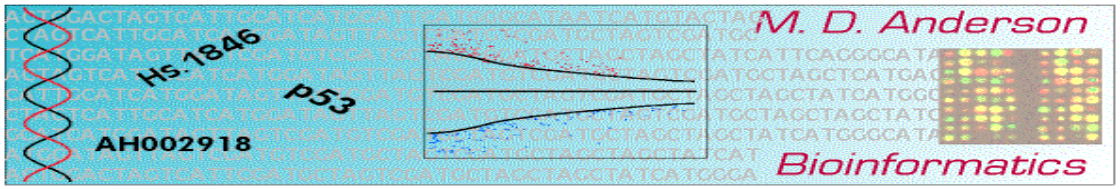
Forensic Bioinformatics

Keith A. Baggerly
Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org

Institute of Medicine, March 31, 2011



What is Forensic Bioinformatics?

Forensic bioinformatics is the art of using raw data and reported results to infer what the methods must have been.

Ideally, it shouldn't be required. Empirically, it is.

This is problematic in general, but even worse with many variables because our intuition about what "makes sense" is very poor in high dimensions.

To use omics signatures as biomarkers, we need to know how they've been assembled.

Specific Questions

1. What barriers did we encounter?

2. What was our experience with journals?

3. Were your experiences here similar or different from what you've encountered in previous forensic bioinformatic studies?

4. What were your experiences in the OvaCheck case?

5. What motivates a forensic bioinformatics study?

6. What recommendations would you make to the committee?

To get to these questions, I'm going to review the Duke case prior to our *Annals* paper.

The Beginning

The Potti et al. paper on deriving genomic signatures of drug sensitivity appeared in the Nov 2006 *Nature Medicine*.

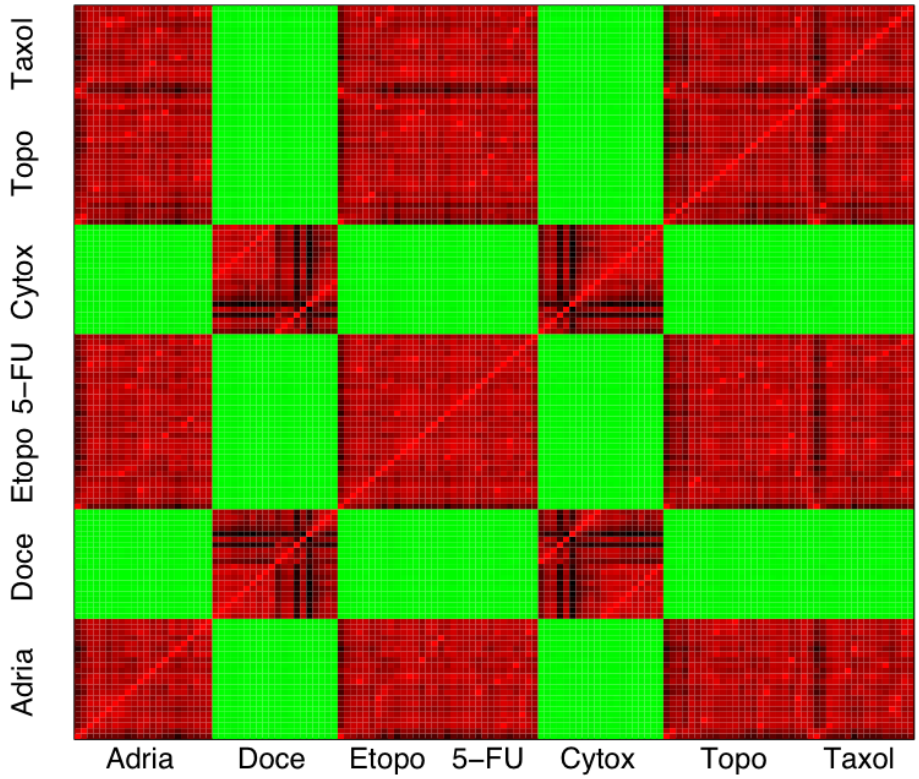
Potti et al. caused enough of a stir that several of our colleagues asked if we could help them implement related approaches at MD Anderson.

Nov 8, 2006: We send our first email to Nevins.

Nov 16: Nevins replies, and Potti sends data.

