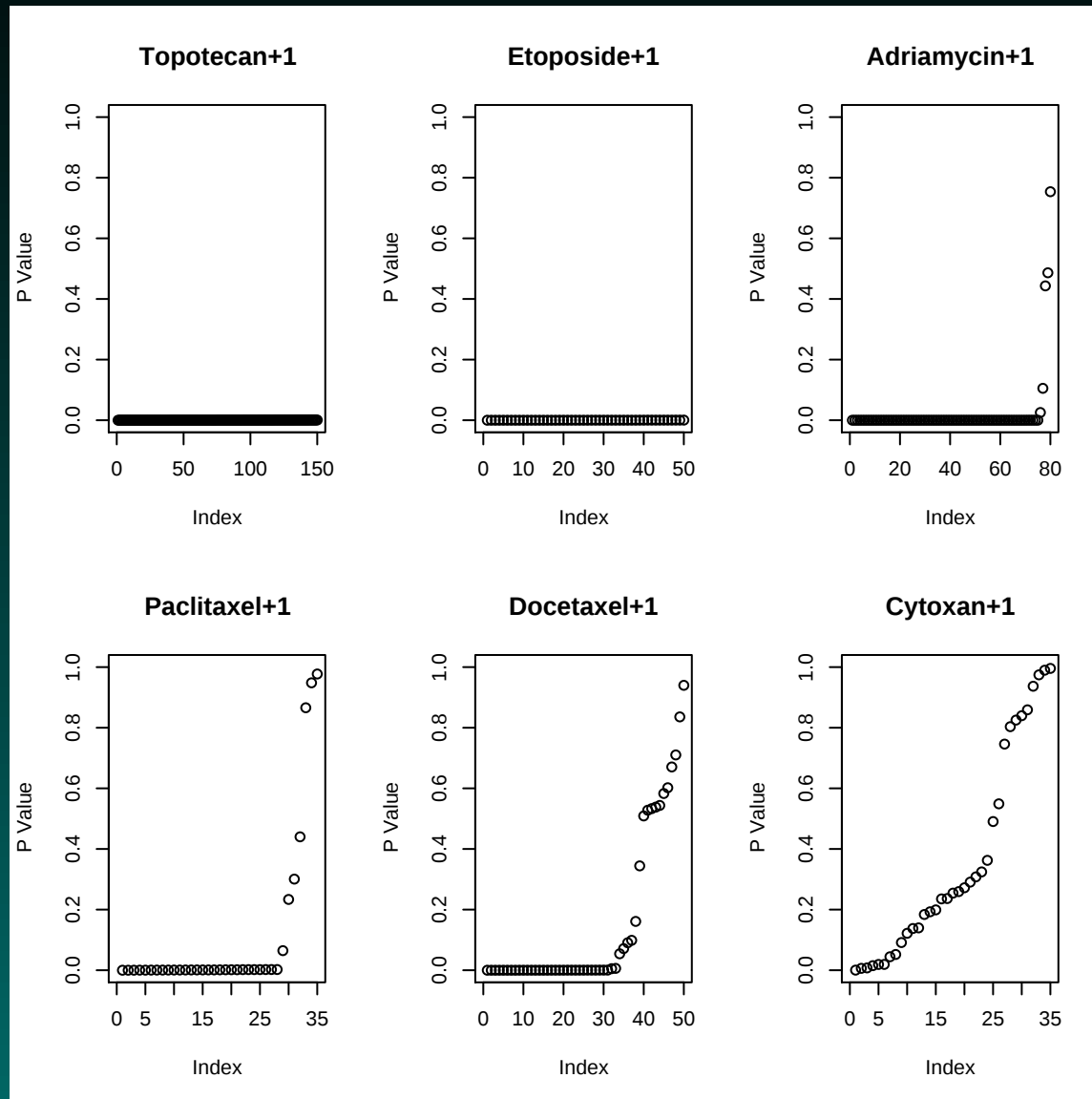
```
> temp <- cbind(  
  sort(rownames(pottiUpdated)[fuRows]),  
  sort(rownames(pottiUpdated)[  
    fuTQNorm@p.values <= fuCut]));  
> colnames(temp) <- c("Theirs", "Ours");  
> temp
```

	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"
	...	

Offset P-Values: 5FU



Offset P-Values: Other Drugs



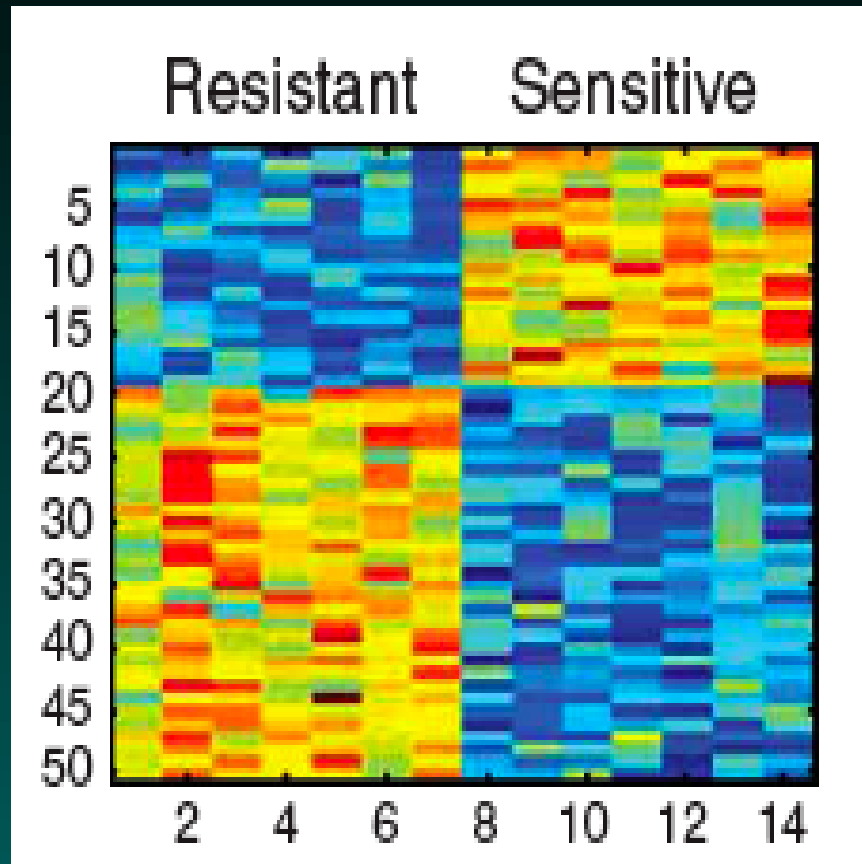
Using Their Software

Their software requires two files (and parameter values):

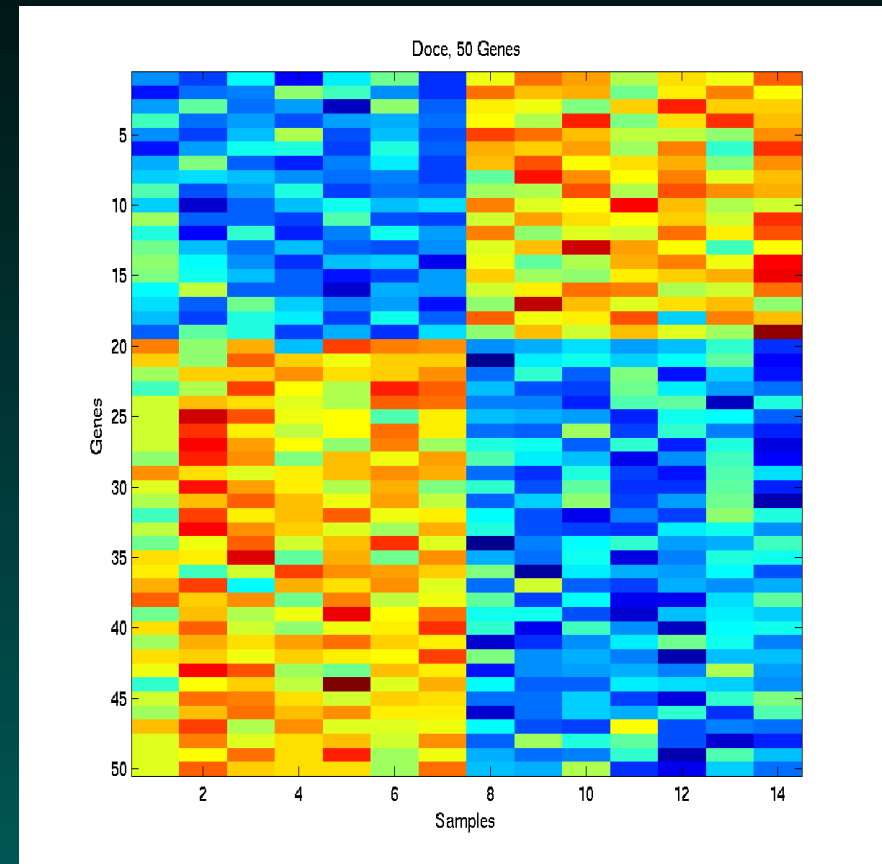
1. a quantification matrix, genes by samples, with a header row giving the classification (0 = Resistant, 1 = Sensitive, 2 = Test)
2. a list of probeset ids in the same order as the quantification matrix. *The list of probeset ids should not have a header row.*

What do we get?

Heatmaps Match Exactly for Docetaxel!



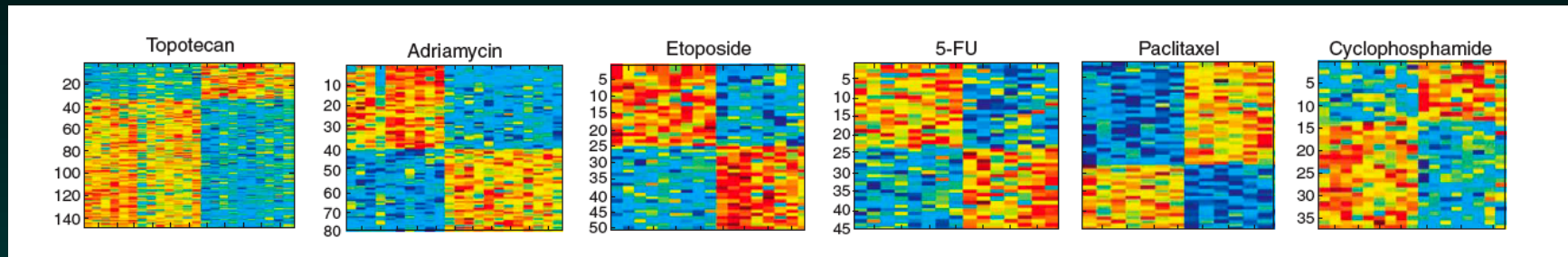
From Potti et al, Figure 1



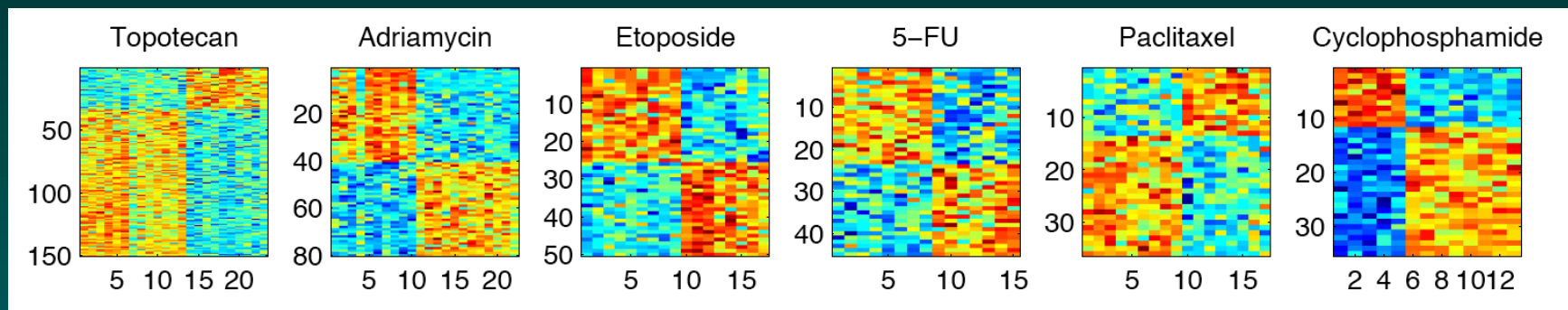
From the software

Heatmaps Match Exactly for 5 Others!

