Genomic signatures based on the NCI60 cell lines do not predict patient response to chemotherapy

Kevin R. Coombes[1], Jing Wang[1], and Keith A. Baggerly[1]

[1]*Department of Bioinformatics and Computational Biology, University of Texas M.D. Anderson Cancer Center, Houston TX 77030 USA*

We have systematically attempted to use *in vitro* drug sensitivity data coupled with Affymetrix microarray data to develop genomic signatures that predict sensitivity to individual chemotherapeutic agents. By randomly selecting cell lines, we have found that signatures developed using the most sensitive and most resistant NCI60 cell lines do not perform better than chance on patient data. Thus, our result fails to reproduce a previous report in *Nature Medicine*. We discuss some of the reasons for this failure, and discuss methods for improving the reproducibility of analyses of large data sets.

Predicting whether a tumor will respond well to therapy remains one of the biggest opportunities in clinical oncology, allowing us to realize the promise of personalized treatment with improved outcomes and decreased toxicity. Recently, Potti and colleagues [1] published an article in *Nature Medicine* that appeared to offer a significant breakthrough on this front. Using publicly available data, they assembled microarray profiles from the NCI60 cancer cell lines with known *in vitro* sensitivity or resistance to a particular drug. They then used these profiles to predict *in vivo* chemotherapeutic response. They reported good success using this approach with seven drugs. Unfortunately, our group has been unable to reproduce their findings. The purpose of this paper is to examine the causes of this apparent irreproducibility, and to demonstrate a method for reporting a reproducible analysis of similar data.

Potti and colleagues broadly outlined a plausible strategy for trying to discover and validate genomic signatures of drug sensitivity. Conceptually, the plan is straightforward:

1. **Cell line selection:** Choose cell lines, using dose response data, to represent the extremes of sensitivity and resistance.
2. **Feature selection:** Select genes to include in a model, using microarray profiles of the chosen cell lines, by ranking the genes based on univariate t-tests between sensitive and resistant cell lines.
3. **Model training:** Train a probit regression model for predicting sensitivity on the

chosen cell lines, using principal components of the selected features as predictors.

4. **Model testing:** Test the model on an independent data set to determine if the cell line data can predict clinical outcomes.

The challenge is to report the implementation of these steps in sufficient detail that an independent reader can reproduce—or discover flaws in—the reported analysis. Potti and colleagues described their methods in words (rather than equations or computer code) in the published paper and in an online supplement. At each step, when we could not reproduce their results, we were uncertain if we had correctly interpreted their descriptions. We repeatedly contacted the authors and obtained clarification, but remained unable to reproduce their results. For our own analysis, we have made every effort to provide unambiguous descriptions. In addition, in case our written descriptions fail, the full source code for each analysis step is available so that other researchers can reproduce and evaluate our methods carefully—and, we hope, improve upon them.

Ultimately, we believe that the methods we used are a reasonable interpretation of the ones presented by Potti and colleagues. The failure of these methods suggests that the approach used to interpret genomic signatures based only on the NCI60 cell lines cannot be successfully applied in this fashion to predict patient response to chemotherapy.

RESULTS

The GI50 values of their sensitive and resistant cell lines overlap (cell line selection).

In their original paper, Potti and colleagues did not report which cell lines were used to define individual drug sensitivity signatures. But, in response to inquiries, they kindly posted additional information on their web site. For docetaxel, they found seven sensitive cell lines:

```
HL-60(TB), HOP-62, HT29, NCI-H522, SF-539, SK-MEL-2, SK-MEL-5
```

and seven resistant cell lines:

```
786-0, CAKI-1, EKVX, IGROV1, OVCAR-4, SN12C, TK-10
```

We have used these cell lines with their software to reproduce the docetaxel heatmap in their paper (Supplementary Report SR9), confirming that these are the cell lines used in their analysis.

Here, we are trying to understand how they chose those cell lines, and whether the method is reproducible. At the *Nature Medicine* web site*, Potti and colleagues posted Supplementary Methods (P-SM). In that document, they say: "[W]e chose cell lines ... that would represent the extremes of sensitivity to a given chemotherapeutic agent (mean GI50 ± 1 SD).... [T]he log transformed TGI and LC50 dose ... was then correlated with the respective GI50 data.... Cell lines with low GI50 ... also needed to have a low LC50 and TGI...."