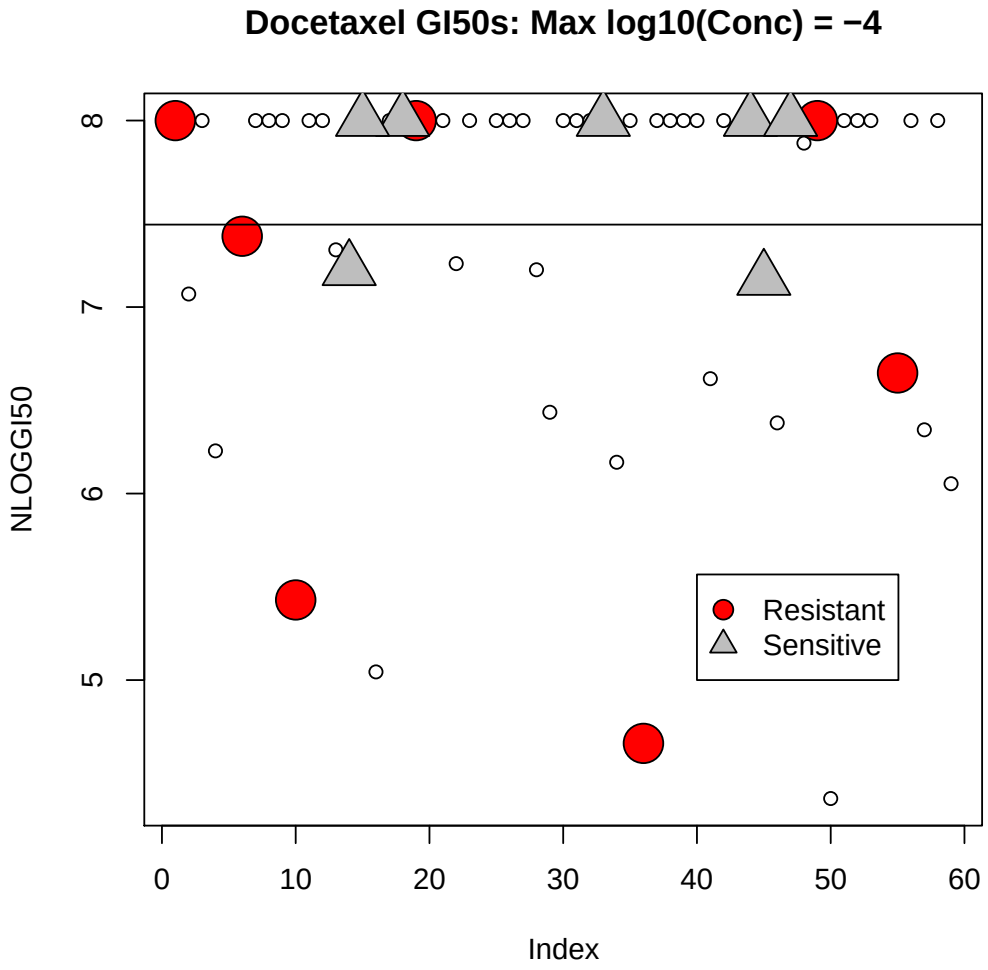
Nov 21: We send back our first report.

Data Inconsistencies



Some cell lines don’t match.

The next iteration: GI50 problems



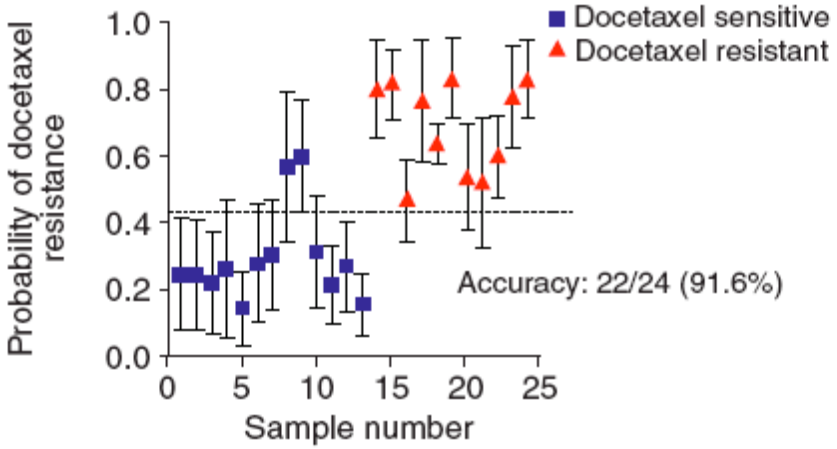
Nov 21: New docetaxel data. Nov 27: Our next report.

