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 9, 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

Genomic signatures to guide the use of chemotherapeutics

Anil Potti¹², Holly K Dressman¹³, Andrea Bild¹³, Richard F Riedel¹², Gina Chan⁴, Robyn Sayer⁴, Janiel Cragun⁴, Hope Cottrill⁴, Michael J Kelley², Rebecca Petersen⁵, David Harpole⁵, Jeffrey Marks⁵, Andrew Berchuck¹⁶, Geoffrey S Ginsburg¹², Phillip Febbo¹³, Johnathan Lancaster⁴ & Joseph R Nevins¹³


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.
Gathering Data


2. Training (NCI60): Affy U95Av2, triplicate runs (http://dtp.nci.nih.gov/mtargets/download.html)

We want the test data to split like this...
But it doesn’t. Did we do something wrong?
Examineing Signatures

Lists of probesets used were given in a supplementary table.

The paper explains why many of these genes make sense.

How were the genes found? Supplementary methods:
“a variance fixed t-test was used to calculate significance”.
5-FU Heatmaps

Nat Med Paper  Our t-tests  Reported Genes
Their List and Ours

```r
> 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: 5FU

5–FU List Pvalues

5–FU List+1 Pvalues

© Copyright 2009-10, KA Baggerly, KR Coombes
Offset P-Values: Other Drugs

Topotecan+1

Etoposide+1

Adriamycin+1

Paclitaxel+1

Docetaxel+1

Cytoxan+1

© Copyright 2009-10, KA Baggerly, KR Coombes
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 Docetaxel!

From Potti et al, Figure 1

From the software
Heatmaps Match Exactly for 5 Others!

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)

Accuracy: 22/24 (91.6%)

Probability of docetaxel resistance

Sample number

Sigmoidal curves for docetaxel sensitive and resistant samples.

Probability of docetaxel resistance

Sodium levels for docetaxel sensitive and resistant samples.

Histogram of residual tumor sizes for sensitive and resistant samples.
Predicting Adriamycin (Holleman 04)

- Probability of Adriamycin sensitivity:
  - Adriamycin resistant
  - Adriamycin sensitive

- Accuracy: 99/122 (81%)

- Graph showing percentage of patients by Daunorubicin LC50 (µg/ml):
  - Total (n=572)
  - Study group (n=148)
  - Sensitive
  - Resistant

© Copyright 2009-10, KA Baggerly, KR Coombes
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!


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

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


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

Hervé Bonnefoi, Anil Potti, Mauro Dei Tos, Louis Mauriac, Mario Campone, Michèle Tubiana-Hulin, Thierry Petit, Philippe Rouanet, Jacek Jassem, Emmanuel Blot, Véronique Becette, Pierre Farmer, Sylvie André, Chaitanya R Acharya, Seayan Mukherjee, David Cameron, Jonas Bergh, Joseph R Nevins, Richard D Iggo
Adriamycin 0.9999+ Correlations (Reply)

Redone in Aug 08, “using only the 95 unique samples”
### The First 20 Files Now Named

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GSM44303 RES</td>
</tr>
<tr>
<td>2</td>
<td>GSM44304 RES</td>
</tr>
<tr>
<td>3</td>
<td>GSM9653 RES</td>
</tr>
<tr>
<td>4</td>
<td>GSM9653 RES</td>
</tr>
<tr>
<td>5</td>
<td>GSM9654 RES</td>
</tr>
<tr>
<td>6</td>
<td>GSM9655 RES</td>
</tr>
<tr>
<td>7</td>
<td>GSM9656 RES</td>
</tr>
<tr>
<td>8</td>
<td>GSM9657 RES</td>
</tr>
<tr>
<td>9</td>
<td>GSM9658 SEN</td>
</tr>
<tr>
<td>10</td>
<td>GSM9658 SEN</td>
</tr>
<tr>
<td>11</td>
<td>GSM9694 RES</td>
</tr>
<tr>
<td>12</td>
<td>GSM9695 RES</td>
</tr>
<tr>
<td>13</td>
<td>GSM9696 RES</td>
</tr>
<tr>
<td>14</td>
<td>GSM9698 RES</td>
</tr>
<tr>
<td>15</td>
<td>GSM9699 SEN</td>
</tr>
<tr>
<td>16</td>
<td>GSM9701 RES</td>
</tr>
<tr>
<td>17</td>
<td>GSM9708 RES</td>
</tr>
<tr>
<td>18</td>
<td>GSM9708 SEN</td>
</tr>
<tr>
<td>19</td>
<td>GSM9709 RES</td>
</tr>
<tr>
<td>20</td>
<td>GSM9711 RES</td>
</tr>
</tbody>
</table>
The First 20 Files Now Named

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM44303</td>
<td>RES</td>
</tr>
<tr>
<td>GSM44304</td>
<td>RES</td>
</tr>
<tr>
<td>GSM9653</td>
<td>RES</td>
</tr>
<tr>
<td>GSM9653</td>
<td>RES</td>
</tr>
<tr>
<td>GSM9654</td>
<td>RES</td>
</tr>
<tr>
<td>GSM9654</td>
<td>RES</td>
</tr>
<tr>
<td>GSM9655</td>
<td>RES</td>
</tr>
<tr>
<td>GSM9655</td>
<td>RES</td>
</tr>
<tr>
<td>GSM9656</td>
<td>RES</td>
</tr>
<tr>
<td>GSM9656</td>
<td>RES</td>
</tr>
<tr>
<td>GSM9657</td>
<td>RES</td>
</tr>
<tr>
<td>GSM9657</td>
<td>RES</td>
</tr>
<tr>
<td>GSM9658</td>
<td>SEN</td>
</tr>
<tr>
<td>GSM9658</td>
<td>SEN</td>
</tr>
<tr>
<td>GSM9658</td>
<td>SEN</td>
</tr>
</tbody>
</table>

15 duplicates; 6 inconsistent. (61R, 13S, 6B) vs (22,48,10).
Validation 1: Hsu et al

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


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.
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

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.
Does cytoxan make sense?
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 Timeline Here...

JCO Lung Oct 07*.
Lancet Oncology Breast Dec 07*.
CCR Temozolomide Jan 09*.
(* errors reported to journals.)
... other more recent papers ...

