The Importance of Reproducibility in •High-Throughput Biology: Case Studies in Forensic Bioinformatics

- Keith A. Baggerly

Bioinformatics and Computational Biology

- UT M. D. Anderson Cancer Center
- kabagg@mdanderson.org

Cambridge, 4 September 2010

Why is RR So Important in H-TB?

Our intuition about what “makes sense” is very poor in high dimensions. To use “genomic signatures” as biomarkers, we need to know they’ve been assembled correctly.

Without documentation, we may need to employ *forensic bioinformatics* to infer what was done to obtain the results.

Let’s examine some case studies involving an important clinical problem: *can we predict how a given patient will respond to available chemotherapeutics?*

Using the NCI60 to Predict Sensitivity

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.

Fit Training Data

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

Fit Testing Data

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

5-FU Heatmaps

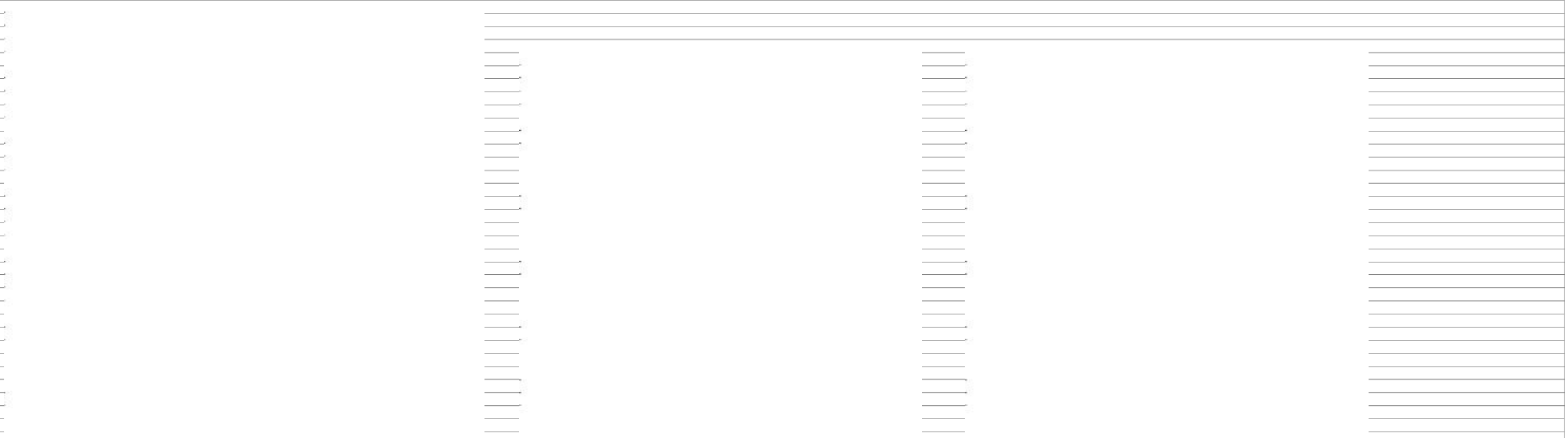
Nat Med Paper

5-FU Heatmaps

Nat Med Paper

Our t-tests

5-FU Heatmaps

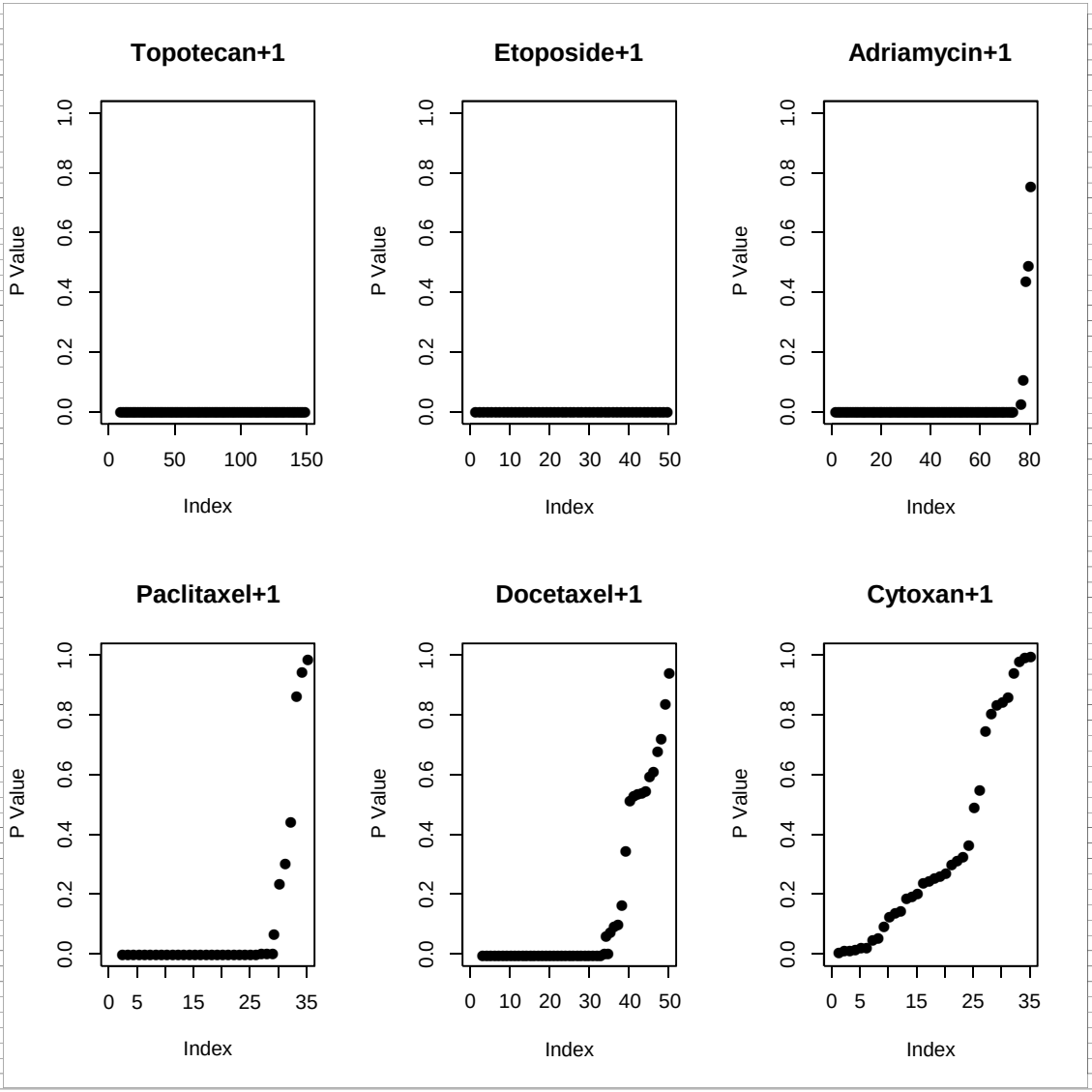


Nat Med Paper Our t-tests Reported Genes

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
...
[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: Other Drugs



Using Their Software

Their software requires two input files:

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

2. *a list of probeset ids* in the same order as the quantification matrix. *This list must not have a header row.*

What do we get?

Heatmaps Match Exactly for Most Drugs!

From the paper:

From the software:

Heatmaps Match Exactly for Most Drugs!

From the paper:

From the software:

We match heatmaps but not gene lists? We'll come back to this, because their software also gives *predictions*.

Predicting Docetaxel (Chang 03)

Predicting Adriamycin (Holleman 04)

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 Chang et al’s list comprise 14/19 outliers.

The others: ERCC1, ERCC4, ERBB2, BCL2L11, TUBA3.
These are the genes named to explain the biology.

RR Theme: 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 (*and from Nat Med):

http://bioinformatics.mdanderson.org/Supplements/ReproRsch-Chemo.

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!)

Adriamycin 0.9999+ Correlations (Reply)

Adriamycin 0.9999+ Correlations (Reply)

Redone in Aug 08, “using only the 95 unique samples”

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

15 duplicates; 6 inconsistent. (61R, 13S, 6B) vs (22,48,10).

Validation 1: Hsu et al

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

Same approach, using Cisplatin and Pemetrexed.

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

With some work, we matched the heatmaps. (Gene lists?)

The 4 We Can't Match (Reply)

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

The last two probesets are special.

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

Some Timeline Here...

Nat Med Nov 06*, Nov 07*, Aug 08. JCO Lung Oct 07*.
Lancet Oncology Breast Dec 07*. (* errors reported)

Some Timeline Here...

Nat Med Nov 06*, Nov 07*, Aug 08. JCO Lung Oct 07*.
Lancet Oncology Breast Dec 07*. (* errors reported)

May/June 2009: we learn clinical trials had begun.
2007: pemetrexed vs cisplatin, pem vs vinorelbine.
2008: docetaxel vs doxorubicin, topotecan vs dox (Moffitt).

Some Timeline Here...