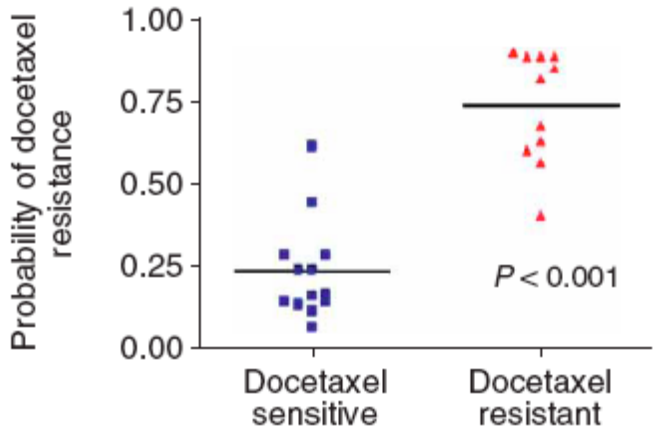
The next iteration: Off by one

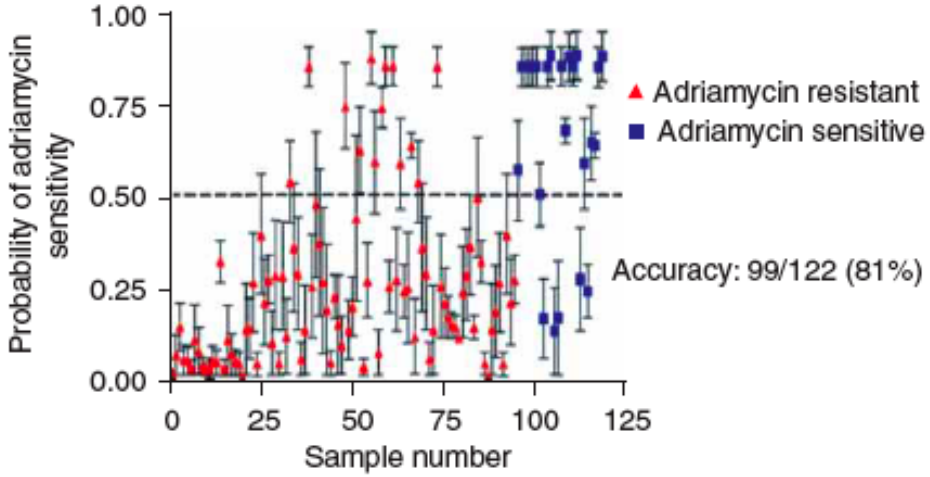
```
> temp <- cbind(
    sort(rownames(pottiUpdated)[fuRows]),
    sort(rownames(pottiUpdated)[
         fuTQNorm@p.values <= fuCut]);
> colnames(temp) <- c("Theirs", "Ours");
> temp
     Theirs         Ours
...
[3,] "1881_at"      "1882_g_at"
[4,] "31321_at"     "31322_at"
[5,] "31725_s_at"   "31726_at"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

Dec 4: Our next reports.

The next iteration: Numbers Sensitive



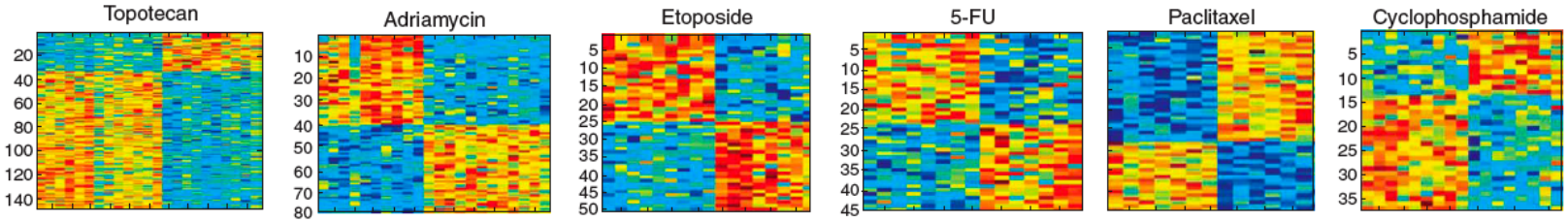




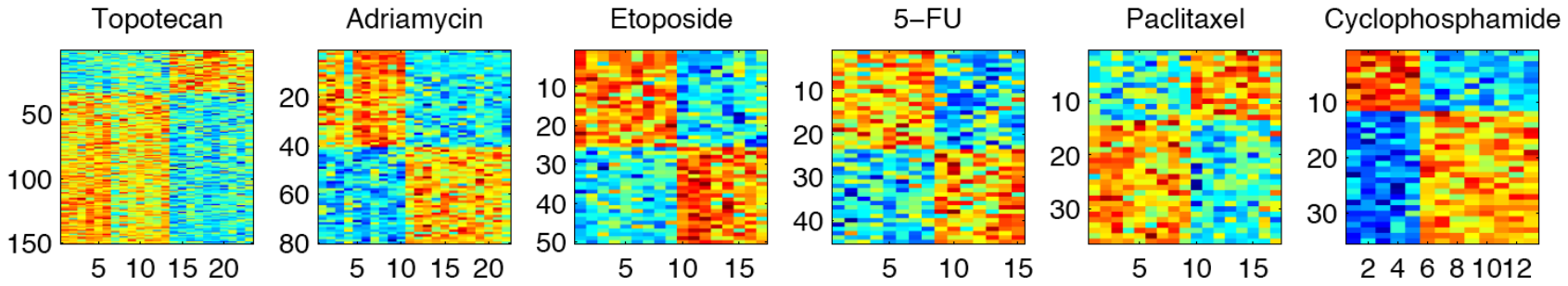
Dec 13: Our next reports.

The next iteration: New Data, Same Questions

From the paper:

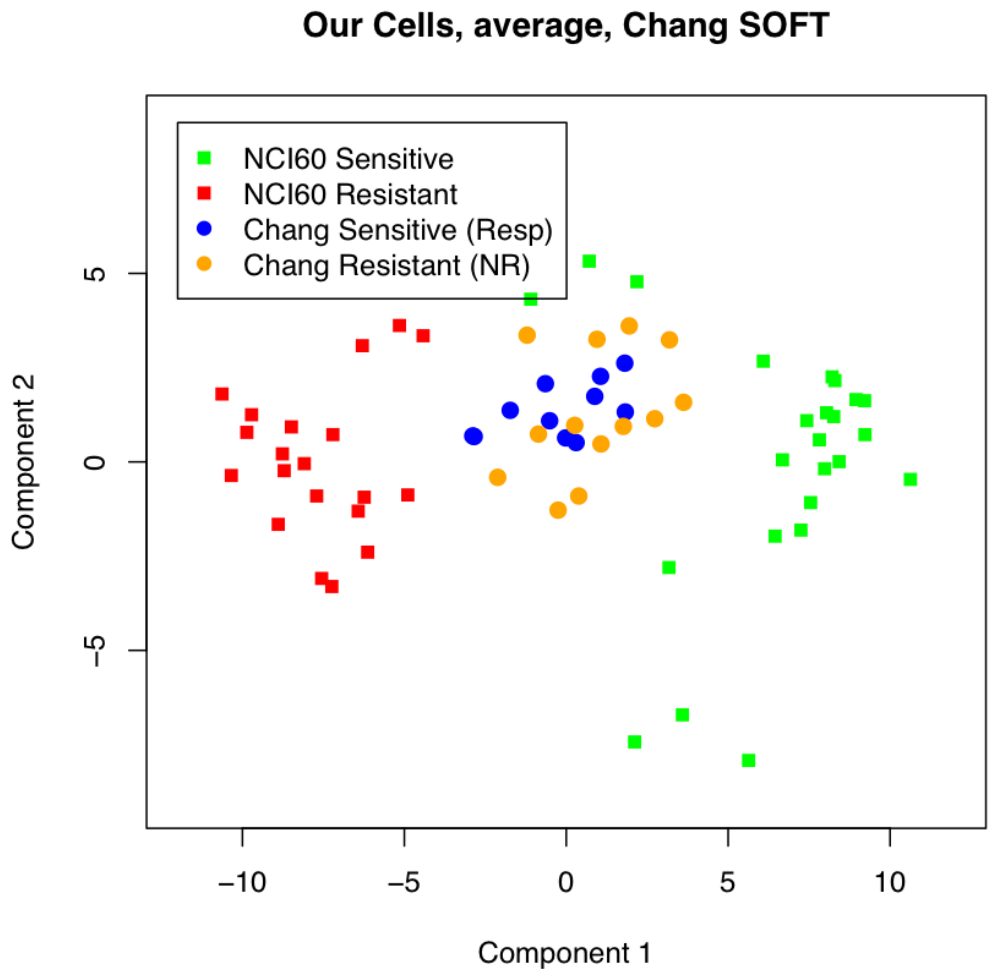


From the software:



Late Dec: New data. These didn't address our questions.
Dec 27: We ask again. Dec 30: We reproduce heatmaps.

The next iteration: Our Docetaxel Attempts



Jan 10, 2007: We try predicting docetaxel with their cell lines.

Further iterations

Jan 22: They acknowledge off by one errors.

Jan 24: We meet with Nevins at MD Anderson.

Feb 27: They post new data to address our concerns.

The data changed (28 cell lines for adria vs 22 before).

The rules aren't rules (detailed constructor mentions dropping the most sensitive cell line altogether).

Mid March: Predictions with randomly chosen cell lines do as well as predictions with those reported.

Active Disagreement

April 1: We tell them we don't think it works.

April 25: We show them a draft note to Nat Med.

May 1: They claim success, and suggest standardizing.

May 8: We note standardizing doesn't help.

May 16: We note the outliers are named in the paper.

May 26: They "bring this to a close", insist it works, and argue that the outliers are driven by stochastic noise.

May 30: We report the lists aren't stochastic.

A Matter of Timing