To apply this description, we downloaded the dose response data from the DTP web site (see Methods and Supplementary Report SR1). The web site contains data on three sets of dose response experiments for docetaxel. These experiments differ based on the (maximum) starting concentration used, which was either 10^{-4} M, 10^{-6} M, or 10^{-7} M. For each of the three starting concentrations, we began with cell lines whose GI50 values were 1 SD above or 1 SD below the mean. We only kept cell lines for which the LC50 and TGI values were both above (resp., below) their median levels whenever the GI50 value was above (resp., below) the mean ± 1 SD cutoff. Using this interpretation, we found no cell lines that were resistant to docetaxel and only one cell line (`COLO 205`) that was sensitive. If we weakened the criteria to allow the inclusion of cell lines for which the TGI or LC50 was equal to its median value, then more cell lines were found but none of the experiments produced lists of cell lines that matched the ones reported by Potti and colleagues (Supplementary Report SR3).

To improve upon this result, we looked more carefully at the drug response data. The two lower starting concentrations provided no useful information about the LC50 values of the cell lines. Moreover, 41 of the 59 cell lines tested have LC50 values greater than the highest starting concentration of 10^{-4} M (SR3). As a result, we decided to ignore the LC50 data for docetaxel. Similarly, the lower starting concentration experiments were not useful for estimating TGI (SR3). So, we used the experiment with starting concentration 10^{-4} M to estimate TGI values for docetaxel. Both of the lower concentration experiments appeared to give reasonable (although not perfectly congruent) estimates of GI50 values for the response of cell lines to docetaxel. We averaged these estimates to provide pooled estimates of GI50.

Although the GI50 values appeared to be normally distributed, the TGI values were highly skewed. Consequently, we chose to work with distribution-free descriptions (i.e., me-