From the **paper**:

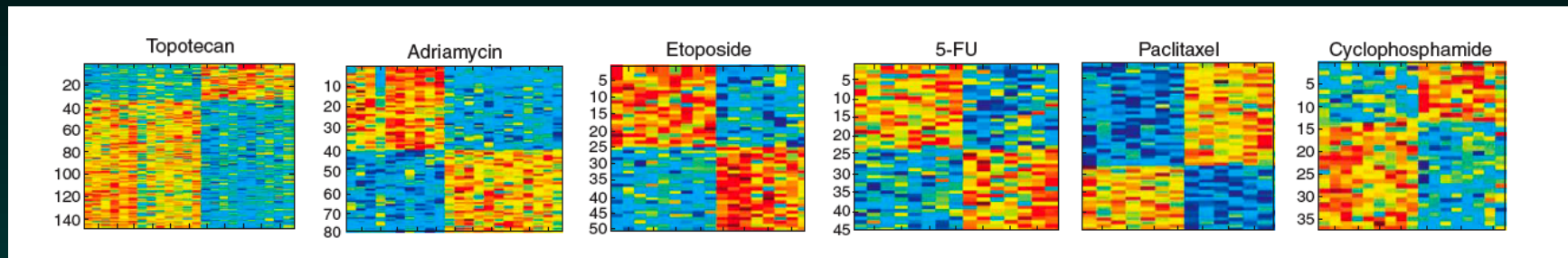


From the **software**:

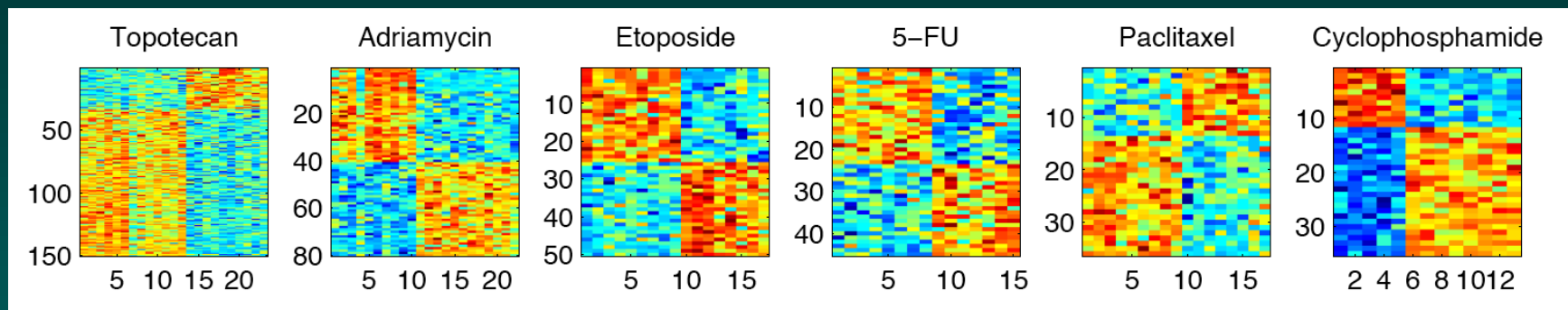


Heatmaps Match Exactly for 5 Others!

From the **paper**:

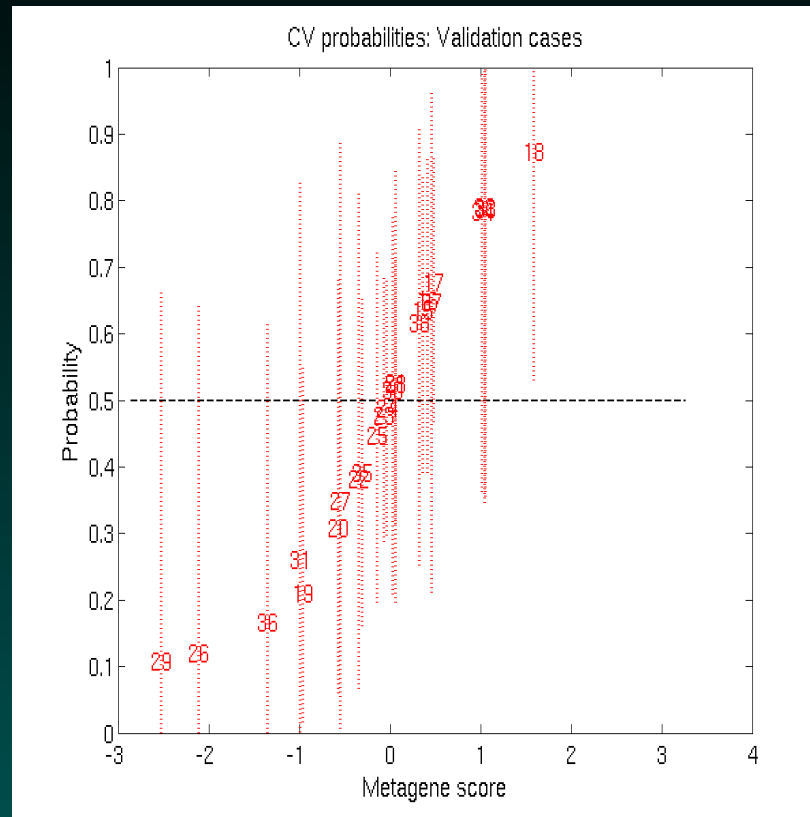


From the **software**:



We match heatmaps but not gene lists?

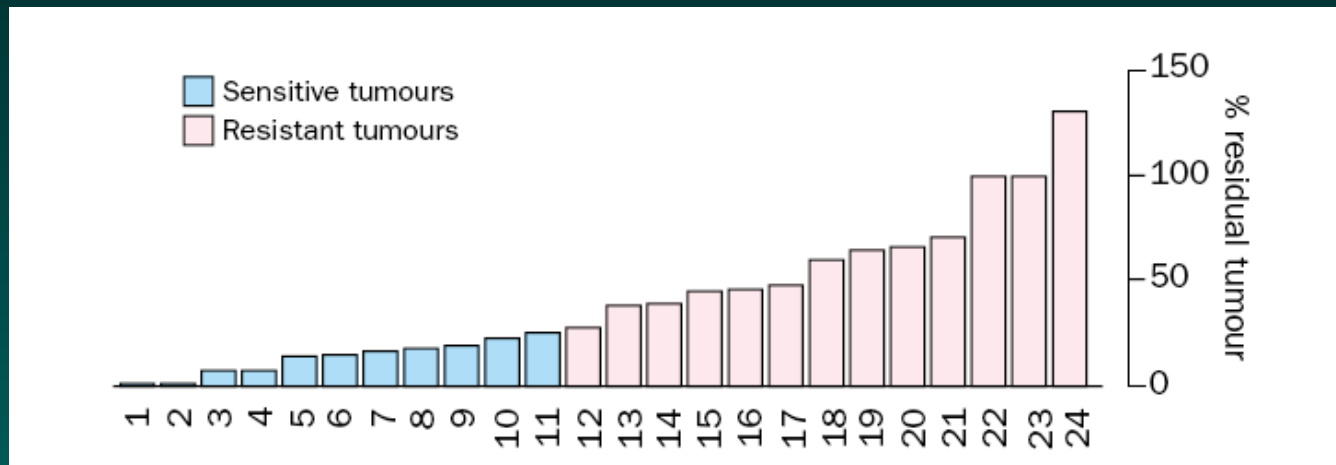
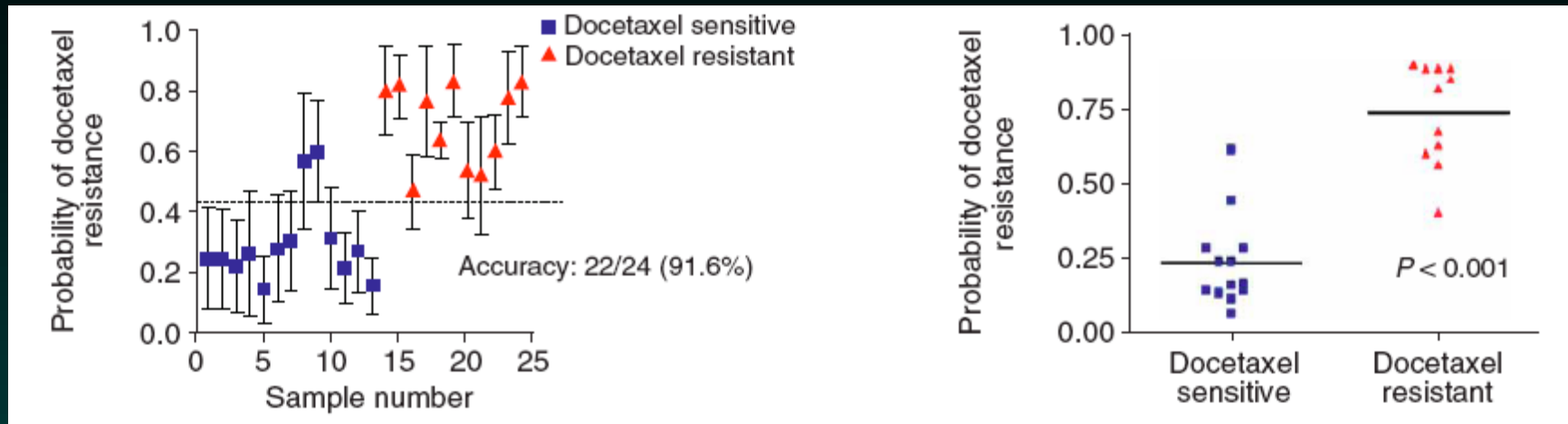
The Software Also Gives Predictions...