6 months after publication of their major article,

6 months after we started asking for clarification,

After the first clinical trial has begun,

They're sending inaccurate descriptions of how their method works.

They don't know how the method they're using works, but they're sure it does.

We Write to Nature Medicine

Jun 13: we submit our article to Nature Medicine

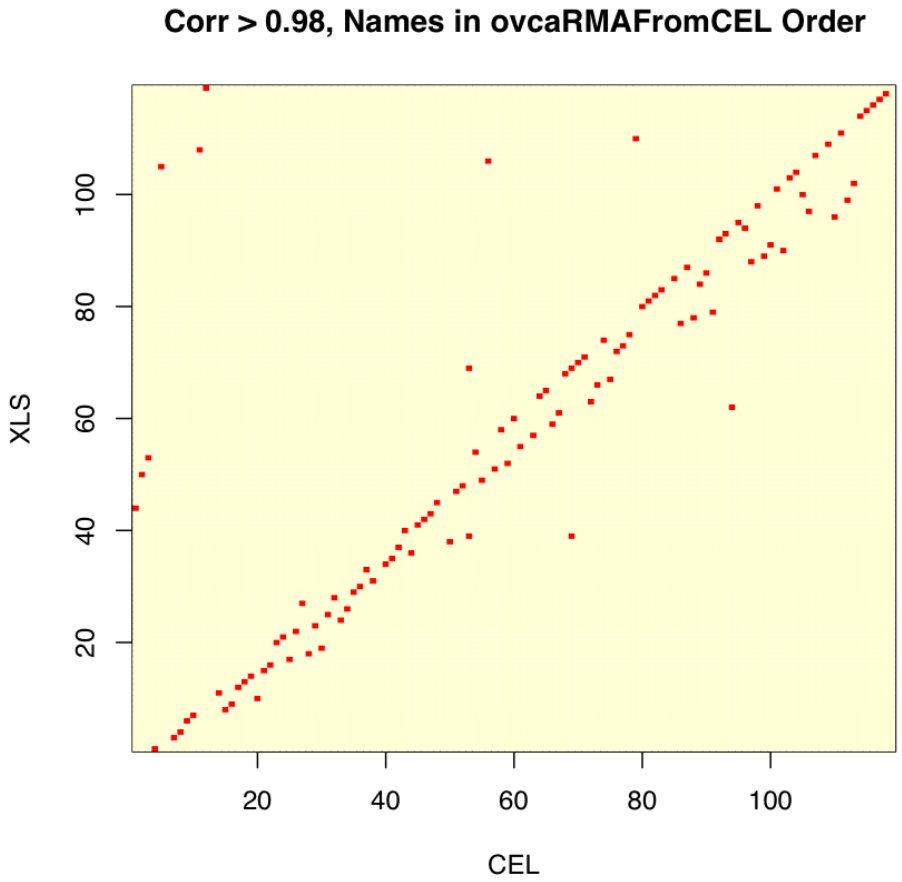
Jun 22: Nevins acknowledges lists aren't stochastic but still claims results hold.

The gene lists are "fixed".

The fixed gene lists are wrong, due to errors introduced by manual adjustment.

The gene list for cytoxan disappears.

Dressman et al (JCO, 2007): May-July



mid-May: We begin. About 3/4 of the data are mislabeled.
mid-July: We send reports to Dressman and Lancaster.

Hsu et al (JCO 2007): October

We match 41/45 cisplatin genes after dealing with off-by-one errors.

We can't match genes named to convey plausibility:

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

The last two probesets aren't on the U133A arrays that were used. They're on the U133B.

Second clinical trial begins.

JCO Submissions: November 5

Dressman et al:

Letter outlining problems with ovarian data – most pathway scores appear driven by batch effects.

End of November: Letter accepted.

Hsu et al:

Letter re problems with lung data, gene lists, heatmaps.

Mid-December: Letter rejected (no explanation).

We request clarification.

January, 2008: Second rejection (no explanation).

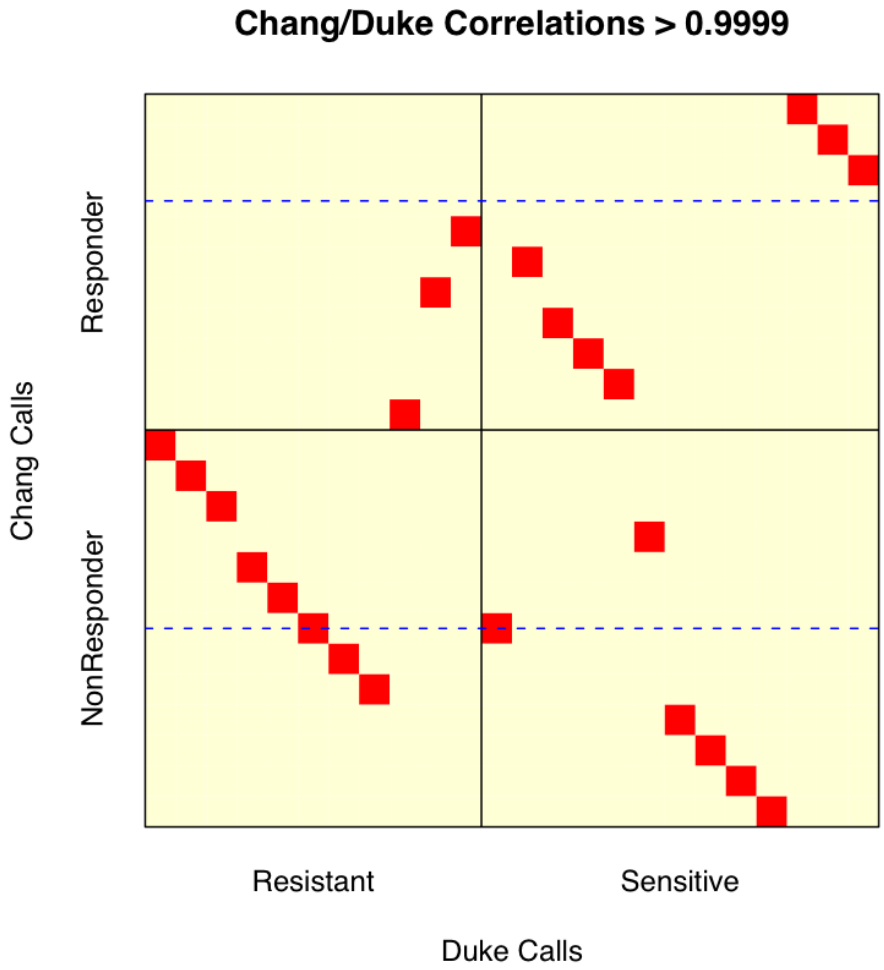
Nat Med Letter and Rebuttal: November 6

Assertions: We got it wrong, because:

1. We get their results when we use their methods
2. Their labels for docetaxel are correct (data posted)
3. Their labels for adriamycin are correct (data posted)
4. They've gotten it to work again (Hsu et al)
5. They've gotten it to work again, blinded (Bonnefoi et al, then in press)