*http://www.nature.com/nm/journal/v12/n11/extref/nm1491-S9.pdf

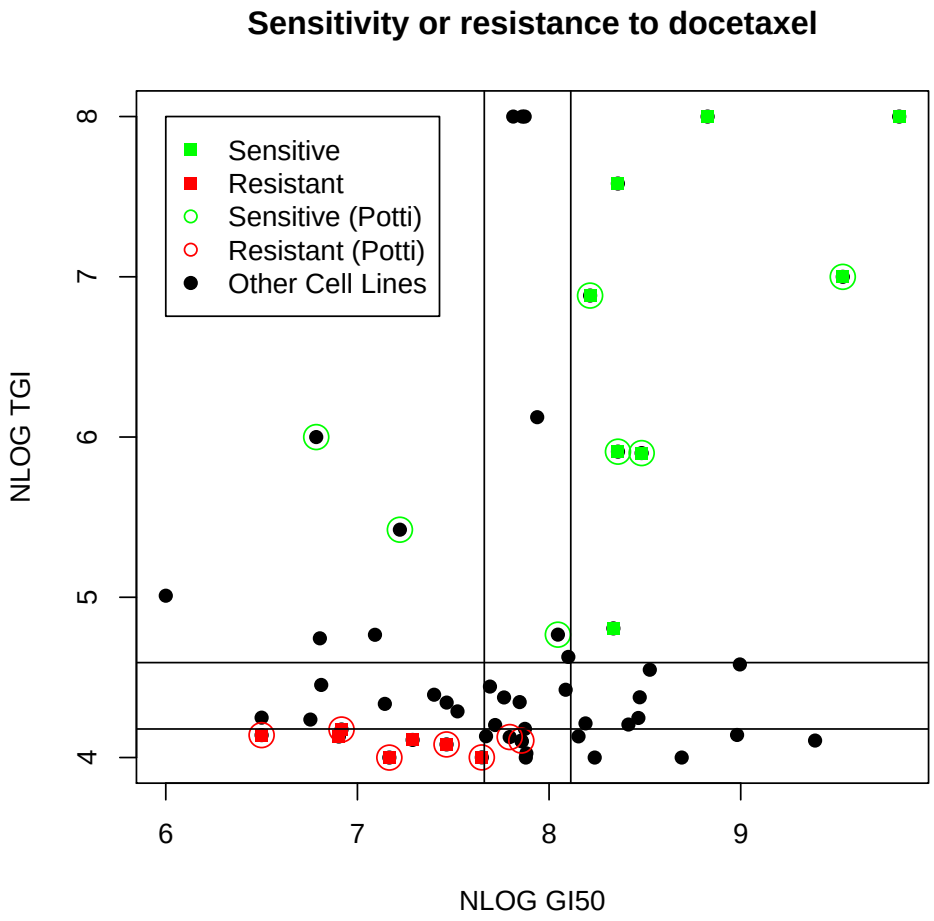

Figure 1: Scatter plot of the negative base-ten logarithm (NLOG) of the GI50 and TGI values for 59 cell lines. The observed values were separated into thirds. Genes selected as sensitive (resp., resistant) were in the highest (resp., lowest) NLOG concentration third on both measures.

dian and quantile) rather than parametric descriptions (e.g., mean and standard deviation). Specifically, we divided both the GI50 data and the TGI into thirds (Figure 1). Cell lines that were in the highest third for both negative base-ten log (NLOG) TGI and NLOG GI50 concentrations were called sensitive, and cell lines that were in the lowest third for both NLOG TGI and NLOG GI50 were called resistant. Using these definitions, we found that eight cell lines were sensitive:

`COLO 205, HCC-2998, HL-60(TB), HT29, MDA-MB-435, NCI-H522, RPMI-8226, SF-539`

and seven cell lines were resistant:

`786-0, ACHN, CAKI-1, EKVX, IGROV1, OVCAR-4, SF-268`

We found similar results (12 or 13 of the 15 selected lines) using the unpooled GI50 values from the experiment with starting concentration 10^{-6} M or 10^{-7} M to select cell lines.

While our approach generates lists of resistant and sensitive cell lines that are similar in size to the ones selected by Potti's group, we have poor agreement with the actual lines they selected, with only 4 or 5 of 7 overlapping. In Figure 1, the NCI60 cell lines are plotted according to our TGI and GI50 values. Our cell lines are indicated in color and the Potti group's cell lines are circled. Two of the lines they called sensitive have smaller NLOG GI50 values than half of the lines they called resistant. In addition, the most sensitive line (COLO 205), based on both GI50 and TGI scores, was not included in their set.

Differentially expressed probe sets vary depending on the microarray data set (feature selection).

Novartis ran each cell line on three Affymetrix U95Av2 microarrays (see Methods). These three sets of replicates can be grouped into series "A", "B", and "C" using the labeling supplied. Examining the data posted on the Duke web site, we determined that Potti and colleagues selected features from the A series of Novartis experiments (Supplementary Report SR2), where "a variance fixed t-test was used to calculate significance" (P-SM).

We selected genes that were differentially expressed between the sensitive and resistant cell lines that we chose in the previous section. We used two-sample t-tests to analyze the

A, B, and C series, both separately and jointly. In the joint analysis, we first averaged the replicates and then performed a t-test on the averages. For every analysis, a beta-uniform mixture (BUM) model of the p-values[2] showed evidence of substantial differences in gene expression between the docetaxel-sensitive cell lines and the docetaxel-resistant cell lines. In their published analysis, Potti and colleagues selected the 50 most significant genes to use in model building. The number of genes they selected varied for different chemotherapeutic agents, and criteria for choosing these numbers were not given. Following their lead, we selected the 50 most significant genes from each analysis.

The sets of top 50 genes varied from data set to data set; the numbers of genes in the intersections are listed in Table 1. Using the BUM model to estimate the false discovery rate (FDR)[2,3], we found that the 50 genes from the series A experiments correspond to FDR = 4.0%; series B, FDR = 13.0%; series C, FDR = 28.4%; and the average analysis, FDR = 7.9%. As a result, we decided to use both the list of 50 genes selected from the average data and the 50 from Series A for our further analysis.

We repeated this analysis using the cell lines chosen by Potti and colleagues. The variability of the gene lists was comparable (Supplementary Report SR4). Potti and colleagues also provided lists of the features selected for each drug treatment, indexed by Affymetrix probe set ID and annotated with gene names, symbols, and descriptions. These lists are available on the *Nature Medicine* web site as Supplementary Table 1. The lists as initially reported are wrong, because of an off-by-one indexing error that we discovered (Supplementary Report SR9). After correcting for this error, their list of genes for docetaxel has 29 genes in common with the Series A list we derived using the cell lines that they chose, and 33 of their 50 reported genes have small enough p-values that minor changes in the normalization procedure could account for the difference. The remaining 17 genes, however, have large p-values in our analysis, and we cannot explain how they were selected.