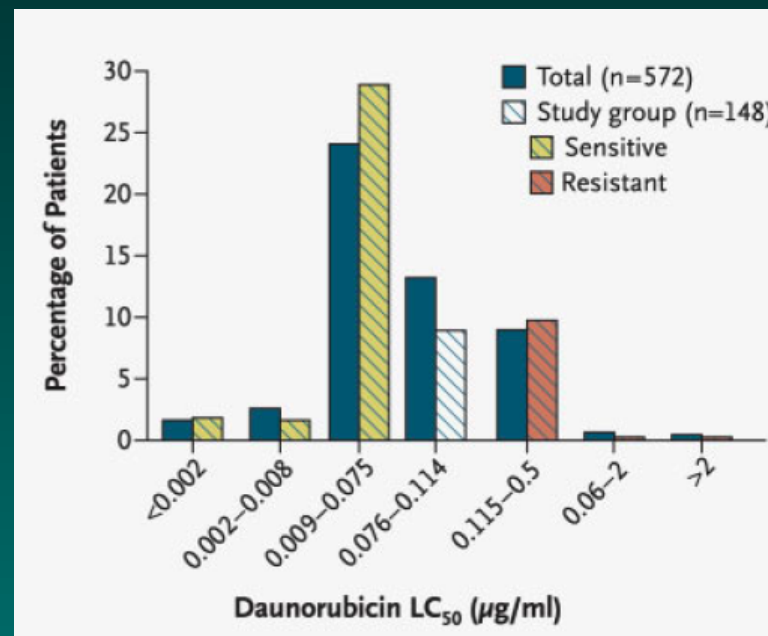
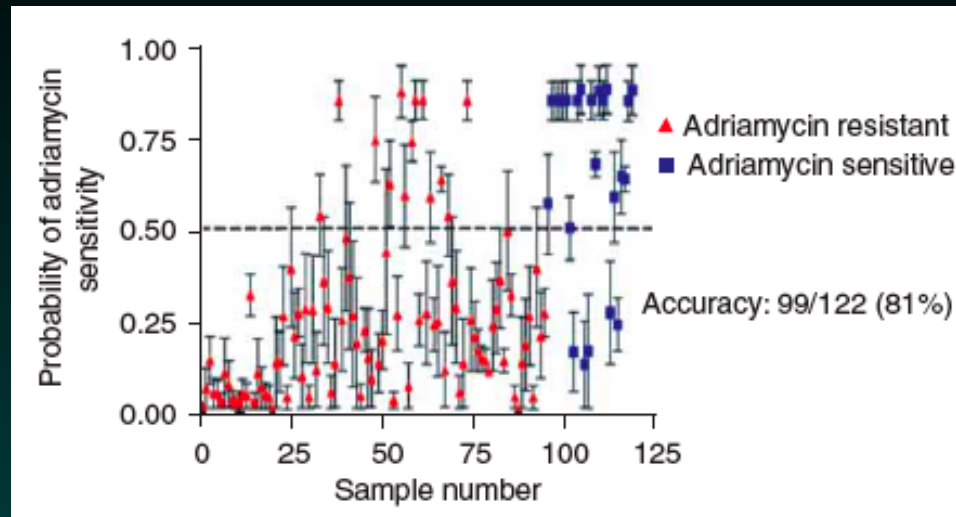
So, how good are the predictions?

How good are the ones they report?

Predicting Docetaxel (Chang 03)



Predicting Adriamycin (Holleman 04)



What's Going On?

One area for potential mixup is in **labeling samples** as “0” or “1” instead of “Sensitive” and “Resistant”.

Another is that the software does something odd in computing metagenes:

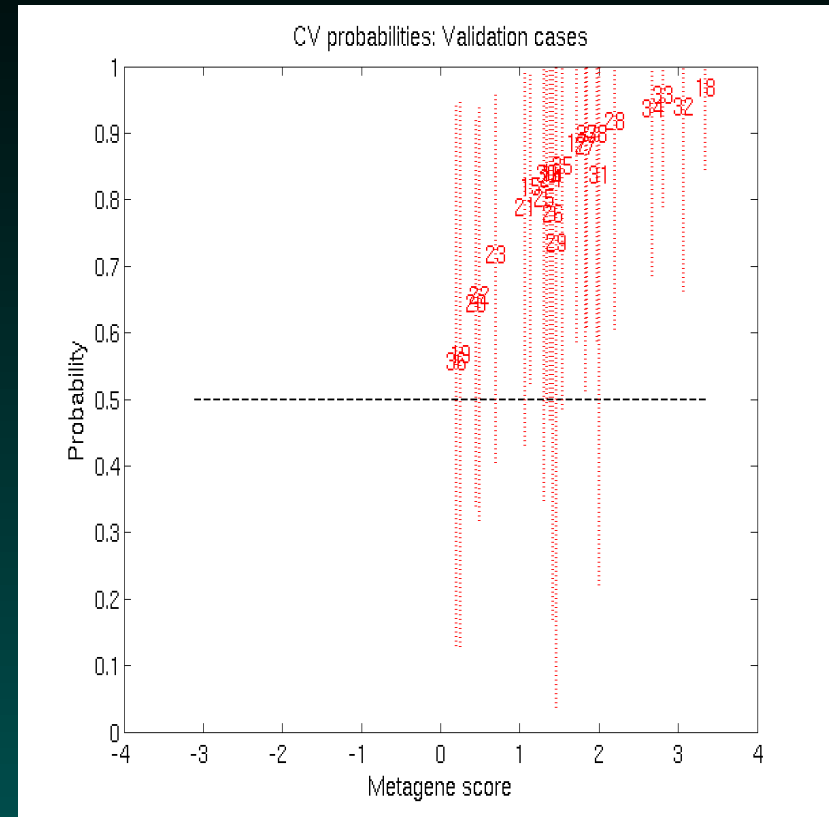
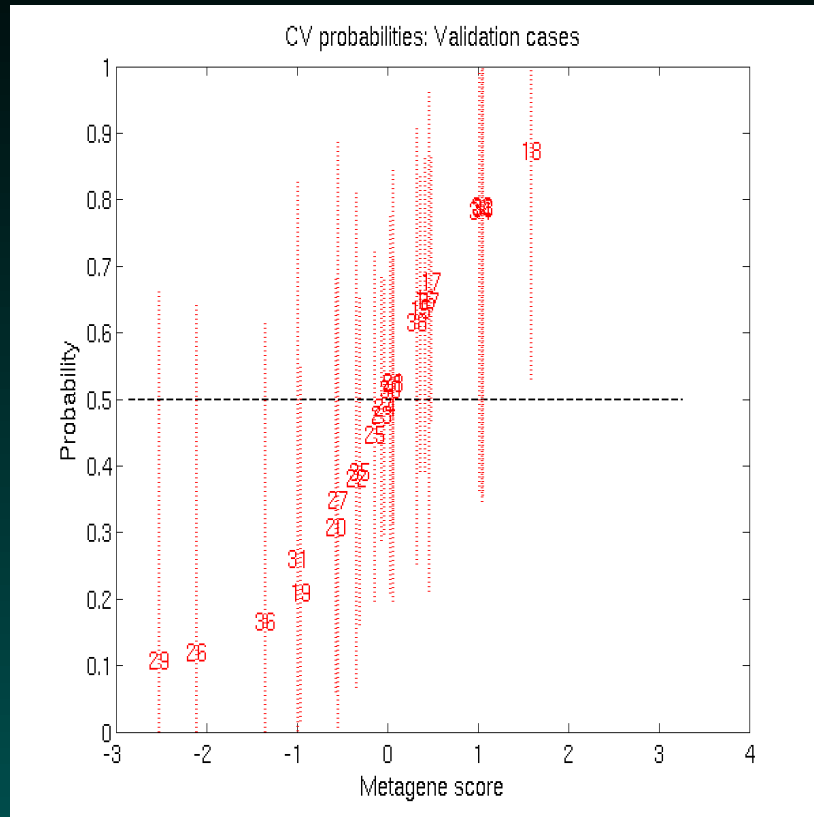
What's Going On?

One area for potential mixup is in **labeling samples as “0” or “1”** instead of “Sensitive” and “Resistant”.

Another is that the software does something odd in computing metagenes:

the metagenes are taken from an SVD applied to both training and test data.

Docetaxel Test Set Predictions



- Training & Test.

- Training Only.

The “combined” predictions are *different*. This doesn’t mean they’re *right*. Now, if we choose genes using only the training data, combining is mostly standardizing. But...

There Were Other Genes...

The 50-gene list for docetaxel has 19 “outliers”.

The initial paper on the test data (Chang et al) gave a list of 92 genes that separated responders from nonresponders.

There Were Other Genes...

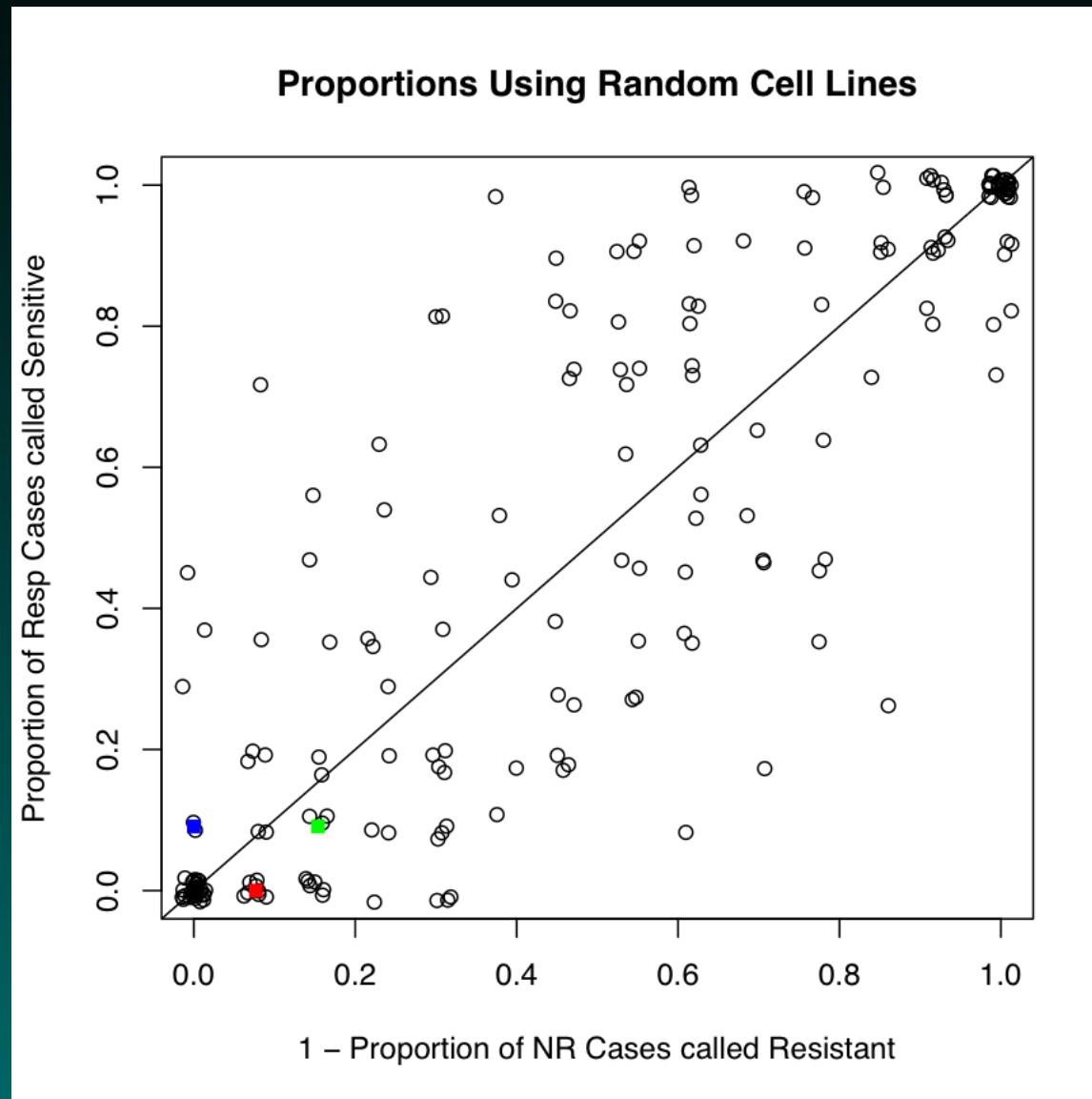
The 50-gene list for docetaxel has 19 “outliers”.