No documentation or worked examples are provided.

All Assertions Are Wrong: November 7



We presented flaws at the NCI the next day.

Combination Therapy: Dec 2007-Feb 2008

Late Nov, 2007: Bonnefoi et al Lancet Oncology appears (Dec issue, early access).

We get data from GEO and European authors. Given the drug predictions, we can match all subsequent results.

For three drug combinations, the combination rules are all wrong, all different, and all unstated.

Late Jan, 2008: Reports sent to European authors. They refer us to Potti.

Feb 1: Questions posed to Potti.

Potti replies that they don't wish to indulge in another interchange with our group.

Ovarian letter and Rebuttal: Feb 2008

Assertions:

* We didn't do what they did.

* Because we didn't do what they did, we cannot comment on reproducibility.

* Sample (and clinical data) mislabelings are clerical errors that don't affect the results.

No documentation is provided.

Corrected array data is posted in July 2009, following separate inquiries by Carey and Stodden.

Communications and Letters: Feb-May 2008

We have extensive interchanges with Mauro Delorenzi, head bioinformatician among the European authors on the Bonnefoi et al Lancet Oncology paper.

Mauro seeks clarification from Duke about the drug prediction results, but is rebuffed.

May 30: Letter submitted to Lancet Oncology.

May 30: Letter submitted to Nature Medicine.

Third trial begins recruiting.

Rejections and "Fixes": Jun-Sep 2008

Jun 9: Lancet Oncology letter rejected: crux of issue is
"a statistical debate with no right or wrong answer".
"you can't have much better publicity than Nature Medicine"
"doesn't fill me with confidence as a referee"
"the withering criticism of your work by Potti et al"

Jun 11: Nature Medicine letter rejected: editors refer us to Nevins and Potti for clarification.