The first principal component suffices to separate resistant from sensitive cell lines (model training).

In P-SM, Potti and colleagues wrote: "The individual drug sensitivity and resistance data from the selected solid tumor NCI60 cell lines was then used in a supervised analysis using binary regression methodologies ... to develop models predictive of chemotherapeutic response.... Each signature summarizes its constituent genes as a single expression profile, and is here derived as the top principal component of that set of genes."

We used singular value decomposition (SVD) to perform principal component analysis

(PCA) on the cell lines and features that we selected, using all replicates in the Novartis data. We used an implementation of the algorithm in version 1.3 of the `ClassComparison` package that is part of a suite of tools for Object-Oriented Microarray and Proteomic Analysis (OOMPA) in R, which we developed and which is available from our web site[†]. Based on a plot of the first two principal components (Figure 2), the first principal component by itself is more than adequate to completely separate the resistant from the sensitive lines. This finding is consistent with the description of the methods in the paper by Potti and colleagues.

We then built a binary probit prediction model using all components *except* the first to try to predict sensitivity in the selected NCI60 cell lines. None of the higher components were statistically significant; the smallest individual p-value was 0.292 (Supplementary Report SR5). We then built another predictive model, including the first principal component, and using the Akaike Information Criterion (AIC) in a step-wise procedure to select the best model incorporating multiple principal components. Only the first principal component was included in the model (SR5). We repeated this entire analysis using features selected from the Novartis Series A experiments, and also did the same thing using the cell lines chosen by Potti and colleagues. The results were comparable (SR5).

The predictions do not validate on a clinical breast cancer data set (model testing).

Potti and colleagues wrote: "Chang and colleagues have published expression ... data and objective response information to docetaxel. Of the 24 patients reported in their study, there were 13 patients with docetaxel sensitivity and 11 patients with resistance. This dataset was used to validate the in vitro predictive model and generate a complementary in vivo model.... Gene selection and identification is based on the training data, and then metagene values are computed using the principal components of the training data *and additional cell line or tumor expression data*. Bayesian fitting of binary probit regression models to the training data then permits an assessment of the relevance of the metagene signatures" (P-SM; emphasis added).

We downloaded the Chang breast cancer data set (GSE349 and GSE350) from the Gene Expression Omnibus (GEO) web site (Supplementary Report SR6). Interestingly, the Chang paper[4] states that there were 13 resistant and 11 sensitive patients (which is the opposite of the numbers used by Potti). Moreover, the data in GEO seems to contain 14 resistant and 10 sensitive samples. Susan Hilsenbeck (personal communication) has informed us that one

[†]http://bioinformatics.mdanderson.org/software.html

sample (#377, GSM4913) was misidentified as resistant when uploaded to GEO, confirming the numbers from the original article.

The Chang data was originally processed using DNA Chip Analyzer (dChip) using the $PM - MM$ algorithm developed by Li and Wong.[5] We processed the CEL files ourselves, using `dchip2006.exe` and the PM-only model. We performed quantile normalization to map the feature intensity distributions in the Chang data onto the same quantiles used to normalize the NCI60 cell line data. We then projected the breast tumor samples from the Chang study onto the principal component (PC) space defined by the docetaxel sensitive and resistant NCI60 cell lines (Figure 2). The tumor samples were projected into the center of the PC space (largely intermediate between the sensitive and resistant cell lines). Moreover, the sensitive tumor samples almost completely overlap the resistant tumor samples, showing no signs of separation and being effectively randomly distributed. This figure suggested that no prediction method based on the first (or even the first and second) principal component from the cell line data could possibly make accurate predictions on the breast tumor data.

To test this, we applied the predictive probit binomial models based on (i) just the first and (ii) the optimal set of principal components chosen from the NCI60 data using AIC. Both results were the same, and were not very convincing (Table 2). We performed similar analyses using models based on all combinations of (a) the cell lines we chose or the cell lines Potti and colleagues chose; (b) features selected from the Novartis A arrays or from the averaged Novartis data; and (c) using our quantifications of the Chang CEL files or the posted quantifications from GEO. None of these eight variants produced results comparable to the ones reported in the paper by Potti and colleagues (Supplementary Report SR7).