The initial paper on the test data (Chang et al) gave a list of 92 genes that separated responders from nonresponders.

Entries 7-20 in the Chang et al list (a contiguous bloc) comprise 14 of the 19 outliers.

The other 5 are ERCC1, ERCC4, ERBB2, BCL2L11, and TUBA3. These are the genes named to explain the biology.

Predictions With Random Lines



So, Do We Think it Works?

Pause here for dramatic tension...

So, Do We Think it Works?

Pause here 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 responsiveness.
- the training and test sets were run at different times and under different conditions with different array platforms and definitions of sensitivity.

We think it *appears* to work due to poor bookkeeping and documentation.

Don't Take My Word For It!

Read the paper! Coombes, Wang & Baggerly, Nat Med, Nov 6, 2007, 13:1276-7, author reply 1277-8.

Try it yourselves! All of the raw data, documentation, and code is available from our web site:

<http://bioinformatics.mdanderson.org/Supplements/ReproRsch-Chemo>; All of our code and documentation is likewise available from Nature Medicine.

Potti/Nevins Reply (Nat Med 13:1277-8)

Labels for Adria are correct – details on their web page. (*)

They've gotten the approach to work again. (Twice!)*

Pharmacogenomic Strategies Provide a Rational Approach to the Treatment of Cisplatin-Resistant Patients With Advanced Cancer

David S. Hsu, Bala S. Balakumaran, Chaitanya R. Acharya, Vanja Vlahovic, Kelli S. Walters, Katherine Garman, Carey Anders, Richard F. Riedel, Johnathan Lancaster, David Harpole, Holly K. Dressman, Joseph R. Nevins, Phillip G. Febbo, and Anil Potti

Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial

Hervé Bonnefoi, Anil Potti, Mauro Delorenzi, Louis Mauriac, Mario Campone, Michèle Tubiana-Hulin, Thierry Petit, Philippe Rouanet, Jacek Jassem, Emmanuel Blot, Véronique Becette, Pierre Farmer, Sylvie André, Chaitanya R Acharya, Sayan Mukherjee, David Cameron, Jonas Bergh, Joseph R Nevins, Richard D Iggo

...and Even More...

An Integrated Genomic-Based Approach to Individualized Treatment of Patients With Advanced-Stage Ovarian Cancer

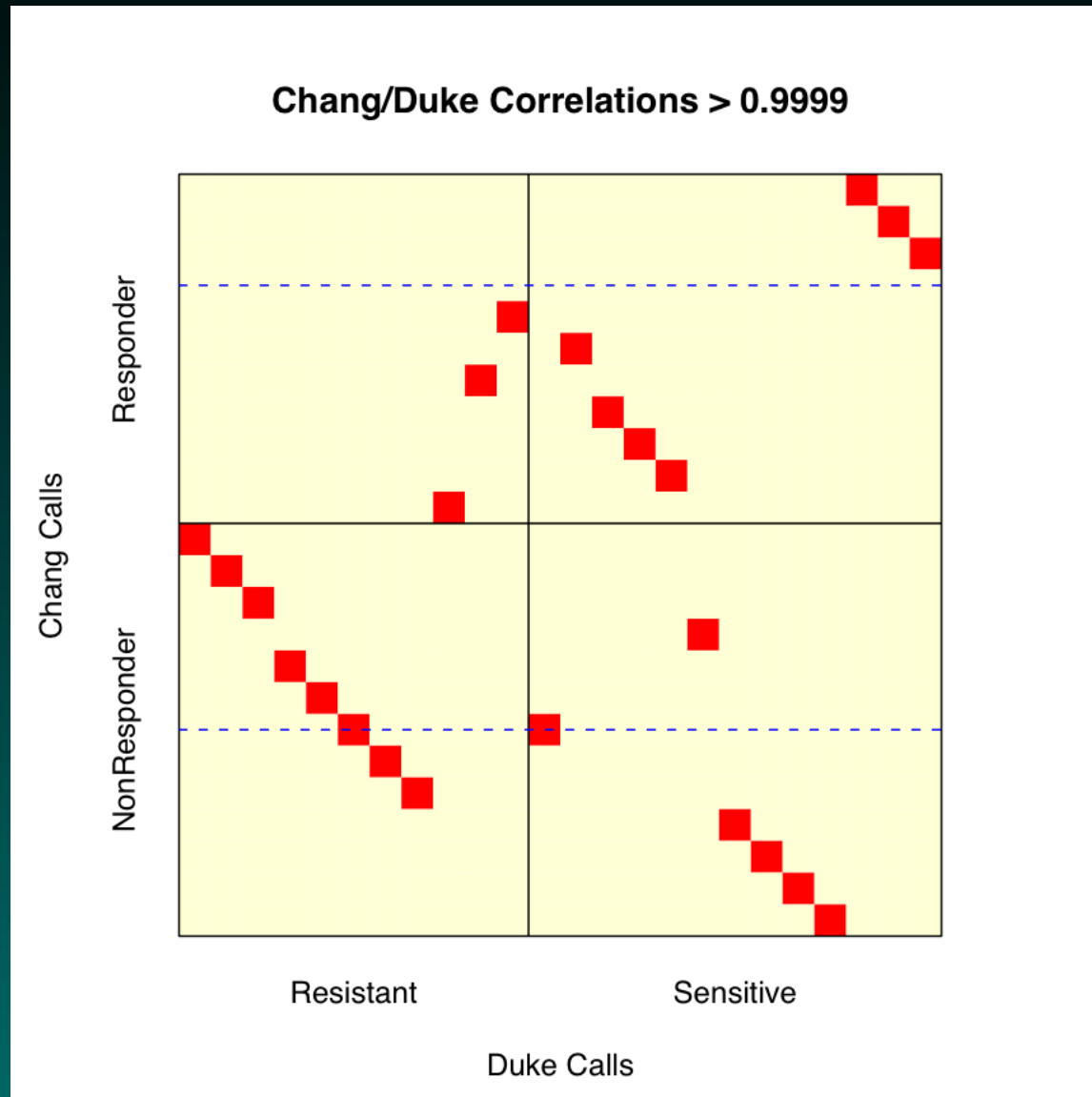
Holly K. Dressman, Andrew Berchuck, Gina Chan, Jun Zhai, Andrea Bild, Robyn Sayer, Janiel Cragun, Jennifer Clarke, Regina S. Whitaker, LiHua Li, Jonathan Gray, Jeffrey Marks, Geoffrey S. Ginsburg, Anil Potti, Mike West, Joseph R. Nevins, and Johnathan M. Lancaster

Genomic and Molecular Profiling Predicts Response to Temozolomide in Melanoma

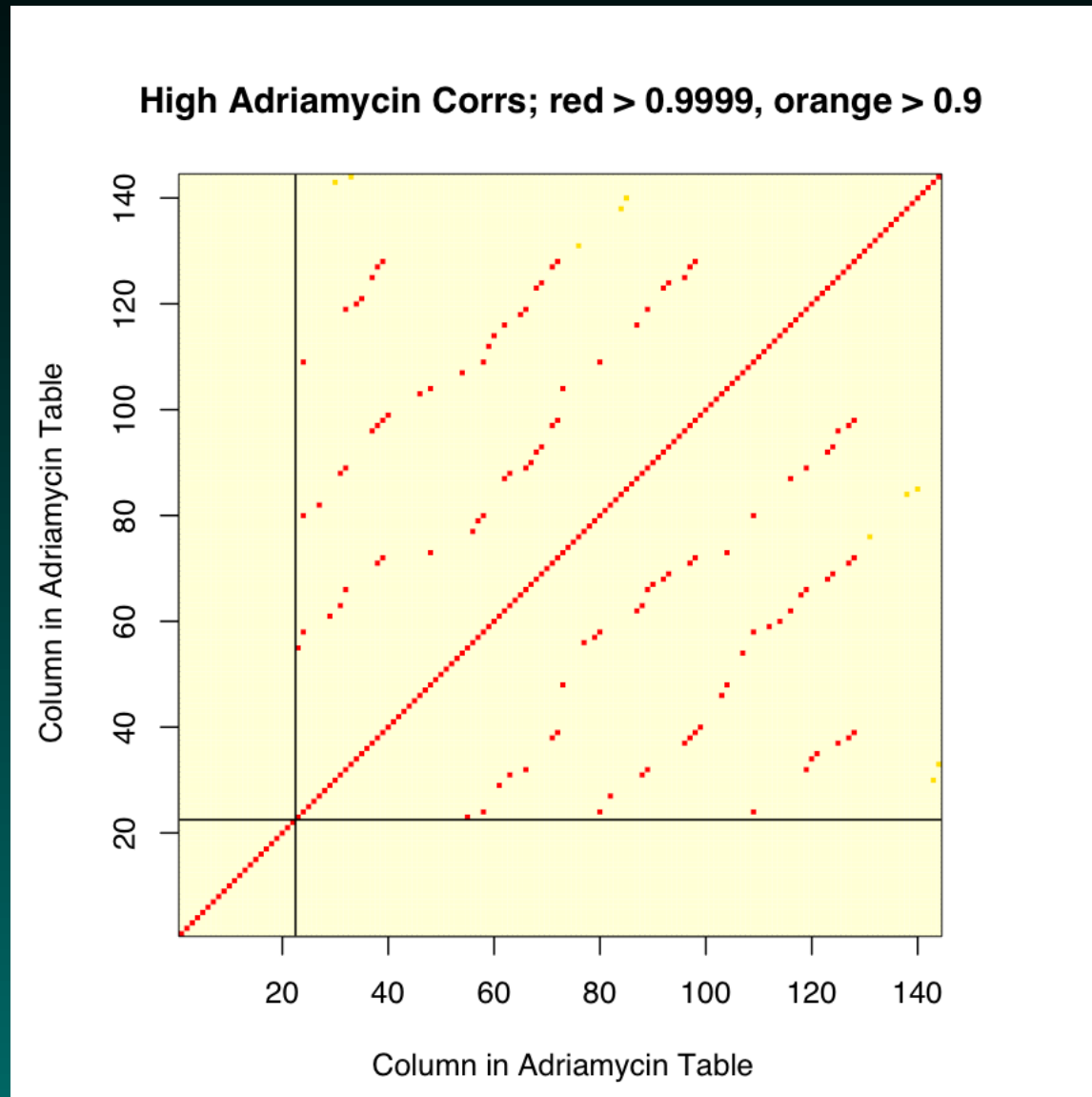
Christina K. Augustine,^{1,5} Jin Soo Yoo,¹ Anil Potti,^{2,4} Yasunori Yoshimoto,^{1,5} Patricia A. Zipfel,^{1,5} Henry S. Friedman,¹ Joseph R. Nevins,^{3,4} Francis Ali-Osman,¹ and Douglas S. Tyler^{1,5}

Is everybody ready?