Nat Med Nov 06*, Nov 07*, Aug 08. JCO Lung Oct 07*.
Lancet Oncology Breast Dec 07*. (* errors reported)

May/June 2009: we learn clinical trials had begun.
2007: pemetrexed vs cisplatin, pem vs vinorelbine.
2008: docetaxel vs doxorubicin, topotecan vs dox (Moffitt).

Sep 1. Paper submitted to *Annals of Applied Statistics*.
Sep 14. Paper online at *Annals of Applied Statistics*.
Sep-Oct: Story covered by *The Cancer Letter*, Duke starts internal investigation, suspends trials.
So, what happened next?

Jan 29, 2010

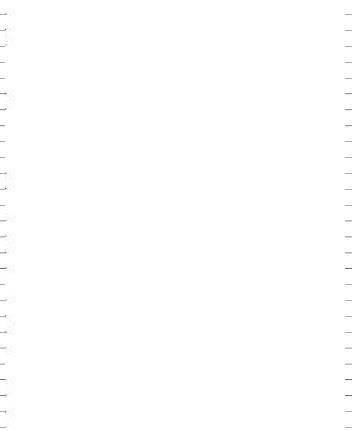
Their investigation's results *"strengthen ... confidence in this evolving approach to personalized cancer treatment."*

Why We're Unhappy...

"While the reviewers approved of our sharing the report with the NCI, *we consider it a confidential document*" (Duke). A *future paper* will explain the methods.

Why We're Unhappy...

"While the reviewers approved of our sharing the report with the NCI, *we consider it a confidential document*" (Duke). A *future paper* will explain the methods.

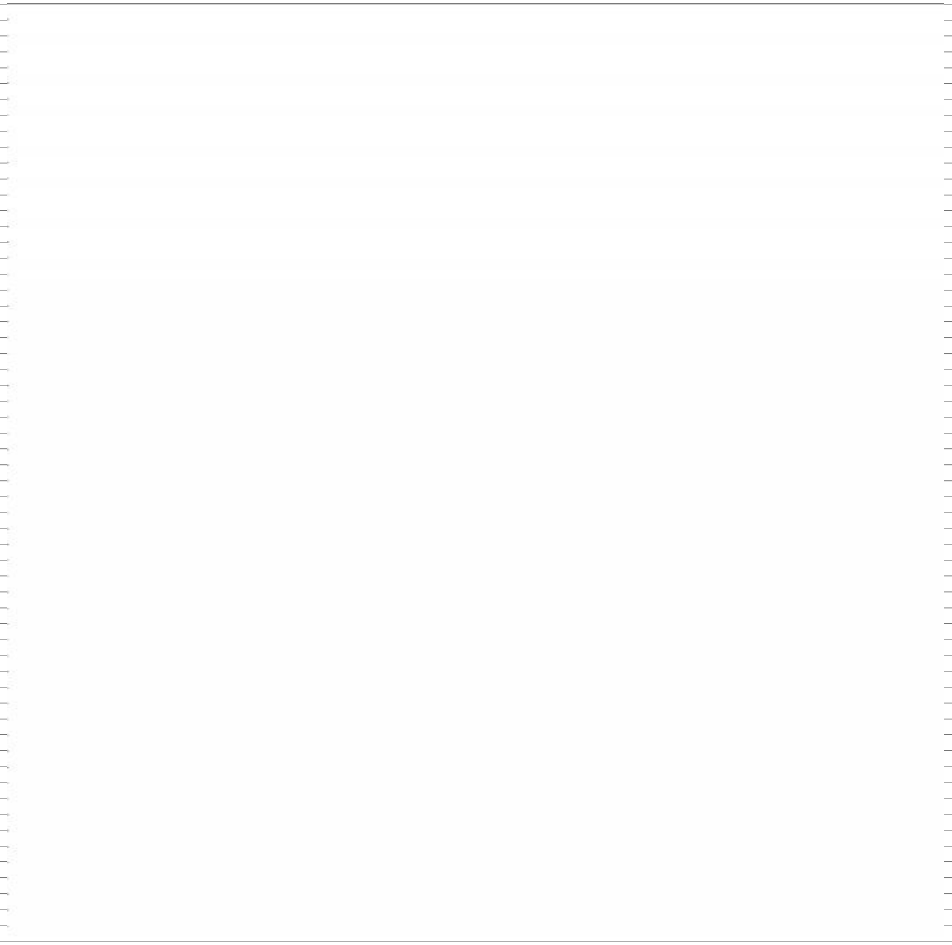


oh, there's just one more thing...

In mid-Nov (mid-investigation), the Duke team posted new data for cisplatin and pemetrexed (in trials since '07).

These included quantifications for 59 ovarian cancer test samples (from GSE3149) used for predictor validation.