The predictions are no better than those made using random cell lines.

At the heart of the paper by Potti and colleagues is the hypothesis that selecting cell lines that represent the extremes of sensitivity and resistance by multiple measures should make it possible to discover genomic signatures of chemotherapeutic response. This hypothesis can be tested by selecting the same number of cell lines randomly, arbitrarily labeling them as "sensitive" or "resistant", and then applying the same methods to discover and validate signatures.

We performed this random cell line selection 200 times, selecting 7 cell lines to call resistant and 7 to call sensitive. The results are displayed in Figure 3, which is analogous to the usual plot of an ROC curve. We can summarize the performance of models derived from each set of random cell lines using the average of the two proportions, which can be

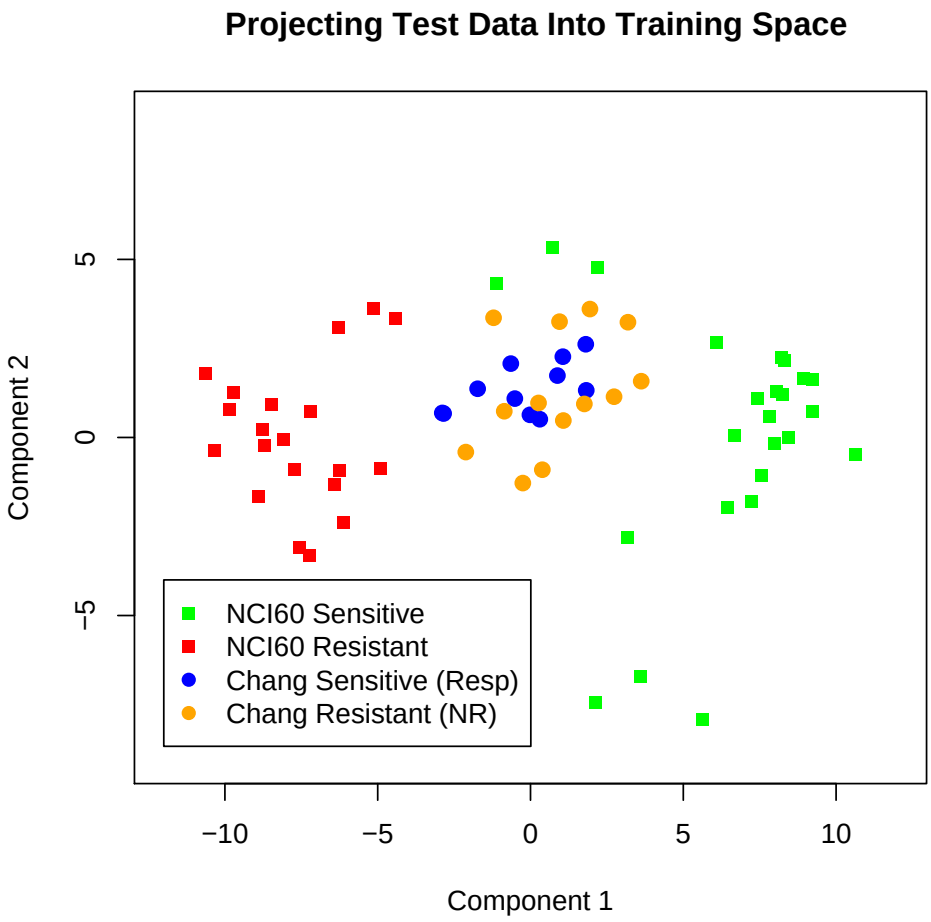

Figure 2: Plot of the first two principal components from the NCI60 training set, into which the Chang validation set has been projected.

