Microarrays: Retracing Steps (Again)

Keith A. Baggerly[1*] and Kevin R. Coombes[1]

[1] Department of Bioinformatics & Computational Biology

UT M.D. Anderson Cancer Center, Houston, TX

This work was partially supported by NIH/NCI grants P50-CA116199, P50-CA070907, and P50-CA083639.

[*] To whom correspondence should be addressed:

Keith A. Baggerly

Department of Bioinformatics & Computational Biology, Unit 237

UT M.D. Anderson Cancer Center

1515 Holcombe Blvd

Houston, TX, 77030

kabagg@mdanderson.org

tel: (713) 563-4290

fax: (713) 563-4243

Recently, Potti et al.[1] introduced a method for using cell line data to define signatures of drug sensitivity. By looking for these signatures in patient microarray profiles, they predicted response to several drugs. However, on reexamining the data, Coombes et al.[2] found several analysis errors and concluded that the approach didn't work.

In reply, Potti and Nevins[3] admit minor errors, contend others are irrelevant, and claim the Coombes et al.[2] analysis is flawed. They assert that data now posted on their website is correct, further note getting the approach to work again[4,5], and conclude that their method is reproducible and robust.

Data now posted includes processed data for docetaxel and adriamycin, a document describing cell line selection for docetaxel, and lists of cell lines used to assemble each drug signature. Unfortunately, examination of the new data and papers shows problems. We sketch these below; details are given in supplementary reports rep01-09.

1. Both test and training data for docetaxel are mislabeled. Potti et al.[1] predict docetaxel response using 24 test samples[6], roughly evenly divided between responders and nonresponders. They report getting 22/24 correct. However, mapping posted data to patient information shows that they misspecified responder/nonresponder status for 10/24 samples before modeling began (Figure 1A). One sample is omitted. Another is included twice, labeled both resistant and sensitive.

Training data for docetaxel involves 14 cell lines. The posted "Description of Predictor Generation" names 7 sensitive to docetaxel and 7 resistant. Matching numbers shows these lines are used with sensitive/resistant labels reversed. (Reports rep01-04.)

2. Test data for adriamycin is mislabeled. Potti et al.[1] predict adriamycin response using 122 test samples. The initially cited source[7] names 94 responders and 28 nonresponders, but Potti et al.[1] find 23 and 99, respectively. Coombes et al.[2] suggested labels might have been reversed. In reply, Potti and Nevins[3] double check the initial labels and find them correct; they claim data now posted clarifies sample sources. The adriamycin file now states "Validation data is from GSE4698, GSE649, GSE650, GSE651, and others". Sample columns are not named, but sensitive/resistant status is indicated for each. However, pairwise correlations show that only 84/122 test samples are distinct (Figure 1B). Some samples are included 2-4 times, with some labeled both ways (e.g., one sample is labeled a responder 3/4 times). The double-checked data is incorrect. (Reports rep05-06.)

3. Important genes are not derived from training data. Potti and Nevins[3] cite Hsu et al.[4], where the same methods are used to derive cisplatin and pemetrexed signatures. Hsu et al.[4] explicitly name ERCC1, ERCC4, and DNA repair genes as important components of the cisplatin signature. Using the Potti et al.[1] software, we can reproduce the cisplatin heatmap perfectly, but can't explain 4 of their 45 probesets: 203719_at (ERCC1), 210158_at (ERCC4), 228131_at (ERCC1), and 231971_at (FANCM, DNA repair). None have small p-values when sensitive and resistant cell lines are contrasted. This statement is vacuous for the last two probesets, as they're not on the U133A arrays used (they're on the U133B). ERCC1 and/or ERCC4 were likewise incorrectly included in gene lists initially reported by Potti et al.[1] for docetaxel, adriamycin, and paclitaxel, and flagged for attention there. (Report rep07.)

4. Cell line sensitive/resistant designations are reversed. Potti and Nevins[3] also cite Bonnefoi et al.[5], where response is predicted for combination chemotherapy. Combination components include taxotere (docetaxel), epirubicin, 5-fluorouracil, and cyclophosphamide. Bonnefoi et al.[5] list the cell lines designated sensitive and resistant to each drug. These match lists from Potti et al.[1], with status reversed throughout: lines called "sensitive" by Bonnefoi et al.[5] are called "resistant" by Potti et al.[1], etc. (Reports rep08-09.)

We don't understand how valid signatures can be derived from mislabeled data. We don't understand how important genes can be identified if not returned by the software used, or not measured. We don't understand how signatures can remain valid when directions are reversed. Consequently, we are not yet persuaded by the approach.

These analyses are complex. Reproducibility is important, since anyone can make mistakes. To help identify specific errors on our part, our reports, code, and data are at `http://bioinformatics.mdanderson.org/Supplements/ReproRsch-Chemo2`. Running these reports through the software program R[8] gives the results reported here.

References

1 Potti A, Dressman HK, Bild A, et al: Genomic signatures to guide the use of chemotherapeutics. *Nat Med*, **12**:1294-1300, 2006.

2 Coombes KR, Wang J, Baggerly KA: Microarrays: retracing steps. *Nat Med*, **13**:1276-7, 2007.

3 Potti A, Nevins JR: Reply to Microarrays: retracing steps. *Nat Med*, **13**:1277-8, 2007.

4 Hsu DS, Balakumaran BS, Acharya CR, et al: Pharmacogenomic strategies provide a rational approach to the treatment of cisplatin-resistant patients with advanced cancer. *J Clin Oncol*, **25**:4350-4357, 2007.

5 Bonnefoi H, Potti A, Delorenzi M, 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. *Lancet Oncology*, **8**:1071-8, 2007.

6 Chang JC, Wooten EC, Tsimelzon A, et al.: Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet*, **362**:362-369 (2003).

7 Holleman A, Cheok MH, Den Boer ML, et al: Gene-expression patterns in drug-resistant acute lymphoblastic leukemia cells and response to treatment. *N Engl J Med*, **351**:533-42, 2004.

8 R Development Core Team. *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna. `http://www.R-project.org`.

Figure 1: **A.** Mapping from Chang et al.[6] to Potti et al.[1]. Ten samples are mislabeled: 3 upper left, 6 lower right, 1 omitted. Potti et al.[1] note relabeling two (rows 20 and 22); these split with respect to status so the number wrong is unaffected. Dashed lines indicate one sample excluded (top) and another included twice, labeled both ways (bottom). **B.** High pairwise correlations between adriamycin samples: 22 training (lower left) and 122 test (upper right). Dots are only expected along the main diagonal. “Bands” show that only 84 test samples are distinct.