Aug: A correction from Potti et al. appears in Nat Med.
Old incorrect data is stripped from the web.
New incorrect data is posted.

Sep: An erratum appears in Lancet Oncology, correcting sensitive and resistant labels for cell lines.

Fallow: Through Sep 09

Many other papers in 2009.

Feb: Heatmap reuse noted. JCO, CCR alerted.

Apr: Incorrect correction posted to CCR.

Jun: We learn of clinical trials underway.

Jul: We circulate draft for informal comment: “too negative”.

Sep: Paper sent to *Annals of Applied Statistics*.
Published online.

Barriers we Encountered 1/4 (Data)

Data were never clearly provided.

Data were never clearly identified.

Clinical data were not supplied.

Data processing was not described.

Data changed over time.

There was no point at which the data were locked down (frozen) with a clear record of provenance.

Barriers we Encountered 2/4 (Questions)

Specific questions were unanswered.

Specific algorithms were not supplied.

Worked examples were not supplied.

Specific and documented objections were countered with assertions without evidence.

Authors were never required to substantiate their claims,
(a) to us,
(b) to the journals, or
(c) to Duke's internal review.

Barriers we Encountered 3/4 (Duke Review)

The Duke reviewers didn't verify provenance.

The Duke report wasn't published.

The Duke data weren't released.

Members of the Duke administration and IRB withheld information from the reviewers.

The review was neither complete nor transparent.

Barriers we Encountered 4/4 (Appeals)

Questions posed by the ORI: can you prove fraud? patient harm? This is what was required before they could get involved.

How long should we fight?

What was the NCI doing?

Where could we have gone next?

Interactions with Journals 1/2

Nature Medicine letter

Nature Medicine rebuttal

JCO rejection (lung)

JCO letter (ovarian)

JCO rebuttal