Mapping Potti to Chang (Reply)



Adriamycin 0.9999+ Correlations (Reply)



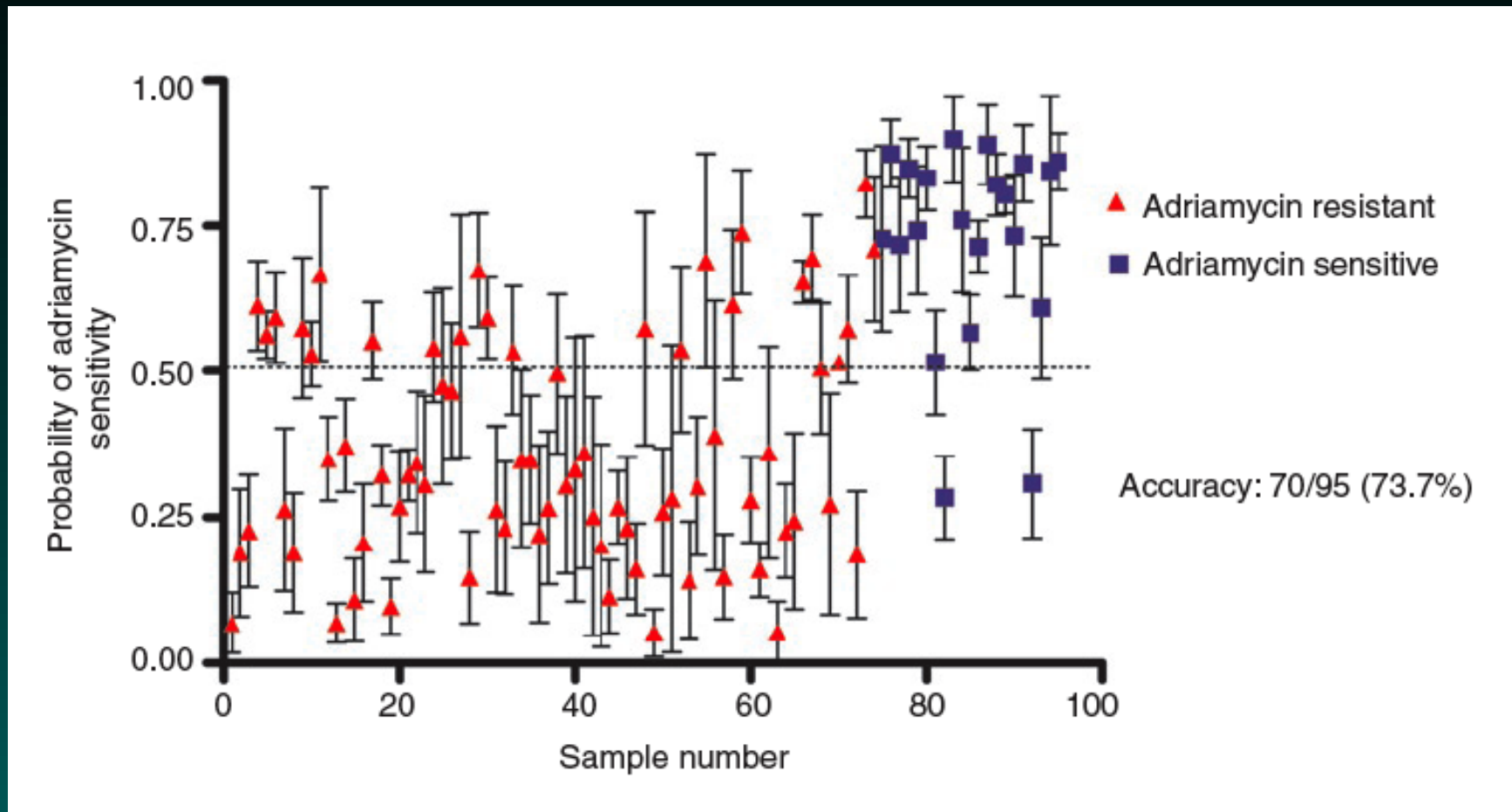
Nat Med, Take 2

- Breast data (n = 24).txt (gone)
- Adria_ALL (n = 122).txt (replaced with Adria_ALL_data1_n95.doc)

“In the version ... initially published ... 27 samples were replicated... The authors have reanalyzed ... using only the 95 unique samples...”

“the authors have added two more accession numbers (GSE2351 and GSE649)”

The New Adria Plot



74 resistant, 21 sensitive

The First 20 Files Now Named

Sample ID Response

1	GSM44303	RES	11	GSM9694	RES
2	GSM44304	RES	12	GSM9695	RES
3	GSM9653	RES	13	GSM9696	RES
4	GSM9653	RES	14	GSM9698	RES
5	GSM9654	RES	15	GSM9699	SEN
6	GSM9655	RES	16	GSM9701	RES
7	GSM9656	RES	17	GSM9708	RES
8	GSM9657	RES	18	GSM9708	SEN
9	GSM9658	SEN	19	GSM9709	RES
10	GSM9658	SEN	20	GSM9711	RES

The First 20 Files Now Named

Sample ID		Response	
1	GSM44303	RES	
2	GSM44304	RES	
3	GSM9653	RES	
4	GSM9653	RES	
5	GSM9654	RES	
6	GSM9655	RES	
7	GSM9656	RES	
8	GSM9657	RES	
9	GSM9658	SEN	
10	GSM9658	SEN	
11	GSM9694	RES	
12	GSM9695	RES	
13	GSM9696	RES	
14	GSM9698	RES	
15	GSM9699	SEN	
16	GSM9701	RES	
17	GSM9708	RES	
18	GSM9708	SEN	
19	GSM9709	RES	
20	GSM9711	RES	

15 duplicates; 6 are inconsistent.

Summary (Reply)

Looking at the 80 distinct GSM ids, Potti et al class

61 Resistant,

13 Sensitive,

6 Both Ways.

Using LC50 values with specified cutoffs, Holleman et al class

22 Resistant,

48 Sensitive,

10 Intermediate.

One new GEO dataset listed as a source (GSE649) flags samples that are resistant to vincristine, not adriamycin.

Validation 1: Hsu et al

Pharmacogenomic Strategies Provide a Rational Approach to the Treatment of Cisplatin-Resistant Patients With Advanced Cancer

David S. Hsu, Bala S. Balakumaran, Chaitanya R. Acharya, Vanja Vlahovic, Kelli S. Walters, Katherine Garman, Carey Anders, Richard F. Riedel, Johnathan Lancaster, David Harpole, Holly K. Dressman, Joseph R. Nevins, Phillip G. Febbo, and Anil Potti

J Clin Oncol, Oct 1, 2007, 25:4350-7.

Same approach, using **Cisplatin** and **Pemetrexed**.

For cisplatin, U133A arrays were used for the training set, and **ERCC1**, **ERCC4** and **DNA repair** genes are identified as being “important”.

With some work, we matched the heatmaps.

What About the Gene Lists?

The list for **pemetrexed** was generated from training data where the **sensitive and resistant labels were reversed**.

But we can match all 85 of the genes for pemetrexed!

What About the Gene Lists?

The list for **pemetrexed** was generated from training data where the **sensitive and resistant labels were reversed**.

But we can match all 85 of the genes for pemetrexed!

After we account for the off-by-one error.

What About the Gene Lists?

The list for **pemetrexed** was generated from training data where the **sensitive and resistant labels were reversed**.

But we can match all 85 of the genes for pemetrexed!

After we account for the off-by-one error.

We can match 41 of the 45 genes for cisplatin, **after we account for the off-by-one error.**

The 4 We Can't Match (Reply)

203719_at, ERCC1,
210158_at, ERCC4,
228131_at, ERCC1, and
231971_at, FANCM (DNA Repair).

ERCC1 and/or ERCC4 were outliers in the earlier gene lists for Docetaxel, Paclitaxel, and Adriamycin. We find their frequent recurrence disturbing. Even so, the last two here are special.

The 4 We Can't Match (Reply)

203719_at, ERCC1,
210158_at, ERCC4,
228131_at, ERCC1, and
231971_at, FANCM (DNA Repair).

ERCC1 and/or ERCC4 were outliers in the earlier gene lists for Docetaxel, Paclitaxel, and Adriamycin. We find their frequent recurrence disturbing. Even so, the last two here are special.

*These probesets aren't on the U133A arrays that were used.
They're on the U133B.*