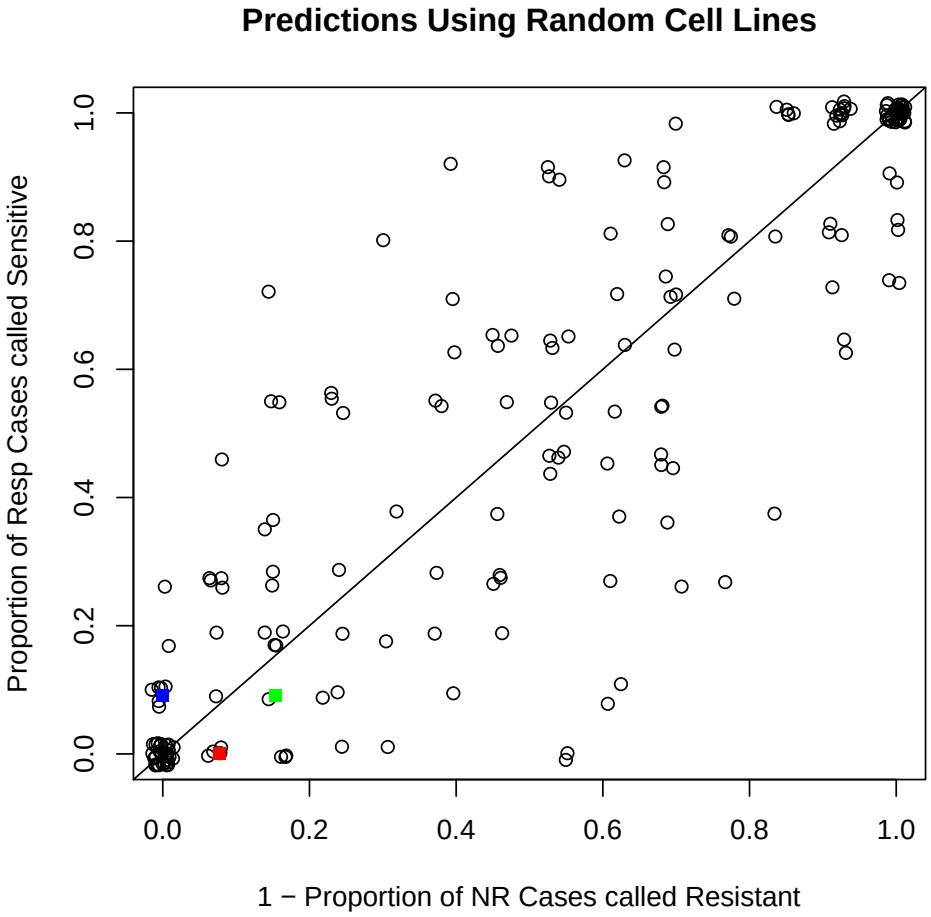

Figure 3: Prediction results using random cell lines to represent "sensitive" or "resistant" cases. True outcomes (Resp= responder, NR=non-responder) are known for the Chang breast cancer samples. Using the cell lines chosen by Potti and colleagues, we marked the performance of the model based on features from the Novartis A data (red square), features from the average Novartis data (green square), and the features they reported (blue square). The indicators of actual performance are concentrated in the lower left of the figure.

interpreted either as the expected accuracy or as the area under an ROC curve (AUC) on which we have only observed one point. The fraction of random data sets with larger AUC values provides an empirical p-value for the observed performance. In the case of features selected using the Novartis A data on the cell lines chosen by Potti and colleagues, this p-value is 0.74 (Supplementary Report SR8).

Their software may use information from the test set while training the model.

We were puzzled by the disparities between our findings and those reported by Potti and colleagues. We were also concerned by the phrase emphasized above, which suggested that information from "additional cell line or tumor expression data" was used during training. To understand what this meant, we reviewed the MATLAB source code from the software that Potti and colleagues posted on the Duke web site. This code performed SVD on the training and test data combined. The Duke group also provided a newer experimental version of the code, which performed SVD on the training data alone, as we did above.

We first ran both versions of their software using just their reported cell line training sets for the 7 drugs studied in their original paper. The results were the same for both versions of the software. We perfectly matched 6 of the published heatmaps; the exception was cytoxan which we could not match at all. We also perfectly matched the list of reported features (after correcting for the off-by-one error) for three drugs (5-fluorouracil, topotecan, and etoposide) and matched 75/80 for adriamycin, 28/35 for paclitaxel, and 31/50 for docetaxel (SR9). We cannot explain the disparities in the feature lists, since the accompanying heatmaps are identical. Of the 19 unexplained genes for docetaxel that appear on their reported list, 14 are listed as useful discriminators in the supplement to the paper by Chang and colleagues.[4]

We then tried to reproduce their predictions for the Chang test data using their selected Novartis A cell line data for training (Figure 4). The top panels of the figure show the two principal components that are most significant in a model built to separate sensitive (blue) from resistant (red) cell lines. The bottom panels show the predicted probability of resistance on the test data.

