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

Gulf Coast Consortium, Jan 18, 2017
Why is Reproducibility Important in H-T B?

Our intuition about what “makes sense” is very poor in high dimensions.

To use “omics-based signatures” as biomarkers, we need to know they’ve been assembled correctly.

Without documentation, we may need to employ (lengthy!) forensic bioinformatics to infer what was done.

Let’s look at examples in the context of a specific problem: can we predict which patients will respond to which chemotherapeutics?
Using Cell Lines to Predict Sensitivity

Genomic signatures to guide the use of chemotherapeutics

Anil Potti¹,², Holly K Dressman¹,³, Andrea Bild¹,³, Richard F Riedel¹,², Gina Chan⁴, Robyn Sayer⁴, Janiel Cragun⁴, Hope Cottrill⁴, Michael J Kelley², Rebecca Petersen⁵, David Harpole⁵, Jeffrey Marks⁵, Andrew Berchuck¹,⁶, Geoffrey S Ginsburg¹,², Phillip Febbo¹,³, Johnathan Lancaster⁴ & Joseph R Nevins¹,³


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

They provide examples using 7 commonly used agents.

This got people at MDA very excited.
Their Gene List and Ours

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

<table>
<thead>
<tr>
<th></th>
<th>Theirs</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>&quot;1881_at&quot;</td>
<td>&quot;1882_g_at&quot;</td>
</tr>
<tr>
<td>4</td>
<td>&quot;31321_at&quot;</td>
<td>&quot;31322_at&quot;</td>
</tr>
<tr>
<td>5</td>
<td>&quot;31725_s_at&quot;</td>
<td>&quot;31726_at&quot;</td>
</tr>
<tr>
<td>6</td>
<td>&quot;32307_r_at&quot;</td>
<td>&quot;32308_r_at&quot;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Predicting Response: Docetaxel

Potti et al (2006), Nature Medicine, 12:1294-300, Fig 1d

Chang et al, Lancet 2003, 362:362-9, Fig 2 top
Predicting Response: Adriamycin

Potti et al (2006), Nature Medicine, 12:1294-300, Fig 2c

Holleman et al, NEJM 2004, 351:533-42, Fig 1
Partial Timeline

2006:
* Nov 8: Our first questions to Potti and Nevins.
* Nov 21: Our first report describing errors.

2007:
* Jan 24: We meet with Nevins at M.D. Anderson. We urge him to review the data.
* Feb-Apr: New data and code are posted. Some numbers change. We tell them we don’t think it works.
* Apr 25: We send Potti and Nevins a draft for comment.
* May: We find problems with outliers. Potti and Nevins continue to insist it works, and want to “bring this to a close”.
Adriamycin 0.9999+ Correlations

Redone Aug 08, “using ... 95 unique samples”.
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

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

Another problem –
The 4 We Can’t Match

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

Another problem –

*The last two 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 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 (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
See Leek et al, Nat Rev Genet, 2010 for more examples.

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.
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.

$$P(\text{TFAC}) = P(T) + P(F) + P(A) + P(C) - P(T)P(F)P(A)P(C).$$
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.

$$P(TFAC') = P(T) + P(F) + P(A) + P(C) - P(T)P(F)P(A)P(C).$$

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

\[
P(\text{TFAC}) = P(T) + P(F) + P(A) + P(C) - P(T)P(F)P(A)P(C).
\]

\[
P(\text{ET}) = \max[P(E), P(T)].
\]

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

\[
P(\text{TFAC}) = P(T) + P(F) + P(A) + P(C) - P(T)P(F)P(A)P(C).
\]

\[
P(\text{ET}) = \max[P(E), P(T)].
\]

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

Each rule is different.
Predictions for Individual Drugs?

Does cytoxan make sense?
Temozolomide Heatmaps

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.
The Reason We Really Care

Jun 2009: we learn clinical trials had begun.
2007: pemetrexed vs cisplatin, pem vs vinorelbine.
2008: docetaxel vs doxorubicin, topotecan vs dox (Moffitt).
The Reason We Really Care

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

Sep 1, 2009: We submit a paper describing case studies to the Annals of Applied Statistics.

Sep 14, 2009: Paper accepted and available online at the Annals of Applied Statistics.