Lancet Oncology rejection

Nature Medicine rejection

Interactions with Journals 2/2

Nature Medicine correction

Lancet Oncology correction

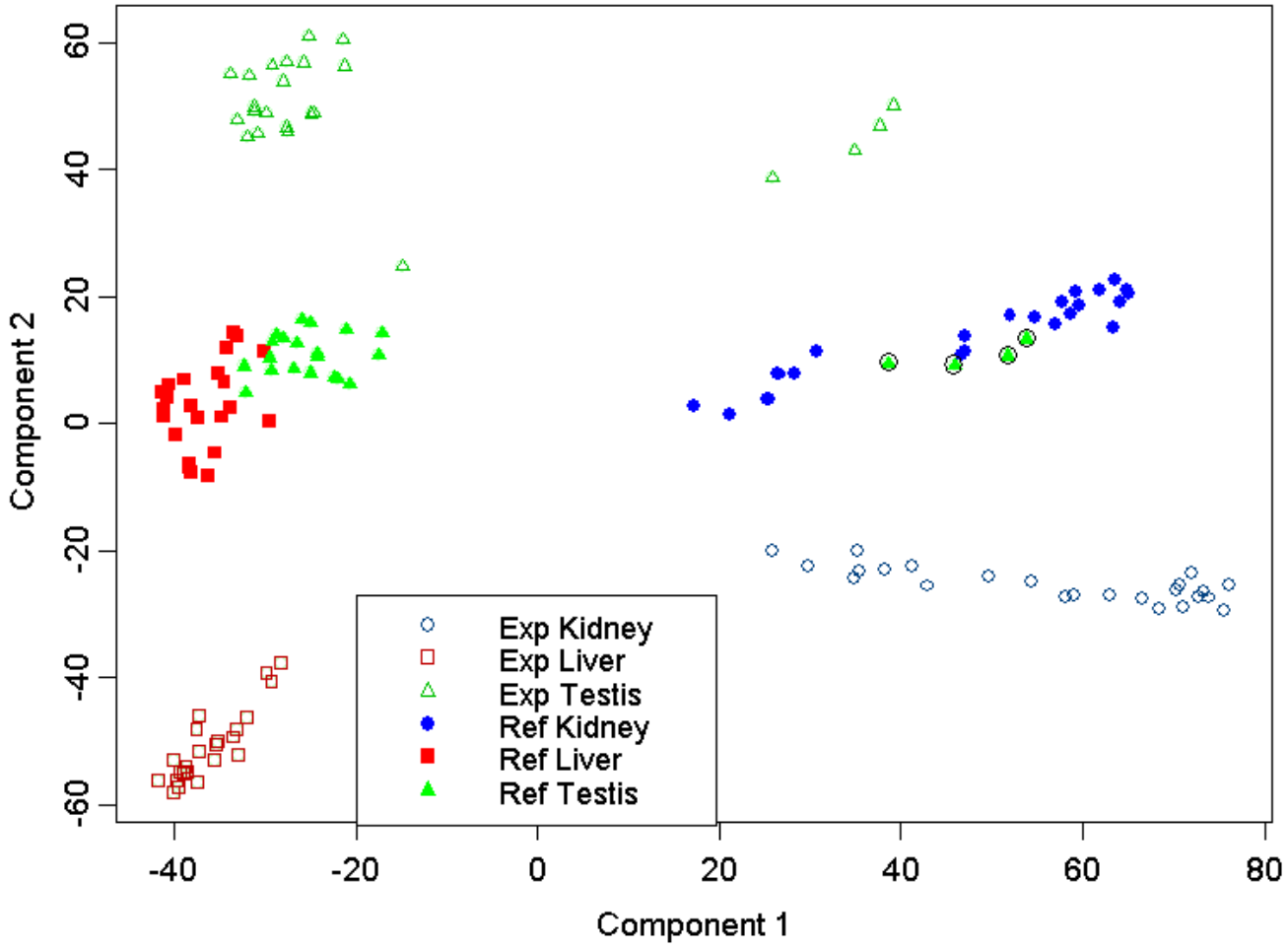
Clin Cancer Research correction

Nature Medicine and The Duke review

Lancet Oncology and blinding

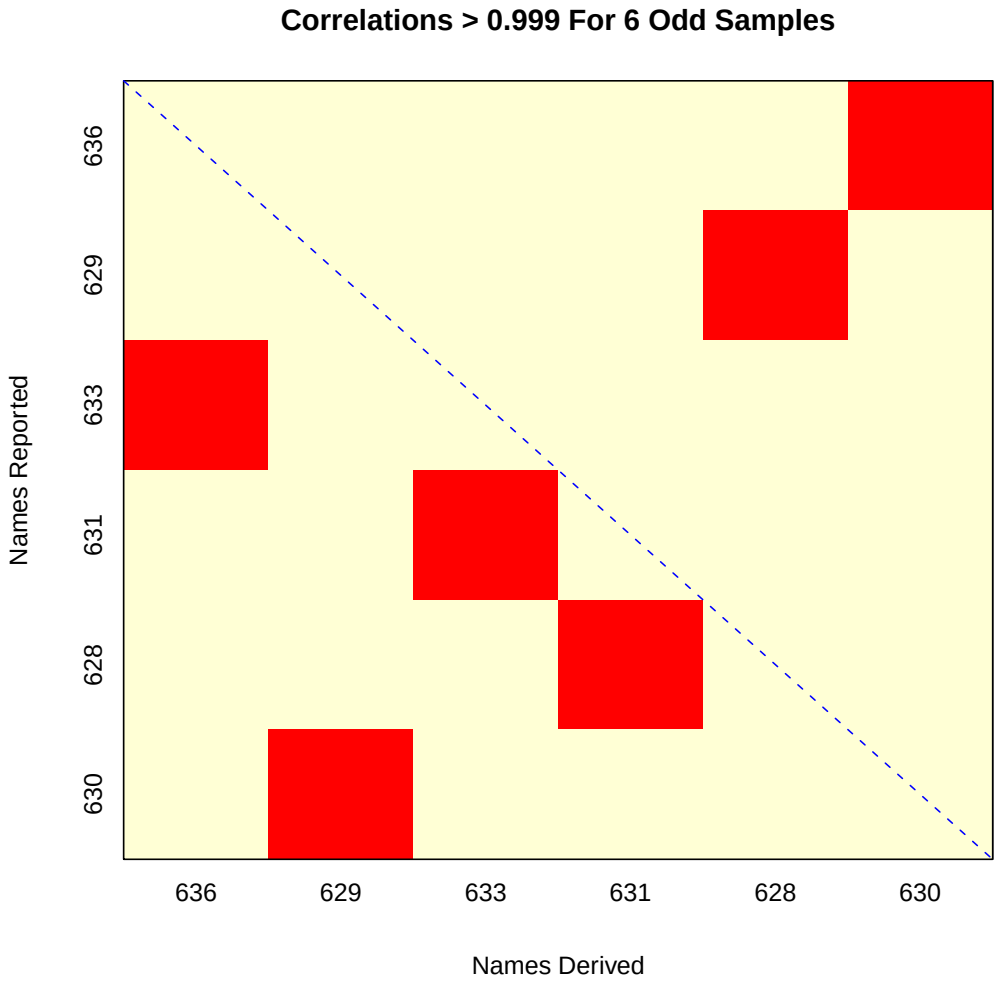
Reluctance of other journals to engage

Similarities: Gene Mislabeling



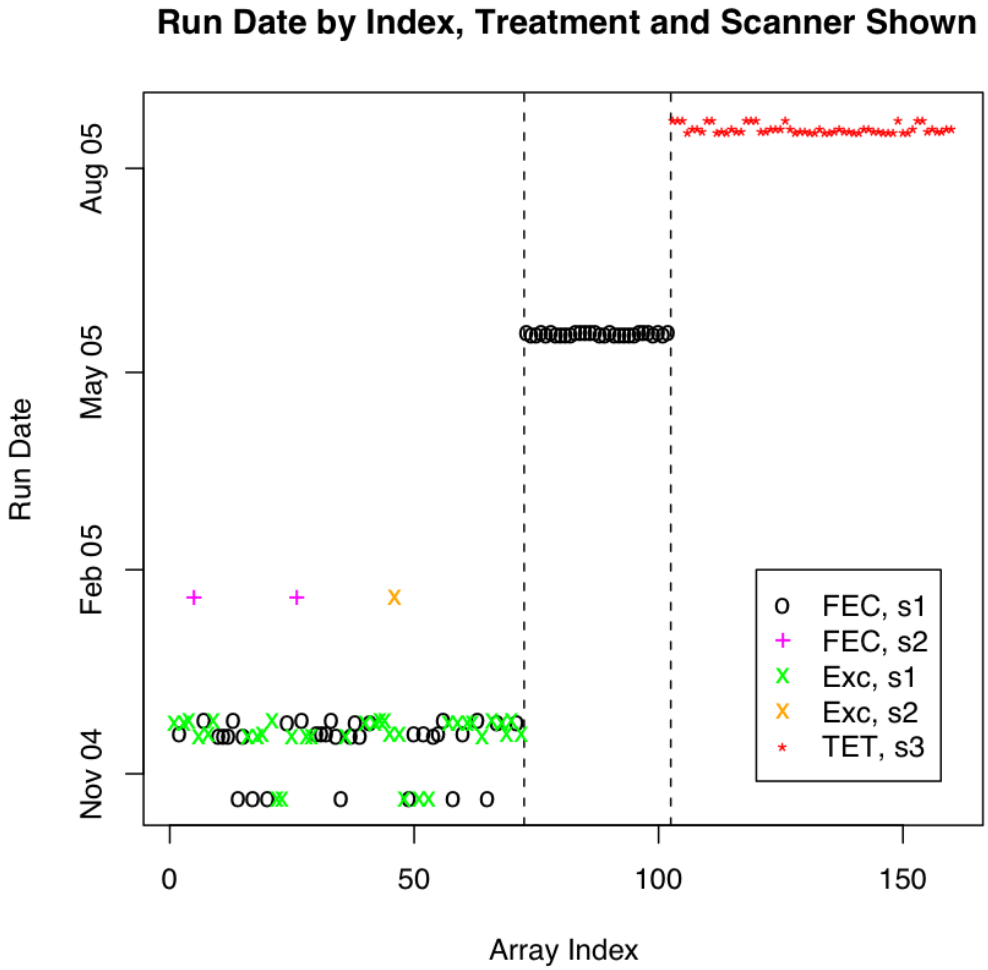
CAMDA 2002: One-row offset midway through collection.

Similarities: Sample Mislabeling



TCGA: Ovarian microRNA, 2010 (Batch 9, now fixed).

Similarities: Design Artifacts



Complete confounding – frequent occurrence.

Similarities: Data and Documentation

Ochsner et al., Nat Meth (2008):

Deposition rates at GEO and ArrayExpress for 20 journals requiring MIAME are below 50%.

Ioannidis et al., Nat Gen (2009):

Reproducibility for 18 quantitative array papers over two years (during MIAME compliance): success with 2, impossible even in principle for 10.

Individual problems of the types seen above are not uncommon.

Differences Setting this Case Apart

* Many more mistakes and papers were involved simultaneously.

* This made the story complex, hard to tell clearly, and easier to counter superficially.

* We encountered institutional pushback from Duke.

* There were repeated claims of blinded validation, which required contradiction by coauthors to counter (NEJM untestable).

The signatures progressed to clinical trials.

OvaCheck (pre-objections)

* Paper (Petricoin et al, Lancet, 2002) was very splashy, and aroused lots of interest.

* Data posted weren't what was analyzed.

* Data showed batch effects and design problems.

* The software wasn't properly used (calibration).

* The Lancet rejected our letter as "too technical for our readership".

* A clinical test was advertised.

OvaCheck (post-objections)