When only the training data is used in the SVD (experimental software; right panels), the first principal component (Factor 1, y-axis) is the most important factor for separating the two groups. However, the predictions on the test data for this model put all the samples in the same category, providing no power to separate responders from non-responders. These findings are consistent with our own analysis presented above.

By contrast, when both training and test data are used in the SVD (original software; left panels), the second principal component (Factor 2, y-axis) becomes the most important. Using this model, the predictions appear to separate the test samples. Performing the SVD on the joint data set has produced a drastically different model. The model apparently changes because information from the test samples "leaks" into the model during fitting.

DISCUSSION

Niels Bohr reportedly quipped that "Prediction is hard, especially when it involves the future." In the realm of medical applications of microarrays, we believe that he has a valid point. Developing a predictive model from one microarray study and applying it successfully to an independent microarray study is very difficult. The difficulty arises, in part, because the analysis is inherently complex, requiring a complicated sequence of steps with numerous choices of algorithms and parameters at each step. These analyses are also extremely *fragile*, in the sense that a single error at any one of the steps can invalidate the conclusions. Of course, complexity by itself need not lead to fragility. Living cells, for instance, are highly complex, but they manage to respond succesfully to rapidly changing environmental conditions. Cells and organisms rely on feedback loops, alternate pathways, and homeostasis to achieve a level of robustness that appears to be lacking, as yet, in the analysis of large data sets.

In order to provide feedback on analyses, published results must be reviewed with an eye toward their reproducibility. In the present instance, we found that relying on the written description of the methods, either in the published paper or in the online Supplementary Methods, compounded the difficulties. When we could not reproduce the published results, it was initially unclear if those results were wrong or if we were simply misinterpreting or misunderstanding the descriptions of the methods. We repeatedly contacted the authors and obtained clarification, but were still unable to reproduce their results.

For our own analysis, we have taken what is, perhaps, an extreme view on reproducibility. Our analysis was performed using Sweave, a package that allows analysts to combine the source code (in R, a statistical programming enviroment[6]) and the documentation (in LaTeX, a software tool for text preparation[7]) in the same file. Our source code is freely available; anyone can download it and run it. Moreover, running the code not only reproduces the results; it also generates the figures, tables, and a complete PDF version of this manuscript.