Validation 2: Bonnefoi et al

Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial

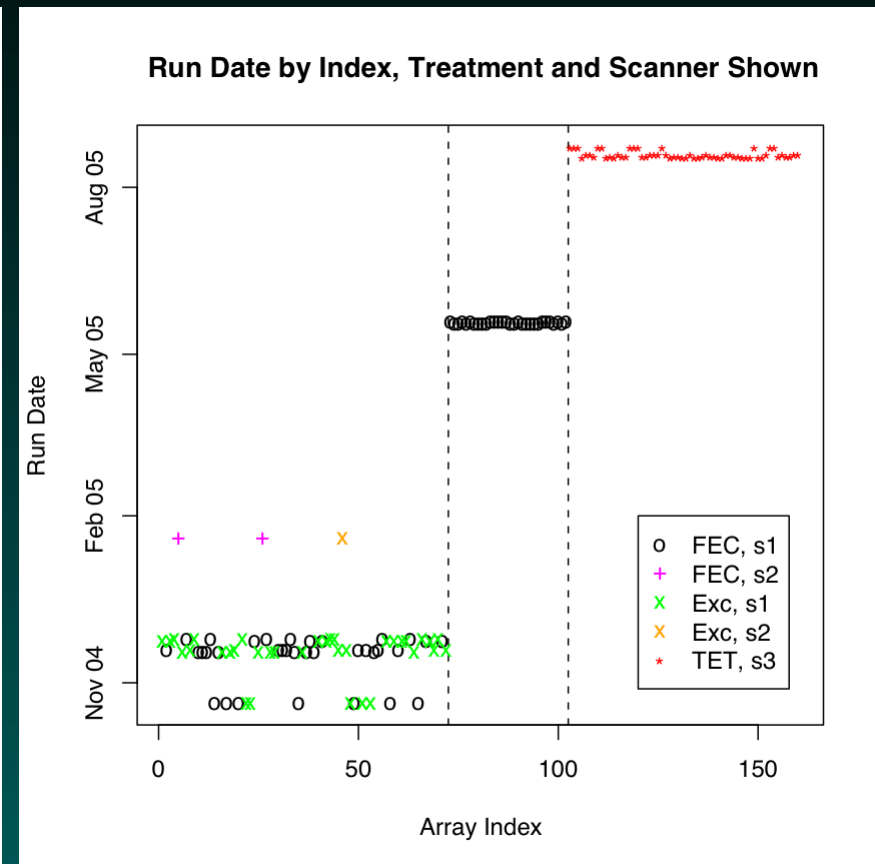
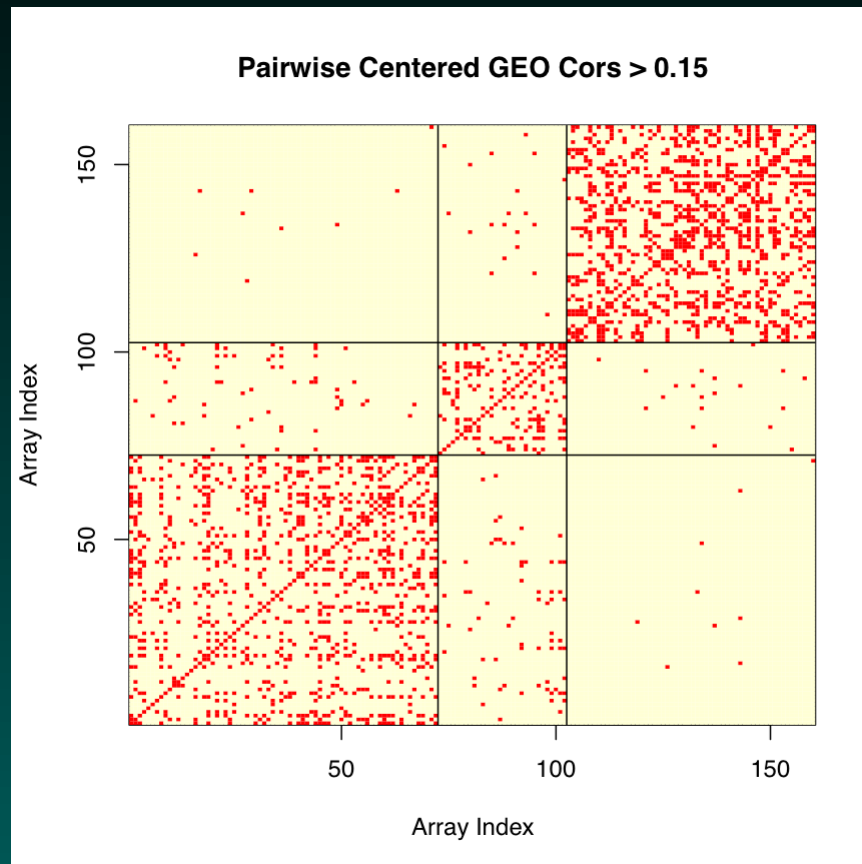
Hervé Bonnefoi, Anil Potti, Mauro Delorenzi, Louis Mauriac, Mario Camponé, Michèle Tubiana-Hulin, Thierry Petit, Philippe Rouanet, Jacek Jassem, Emmanuel Blot, Véronique Becette, Pierre Farmer, Sylvie André, Chaitanya R Acharya, Sayan Mukherjee, David Cameron, Jonas Bergh, Joseph R Nevins, Richard D Iggo

Lancet Oncology, Dec 2007, 8:1071-8. (early access Nov 14)

Similar approach, using signatures for Fluorouracil, Epirubicin (used Adriamycin), Cyclophosphamide, and Taxotere (Docetaxel) to predict response to one of two combination therapies: **FEC** and **TET**.

Potentially improves ER- response from 44% to 70%.

We Might Expect Some Differences...



High Sample Correlations
after Centering by Gene

Array Run Dates

What Cell Lines Were Involved?

Supplementary “webpanel” from Lancet Oncology.

Webpanel: cell lines included in the training set	(Continued from previous column)
Docetaxel	Cyclophosphamide
<i>Sensitive cell lines</i>	<i>Sensitive cell lines</i>
EKVX	K-562
IGROV1	MOLT-4
OVCAR-4	HL-60(TB)
786-0	MCF7
CAKI-1	HCC-2998
SN12C	HCT-116
TK-10	NCI-H460
<i>Resistant cell lines</i>	TK-10
HL-60(TB)	<i>Resistant cell lines</i>
SF-539	SNB-19
HT29	HS 578T
HOP-62	MDA-MB-231/ATCC
SK-MEL-2	MDA-MB-435
SK-MEL-5	NCI-H226
NCI-H522	M14
Doxorubicin	MALME-3M
<i>Sensitive cell lines</i>	SK-MEL-2

The sensitive and resistant categories for docetaxel are reversed relative to Potti et al.

How Are Results Combined?

Potti et al predicted response to TFAC. Bonnefoi et al TET and FEC. Let $P()$ indicate prob sensitive. The rules used are as follows.

How Are Results Combined?

Potti et al predicted response to TFAC. Bonnefoi et al TET and FEC. Let $P()$ indicate prob sensitive. The rules used are as follows.

$$P(TFAC) = P(T) + P(F) + P(A) + P(C) - P(T)P(F)P(A)P(C).$$

How Are Results Combined?

Potti et al predicted response to TFAC. Bonnefoi et al TET and FEC. Let $P()$ indicate prob sensitive. The rules used are as follows.

$$P(TFAC) = P(T) + P(F) + P(A) + P(C) - P(T)P(F)P(A)P(C).$$

$$P(ET) = \max[P(E), P(T)].$$

How Are Results Combined?

Potti et al predicted response to TFAC. Bonnefoi et al TET and FEC. Let $P()$ indicate prob sensitive. The rules used are as follows.

$$P(TFAC) = P(T) + P(F) + P(A) + P(C) - P(T)P(F)P(A)P(C).$$

$$P(ET) = \max[P(E), P(T)].$$

$$P(FEC) = \frac{5}{8}[P(F) + P(E) + P(C)] - \frac{1}{4}.$$

How Are Results Combined?

Potti et al predicted response to TFAC. Bonnefoi et al TET and FEC. Let $P()$ indicate prob sensitive. The rules used are as follows.

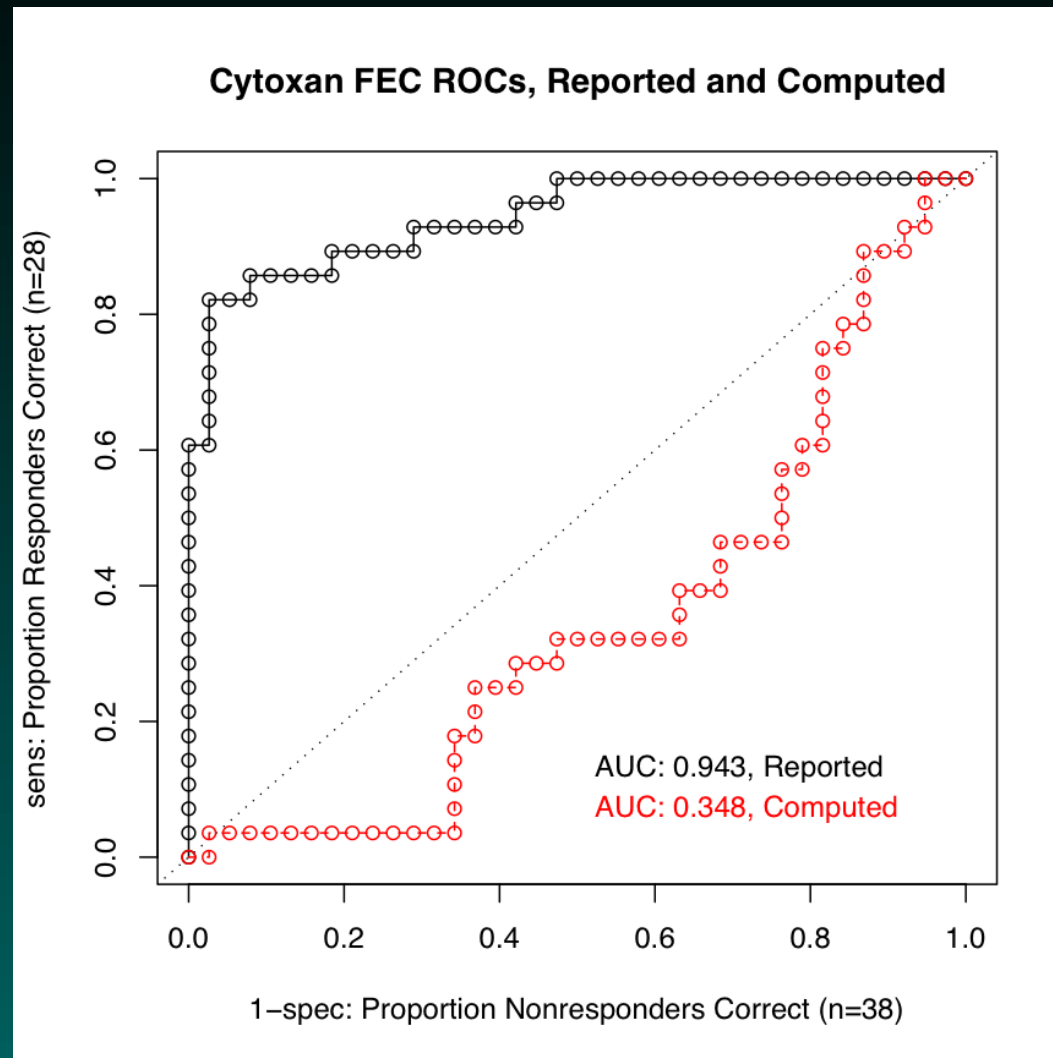
$$P(TFAC) = P(T) + P(F) + P(A) + P(C) - P(T)P(F)P(A)P(C).$$

$$P(ET) = \max[P(E), P(T)].$$

$$P(FEC) = \frac{5}{8}[P(F) + P(E) + P(C)] - \frac{1}{4}.$$

Each rule is different.

Predictions for Individual Drugs? (Reply)



Does cytoxan make sense?

Ovarian Cancer and Pathways

An Integrated Genomic-Based Approach to Individualized Treatment of Patients With Advanced-Stage Ovarian Cancer

Holly K. Dressman, Andrew Berchuck, Gina Chan, Jun Zhai, Andrea Bild, Robyn Sayer, Janiel Cragun, Jennifer Clarke, Regina S. Whitaker, LiHua Li, Jonathan Gray, Jeffrey Marks, Geoffrey S. Ginsburg, Anil Potti, Mike West, Joseph R. Nevins, and Johnathan M. Lancaster

Dressman et al, JCO, Feb 10, 2007.

Looking for pathway deregulation in ovarian cancer.

Using tumor array profiles to predict response to cisplatin.

119 serous tumors, quantifications, CEL files, and clinical information provided.