We Tried Matching The Samples



We correlated the 59 vectors with all samples in GSE3149.
43 samples are mislabeled; 16 don’t match at all.

FOI(L)A!

April 7: Paul Goldberg of *the Cancer Letter* requests “access to and copies of the report (and attendant data)” from the NCI under the Freedom of Information Act (FOIA).

May 3: redacted report supplied.

FOI(L)A!

April 7: Paul Goldberg of *the Cancer Letter* requests "access to and copies of the report (and attendant data)" from the NCI under the Freedom of Information Act (FOIA).
May 3: redacted report supplied.

"*we were unable to identify a place where the statistical methods were described in sufficient detail to independently replicate the findings of the papers*." – review panel

The report makes no mention of the problems with cisplatin/pemetrexed that arose during the investigation.

May 14, 2010

"We have asked [CALGB] to remove the Lung Metagene Score from the trial, because we were unable to confirm the score's utility" – *Jeff Abrams, CTEP director*

(The NCI doesn't directly sponsor the resumed trials.)

July 16, 2010

July 19, 2010

"Duke administrators accomplished something monumental: they triggered a public expression of outrage from biostatisticians."

A Baron, K Bandeen-Roche, D Berry, J Bryan, V Carey, K Chaloner, M Delorenzi, B Efron, R Elston, D Ghosh, J Goldberg, S Goodman, F Harrell, S Hilsenbeck, W Huber, R Irizarry, C Kendziorski, M Kosorok, T Louis, JS Marron, M Newton, M Ochs, G Parmigiani*, J Quackenbush, G Rosner, I Ruczinski, Y Shyr*, S Skates, TPSpeed, JD Storey, Z Szallasi, R Tibshirani, S Zeger

Req to Varmus, DoD, ORI, Duke: suspend trials.

Subsequent Events, and a Caveat

Duke announces trials resuspended

NPR blog, Science blog, Nature blog, NYT blog, article

Lancet Oncology issues Expression of Concern

Varmus & Duke request IOM Involvement

Questions raised about NEJM paper

JCO launches investigation

More awards found to be wrong, COI claims

http://groups.google.com/group/reproducible-research

Correspondence to Nature

Subsequent Events, and a Caveat

Duke announces trials resuspended
NPR blog, Science blog, Nature blog, NYT blog, article
Lancet Oncology issues Expression of Concern
Varmus & Duke request IOM Involvement
Questions raised about NEJM paper
JCO launches investigation
More awards found to be wrong, COI claims

http://groups.google.com/group/reproducible-research
Correspondence to Nature

We've seen problems like these before. CAMDA 2002. Proteomics 2003-5. TCGA current. Others at MDA.

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.

What Should the Norm Be?

For papers?

Things we look for:

1. Data (often mentioned, given MIAME)
2. Provenance
3. Code
4. Descriptions of Nonscriptable Steps
5. Descriptions of Planned Design, if Used.

For clinical trials?

Some Lessons

Is our own work reproducible?

Literate Programming. For the past two years, we have required reports to be prepared in *Sweave*.

Reusing Templates.

Report Structure.

Executive Summaries.

Appendices. Some things we want to know all the time: *SessionInfo*, *Saves*, and *File Location*.

The buzz phrase is *reproducible research*.

Some Acknowledgements

Kevin Coombes

Shannon Neeley, Jing Wang

David Ransohoff, Gordon Mills

Jane Fridlyand, Lajos Pusztai, Zoltan Szallasi

MDACC Ovarian SPORE, Lung SPORE, Breast SPORE

Now in the *Annals of Applied Statistics!* Baggerly and Coombes (2009), 3(4):1309-34.
http://bioinformatics.mdanderson.org/Supplements/ReproRsch-All

Validation 2: Bonnefoi et al

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

Similar approach, using signatures for Fluorouracil, Epirubicin Cyclophosphamide, and Taxotere to predict response to 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

How Are Results Combined?

Potti et al predict response to TFAC, Bonnefoi et al to 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?

What About Blinded Validation?

"Data was made available to us, blinded. All we got was the gene expression data. We ran the predictions and sent it back to the EORTC investigators" – *Joe Nevins, Oct 2.*

What About Blinded Validation?

"Data was made available to us, blinded. All we got was the gene expression data. We ran the predictions and sent it back to the EORTC investigators" – *Joe Nevins, Oct 2.*

```
Sample info supplied:
Arm, Composite label
A, npCREp P- T3 N1 HB01  ...
A, pCR Ep Pp T2 N1 HB04
```

The data weren't blinded.

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.

Index