Things we learned May/June 2009:

clinical trials had begun.
2007: pemetrexed vs cisplatin, pem vs vinorelbine.
2008: docetaxel vs doxorubicin, topotecan vs dox (Moffitt).
More Timeline (2009)


Late Sep. Duke starts internal investigation.
Oct 2. Story covered by *The Cancer Letter*.*
Oct 23. Blinded validation discussed in *The Cancer Letter*. *

(Jan/Feb 2010 - *The IMS Bulletin*!)

* Isn’t all this moot if it works in a blinded validation?
Well, About That “Blinding”

“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, npCR Ep P- T3 N1 HB01 ...
A, pCR Ep Pp T2 N1 HB04

The data weren’t blinded.

“we would not be able to reproduce the reported probabilities with the information we have about how they were obtained.” – Mauro Delorenzi, Oct 23.

Or validated.

So, what happened next?
Duke University said it is in the process of restarting three clinical trials using microarray analysis of patient tumors to predict their response to chemotherapy.

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.

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.
Why Can’t We Match Some At All?

We checked the first 100 probeset intensities across samples. The first 16 don’t match because the genes are mislabeled. We reported this to Duke and to the NCI in mid-November. All data was stripped from the websites within the week.
So, What Next?

The trials resumed.
We waited to see the methods.
We waited.
We tried being patient.

We’re not very good at it.

We know Duke won’t show us the report.
But Duke showed it to the NCI.
Would the NCI show us the report?
Might the NCI have to show us the report?
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). “I look forward to your reply within 20 business days, as the statute requires.”

April 26: NCI agrees in principle to release the report, redacting only the names of the authors. Duke legal is allowed further redactions to protect trade secrets.

May 3: redacted report supplied.

May 7: other statisticians invited to comment.

May 14: story covered in the Cancer Letter.
Some Interesting Things...

“In our review of the methods... we were unable to identify a place where the statistical methods were described in sufficient detail to independently replicate the findings of the papers. Only by examining the R code from Barry were we able to uncover the true methods used.”

The Duke investigators really need to work on “clearly explaining ... the specific statistical steps used in developing the predictors and the prospective sample assignments”

The supporting data and code weren’t sent to the NCI.

The report makes no mention of the problems with cisplatin/pemetrexed that arose during the investigation.
May 14, 2010

NCI Raises New Questions About Duke Genomics Research, Cuts Assay From Trial

By Paul Goldberg

In a new setback to a controversial group of genomics researchers at Duke University, NCI officials eliminated a biomarker test from an ongoing phase III clinical trial.

“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

“When the issues came up with the review by Duke of their studies, we decided to review the LMS score in the trial we sponsored” (CALGB 30506).

(The NCI doesn’t directly sponsor the resumed trials.)
Prominent Duke Scientist Claimed Prizes He Didn't Win, Including Rhodes Scholarship

By Paul Goldberg
July 19, 2010

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


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

© Copyright 2009-10, KA Baggerly, KR Coombes
Subsequent Events

NPR blog
Duke announces trials resuspended
Science blog, Nature blog
NYT blog, article
Lancet Oncology issues Expression of Concern
NEJM states no questions raised
Varmus & Duke request IOM Involvement
Questions raised about NEJM paper
JCO launches investigation
Science news feature
More awards found to be wrong, COI claims
Scientists for RR

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

Correspondence to Nature

Working on White Paper Guidelines
It’s Not Just Them

This is a particularly egregious combination, but we’ve seen many of these problems before.

Critical Analysis of Microarray Data (CAMDA) 2002: annotations in the contest dataset were scrambled due to an Excel error.

Proteomics 2003-5: several studies showed effects driven by design confounding; calibration (annotation) and processing inconsistencies.

TCGA (current): label scrambling going from label 1 (raw) to label 2 (processed) data.

Other examples that never left MD Anderson.
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.
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


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


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

Two RMA Quantifications of Sample 872

CELS vs Tables. We expected better (fewer outliers).
Looking at Their Other Quants

Finding the Best Match, CEL RMA Column 1

Which one would you pick?
Looking at The “Best” Fit

Two RMA Quantifications: 872 From CEL, 2476 From XLS

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.