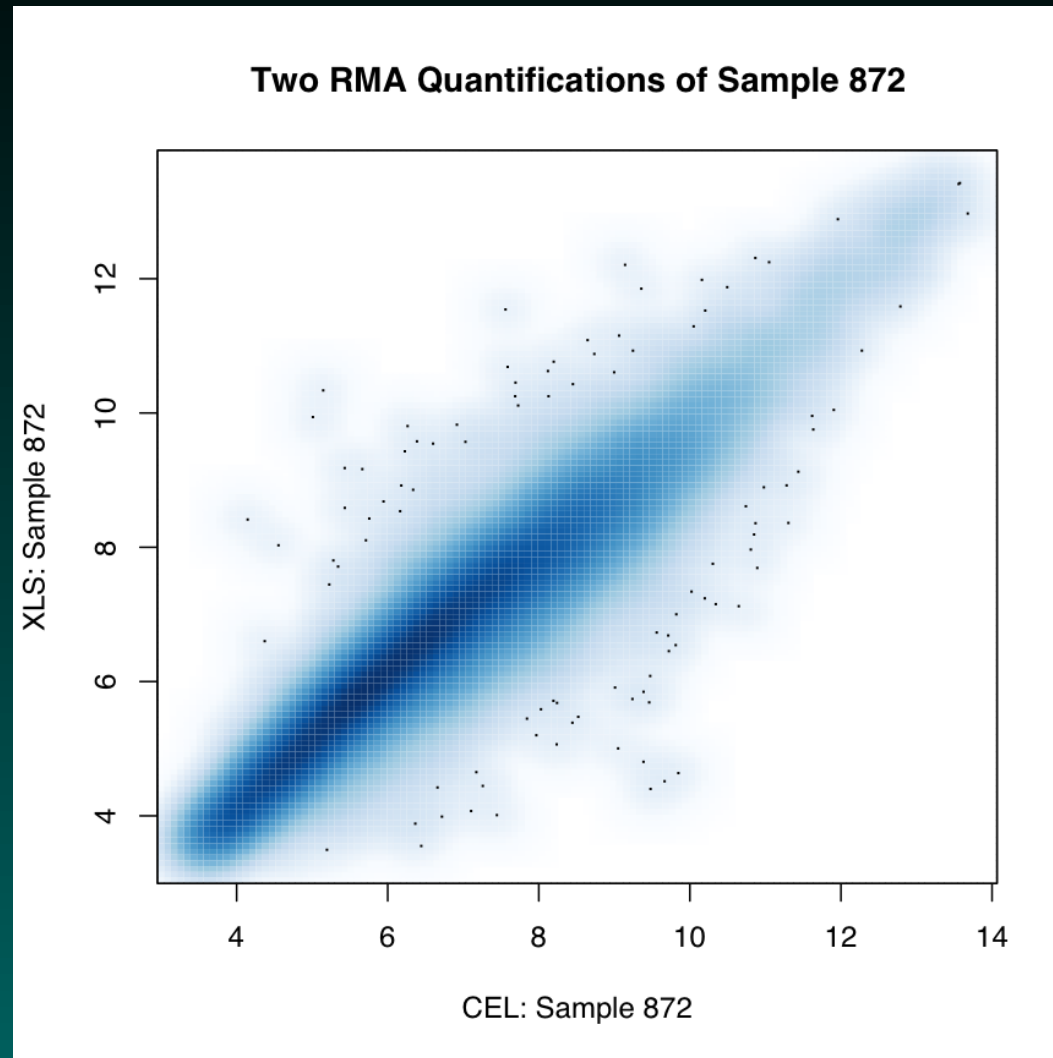
Looking at the Data

We began by looking at the RMA quantifications that they posted for the various arrays.

For each array, expression values were recorded for 22115 probesets. This is a strange number. There are 22283 total probesets on Affy U133A arrays, of which 68 are “controls” that are not often used in signatures. But $22283 - 68 = 22215$.

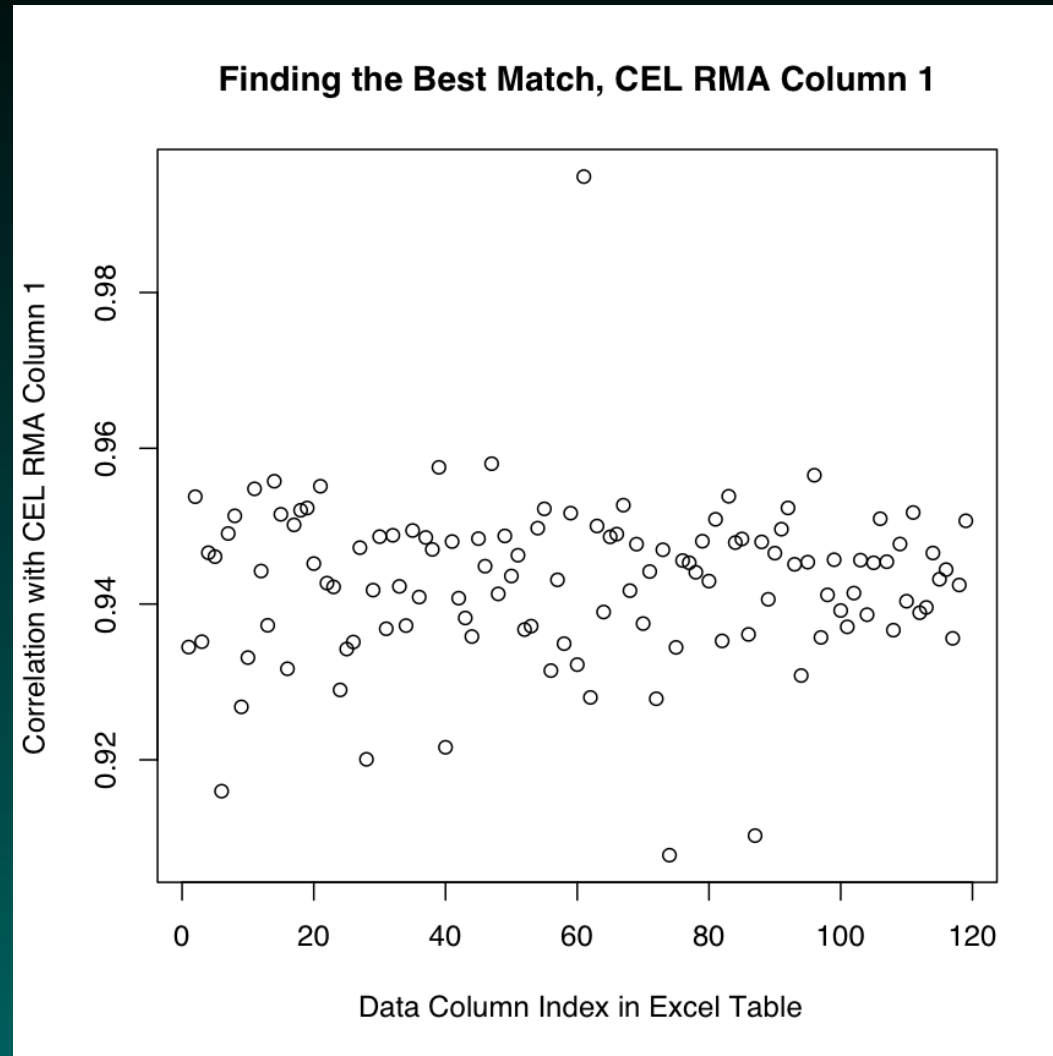
But, they used justRMA, so we could quantify the CEL files ourselves...

Checking Agreement



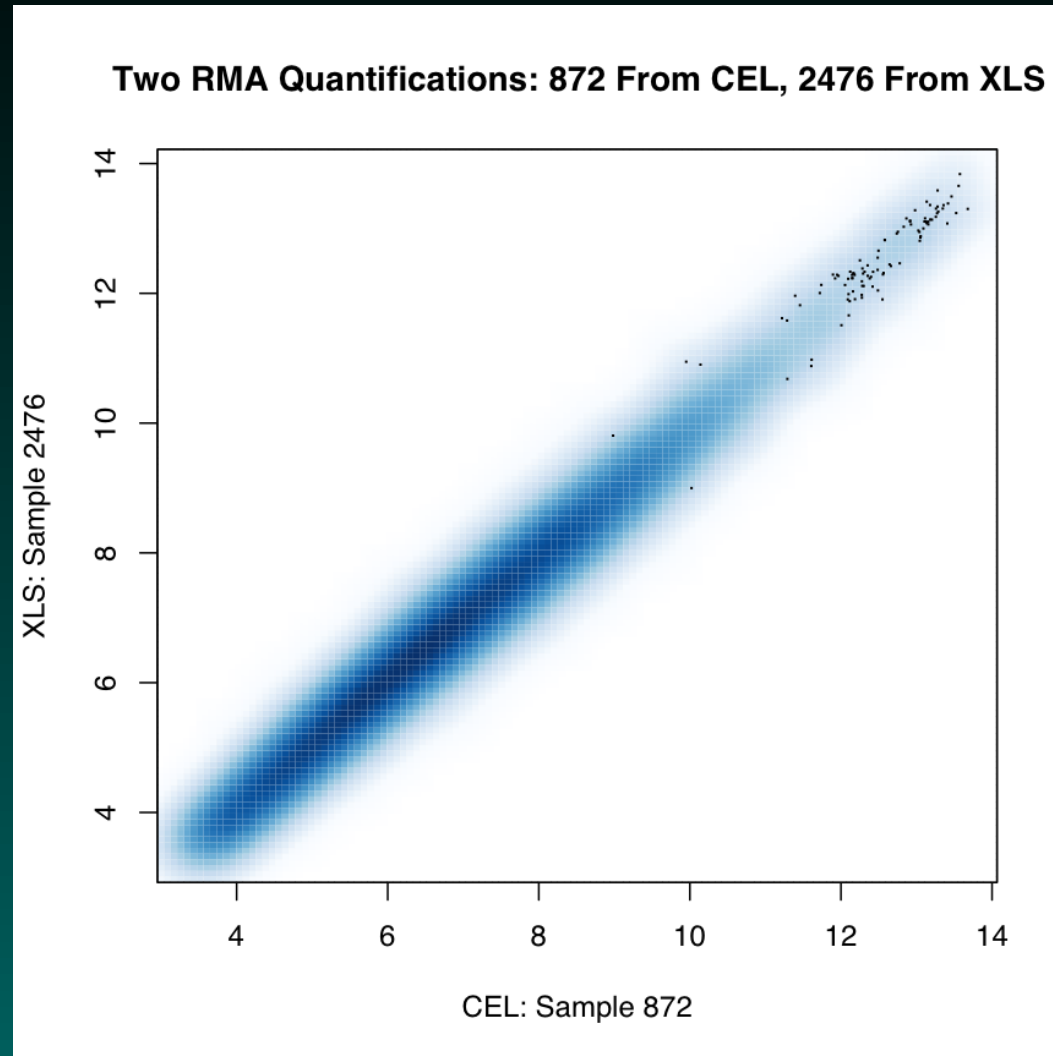
CELs vs Tables. We expected better (fewer outliers).

Looking at Their Other Quants



Which one would you pick?

Looking at The “Best” Fit



Same array. *Different* names (2476 from XLS, 872 from CEL).

How Bad is It?

The names match for **32/119** samples. For all but 3 of the others, we get very good correlations but a mismatch in names.

We don't have a clear “winner” for their quantifications for D1837, M4161, or M444.