Sep-Oct 2009:

NCI raises concerns with Duke’s IRB behind the scenes.
Duke starts internal investigation, suspends trials.
New Data

Early-Nov ’09 (mid-investigation), the Duke team posted new data for cisplatin and pemetrexed (in lung trials since ’07).

These included quantifications for the 59 ovarian cancer test samples (from GSE3149, which has 153 samples) they used to validate their predictor.
We Tried Matching The Samples

43 samples are mislabeled.
16 samples don’t match because the genes are mislabeled.
All of the validation data are wrong.

We reported this to Duke and to the NCI in mid-November.
Jan 29, 2010

Duke In Process To Restart Three Trials Using Microarray Analysis Of Tumors

By Paul Goldberg

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.”
We Asked for the Data

“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.

This did give us one more option...
We Asked for the Data

“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.

This did give us one more option...

In May 2010, we obtained a copy of the reviewers’ report from the NCI under FOIA (Cancer Letter, May 14).

In our assessment (and others’), it did not justify restarting trials.

There was no mention of our Nov 2009 report.
A Catalyzing Event: July 16, 2010

Jul 19/20: Letter to Varmus; Duke resuspends trials.
Oct 22/9: First call for paper retraction.
Nov 9: Duke terminates trials.
Nov 19: call for Nat Med retraction, Potti resigns
Other Developments

117 patients were enrolled in the trials.

**Sep, 2011:** Patient lawsuits filed (11+ settlements).

Misconduct investigation (**ongoing**).

10 retractions, 6+ “partial retractions”

FDA Review, Discussions with Duke IRB

**Jul 8, 2011:** Front Page, NY Times.

**Feb 12, 2012:** 60 Minutes.


**Mar 23, 2012:** IOM Report Released.

Developments in 2015

Jan 9, Cancer Letter highlights whistleblowing report by Brad Perez written in April 2008.

Jan 16, Cancer Letter: “I Hope NCI Doesn’t Get Original Data”, from May 2008 email not produced in discovery

Nevins testimony on time required to identify “abundantly clear” falsification: “It would take you maybe an hour.”

April/May, Lawsuits settled; conditions undisclosed

Nov 9, Office of Research Integrity releases findings of research misconduct

Nov 13, editorial
Some Cautions/Observations

This case is pathological.

But we’ve seen similar problems before.

*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
This is not an Isolated Problem


NCI focus meeting Sep 2012.


NAS meeting Feb 26-7, 2015.

ENAR Webinar Nov 20, 2015

SISBID RR Short Course July 18-20, 2016
Some Cost Breakdowns

What Have We and Others Suggested?

Exploiting a Teachable Moment...


Give us your data, your code, your huddled masses

Records of data provenance

Checking existence as a task for journals and reviewers (are there links? are they live?)

Reasons for Hope

1. Our Own (Evolving!) Experience
2. Better tools (knitr, markdown, GitHub)
3. Journals, Code and Data
4. The IOM, the FDA, and IDEs*
5. The NCI and Trials it Funds
6. OSTP, Congress, Science, Nature
Some Places to Learn More

Karl Broman’s Tools for RR Course

Roger Peng’s Coursera course and notes (2013)

Christopher Gandrud’s book (2e, 2015)

Yihui Xie’s book (2e, 2015)


NAS meeting, Feb 26-7, 2015

ENAR Webinar, Nov 20, 2015

SISBID Reproducible Research Short Course, July 2016
Food for Thought/Questions

Could someone else, working with your notes/reports, get the same results?

What are downsides to reproducibility?

Are there incentives for irreproducibility?

What additional problems come up with replication (getting similar results with new data) as opposed to reproduction (the same results from the same data)?

How should reproducibility/replicability be incorporated into the reward structure?

What data should be required, and when?
Acknowledgments

Kevin Coombes

Yang Zhao, Ying Wang, Shelley Herbrich
Shannon Neeley, Jing Wang
David Ransohoff, Gordon Mills
Jane Fridlyand, Lajos Pusztai, Zoltan Szallasi

M.D. Anderson Ovarian, Lung and Breast SPOREs


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

For updates: http://bioinformatics.mdanderson.org/Supplements/ReproRsch-All/Modified.