* We posted all of our data and code.

* The authors asserted we got it wrong, without proof.

* It was a protracted process involving several papers.

* Press coverage prompted action.

* The FDA stepped in.

* The NCI's BSA stepped in.

What Motivates Forensic Bioinformatics?

Our investigators came to us because they wanted to know if we could help implement these approaches at MD Anderson. Thus, we're driven by clinical relevance.

This is *required* because these checks take a lot of time.

What made (let) us stick with it?

Certainty we were right (simple pictures).

Security provided by being at a major institution.

Ethical issues.

Wanting to improve the process, and an optimism about omics approaches in general.

What are Our Recommendations?

We've outlined some in recent notes: Nature letter, Clin Chem editorial, ENAR notes

We need data.

We need metadata (clinical information, run order, design information).

We need evidence of provenance.

We need the code (MAQC II, NCI experience).

We need auditability before trials begin (Duke TMQF docs).

We need reproducibility.

How Do We Get There?

Investigators need to think of reproducibility as a goal from the outset.

Journals need to ask (and check) for code and data deposition (and be prepared to host code and clinical data).

Agencies need to provide data repositories. They need to check for data and code availability at renewal time. They need to budget for reproducibility audits.

Institutions need to help with training and infrastructure.

Can We Get There?

I think so.

* We've been writing reports in Sweave since 2007.

* Biostatistics has an AE for reproducibility.

* Open-source tools (e.g., git) can be adapted to this purpose.

* Other, more user-friendly tools exist.

* Victoria Stodden is teaching a course to statisticians on reproducibility even now.

* You're here.

I'm happy to take questions.