More Raw Data

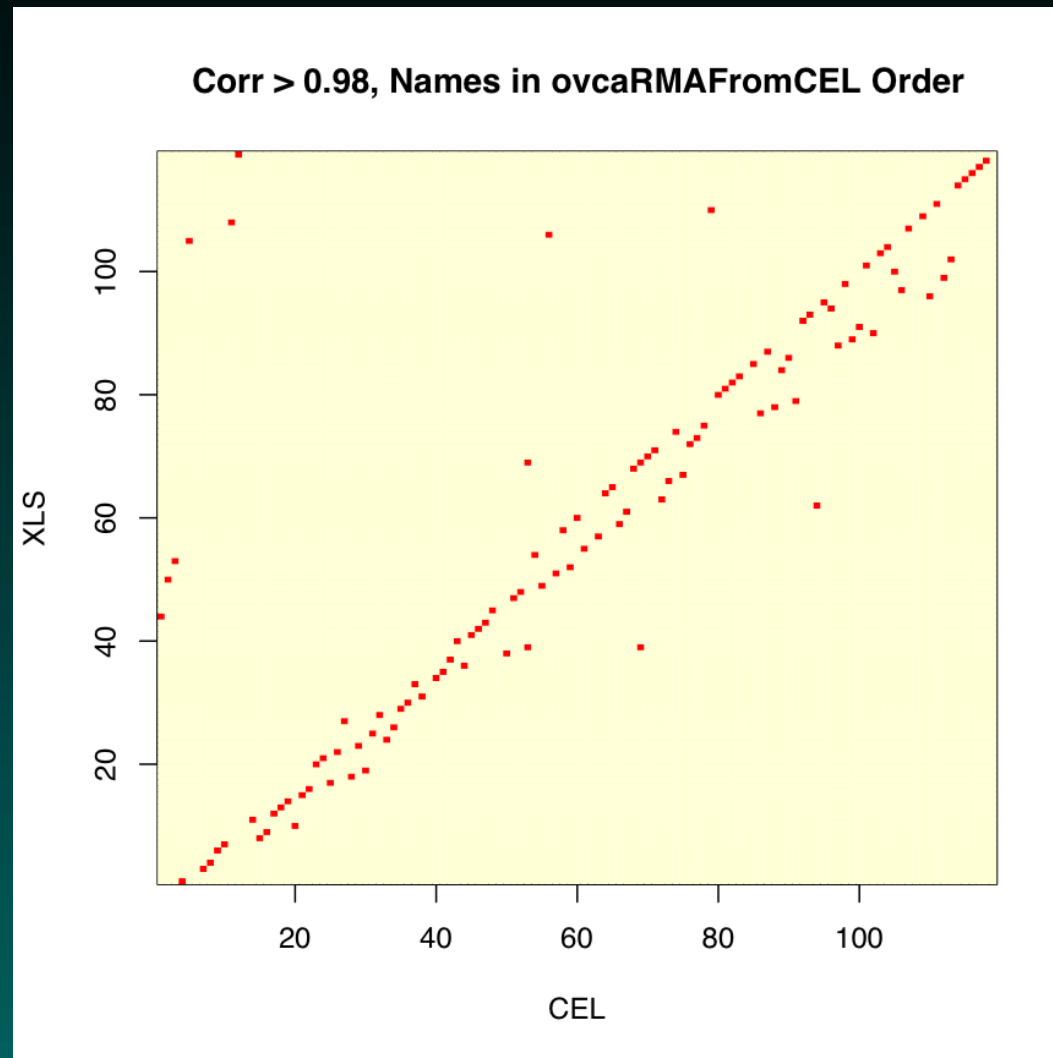
Data from the authors' web site for an earlier paper in Nature (Bild et al, 2006),

<http://data.cgt.duke.edu/oncogene.php>, supplies CEL files and clinical information for 146 ovarian tumor samples, a superset of the ones examined by Dressman et al.

Checking the entire Bild set,
XLS M4161 corresponds to D2159
XLS M444 corresponds to D2171
XLS D1837 corresponds to D2247.

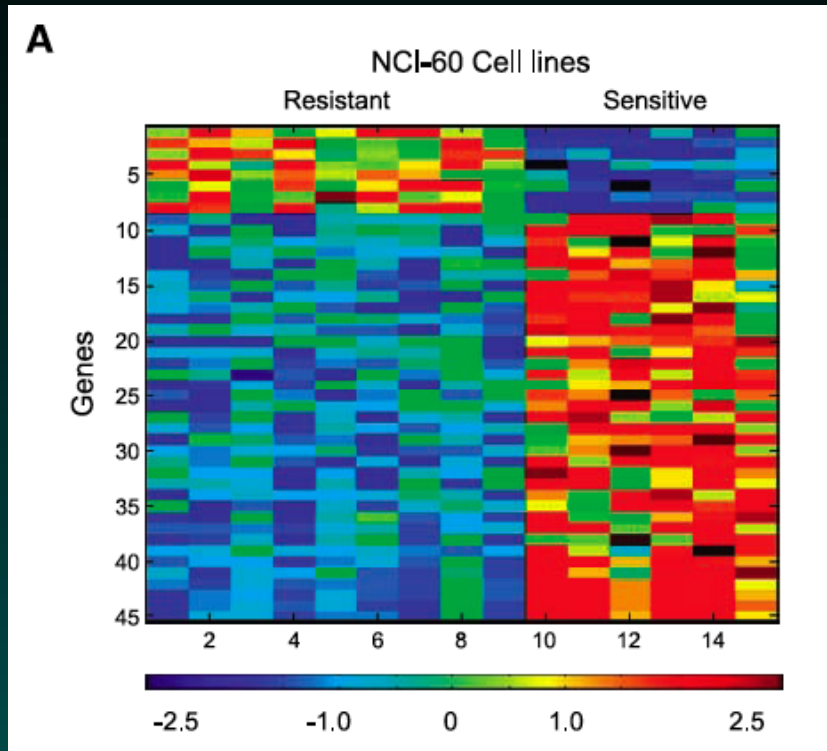
Can we see what happened?

Where the Best Fits Are...



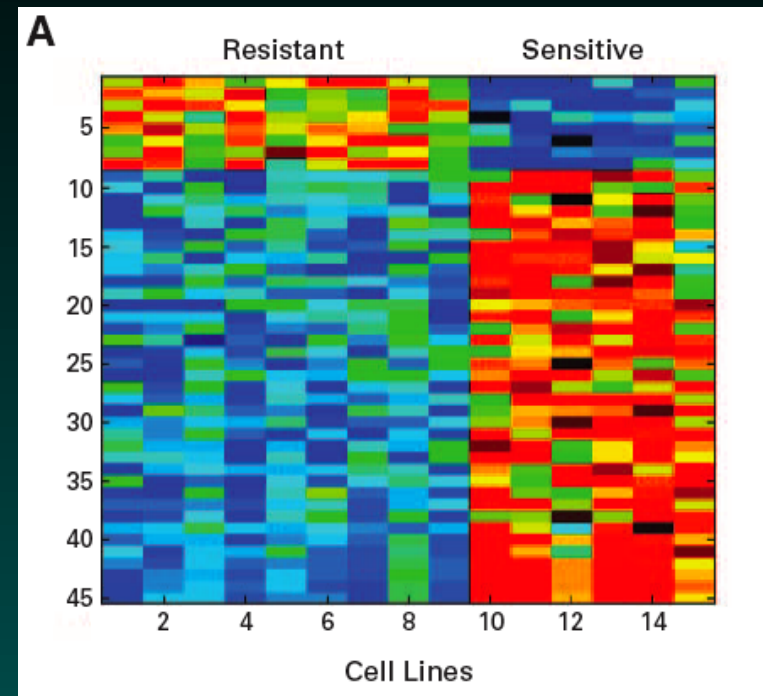
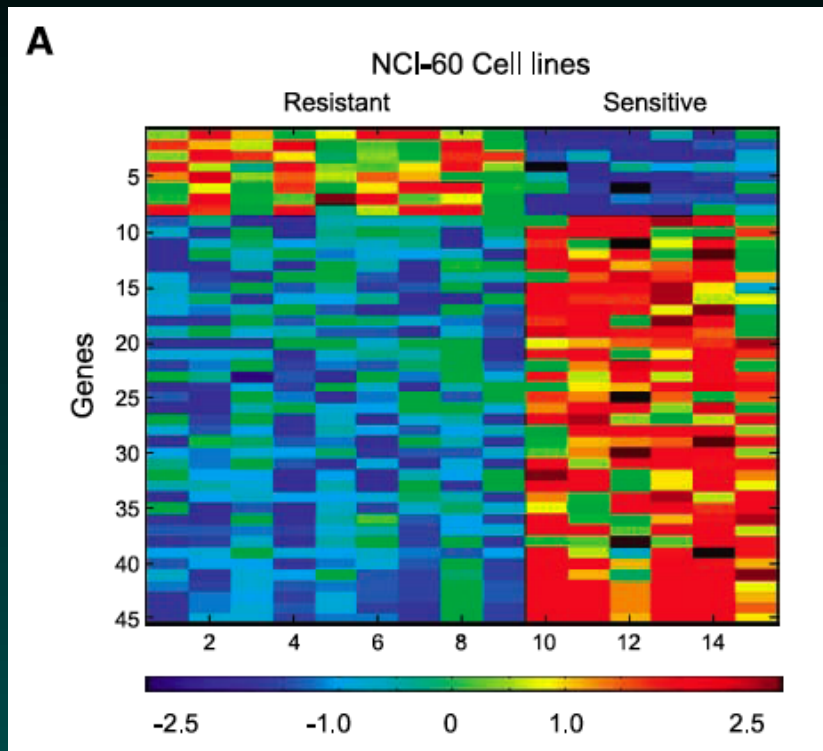
Most of the poor fits are 3 names off.

Temozolomide Heatmaps



Augustine et al., 2009, *Clin Can Res*, 15:502-10, Fig 4A.
Temozolomide, NCI-60.

Temozolomide Heatmaps



Augustine et al., 2009, *Clin Can Res*, 15:502-10, Fig 4A.
Temozolomide, NCI-60.

Hsu et al., 2007, *J Clin Oncol*, 25:4350-7, Fig 1A.
Cisplatin, Gyorffy cell lines.

Some Observations

The most common mistakes are simple.

Confounding in the Experimental Design

Mixing up the sample labels

Mixing up the gene labels

Mixing up the group labels

(Most mixups involve simple switches or offsets)

This simplicity is often hidden.

Incomplete documentation

Unfortunately, we suspect

The most simple mistakes are common.

Steps Towards Reproducibility at MDA

Literate Programming. For the past two years, we have required the analysts in our department to prepare reports in *Sweave*, where documentation has been interweaved with the code used to produce the analysis. There's also *odfWeave* or *SASWeave*, if you prefer.

Reusing Templates. For common types of analyses, we're starting from canonical template analyses that have been assembled in depth. This standardizes and speeds the analysis. We have also produced **RATB** – the “Report and Analysis Template Builder” – for this purpose.

Steps Towards Clarity at MDA

Report Structure. The biologists we work with often don't want to plow through the code, but they do want to understand. This is aided by following a format they're familiar with: Introduction, Methods, Results, Conclusions.

Executive Summaries. Each report is prefaced by a brief textual description following the same outline.

Appendices. Some things we want to know all the time:

- (1) *SessionInfo*, the libraries used,
- (2) *Saves*, the data produced, and
- (3) *File Location*, where the data is.

Acknowledgements

Kevin Coombes

Shannon Neeley, Jing Wang

David Ransohoff, Gordon Mills

Jane Fridlyand, Lajos Pusztai, Zoltan Szallasi

MDACC Ovarian SPORE, Lung SPORE, Breast SPORE