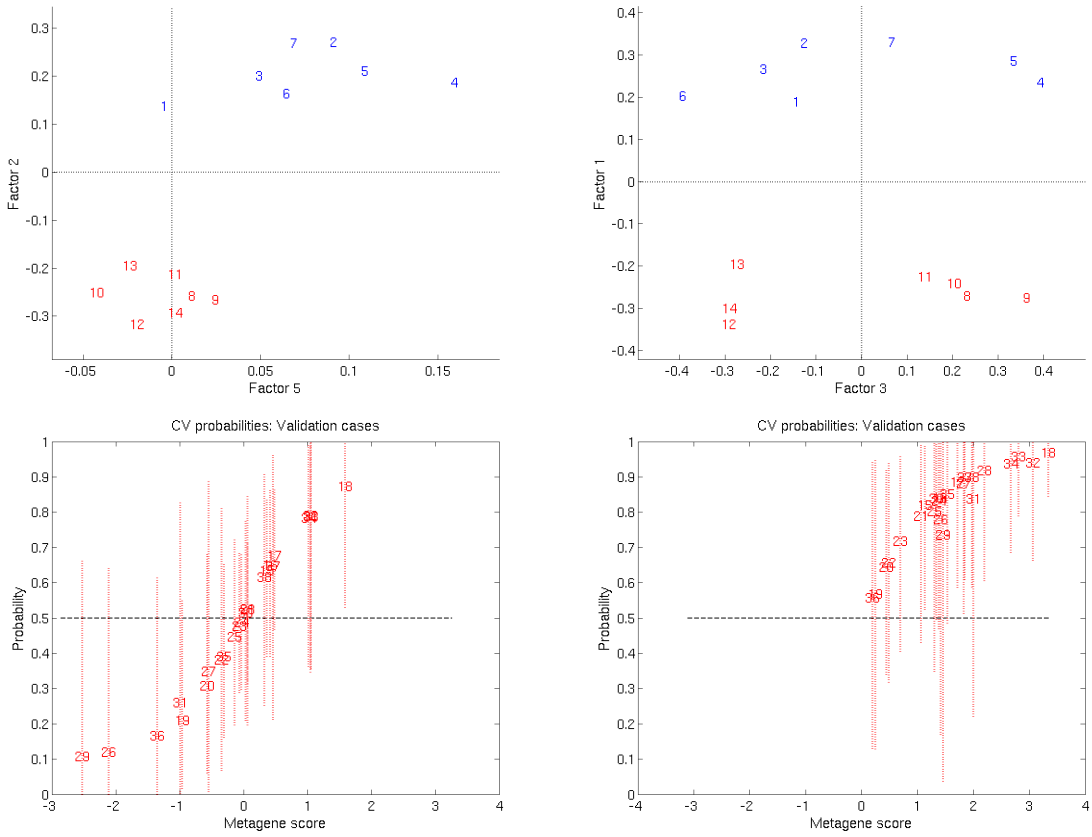

Figure 4: PCA plots for docetaxel training data (top; blue = sensitive, red = resistant) and prediction probabilities on test data (bottom). Panels on the left are from the original version of the software, which performs SVD on the combined training and test data. Panels on the right are from an experimental version of the software, which performs SVD only on the training data.

The results reported in the paper are backed by supplementary reports prepared in Sweave, amounting to 9 PDF files containing about 127 pages of text, code, figures, and tables. The concept of "reproducible research", which advocates that researchers provide complete source code in this way, is beginning to attract attention in many areas of science.[8–12] We have become converts to this philosophy, and see this article as a concrete demonstration that it is possible to provide a complete, fully detailed analysis of a complex microarray study. There have been others.[13,14]

Because Potti and colleagues were not only very responsive to our questions but also made the source code of their analysis available, we were ultimately able to detect what we believe to be flaws in their analysis that may partially explain the disparity between their findings and ours. We suspect that they may have used the original version of the software, which performed PCA using SVD on the combined training and test data. If this is the case, then the independence of the test set was not maintained. The fact that using the test data in this way changes the model is illustrated starkly in Figure 4.

One would expect to see three main sources of variation in the combined cell line (training) and patient (test) data: differences between sensitive and resistant cell lines, differences between responders and non-responders among the patients, and systematic differences between the training set and the test set. This is exactly what happens when the chosen Novartis cell line data are combined with the Chang breast cancer data (SR9). In fact, the first principal component, which measures the largest source of variation, appears to be primarily driven by the differences between the cell line training data and the patient tumor test data. This observation explains why different principal components are significant in Figure 4.

Because of random noise, one would also expect each of the first three PCs to be a combination of the three main sources of variation. As a consequence, information about how to separate the test data may "leak" into the PC that best separates the training data. This effect would be particularly pronounced if genes that separate the test set but not the training set were accidentally included in the input to the SVD. As noted earlier, only 31 of the 50 genes that Potti and colleagues reported to separate docetaxel-sensitive from docetaxel-resistant cell lines can be reproduced using their software. Moreover, 14 of the other 19 genes had previously been reported to have discriminatory ability on the Chang test data. If the SVD were performed on the training set alone, then all 19 genes would get essentially zero weight in the PCs. However, performing the SVD on the joint data would give nonzero weights to the 14 genes that had discriminatory power on the test set, thus inappropriately including them in the PC that separates the training set.

Because these combinations can be driven by random noise, it is also possible that one can get "better than chance" predictions that point the wrong way. This appears to have happened in the paper by Potti and colleagues. Their Figure 2c, center panel, contains predictions for a set of pediatric patients with acute lymphocytic leukemia (ALL) treated with "adriamycin" (actually, daunorubicin; see GEO datasets GSE650 and GSE651, which Potti and colleagues[1] list as their data source on page 1296). That figure shows 99 resistant samples and 23 sensitive samples. The paper that originally reported on these samples[15], consistent with our knowledge about the success rate for treating pediatric ALL, claims that there are 28 resistant and 94 sensitive samples.

The important question, however, is not whether the analysis by Potti and colleagues was flawed, but whether it is possible to learn genomic signatures of chemoresponse from the NCI60 cell lines and apply them to predict which patients will respond to chemotherapy. In this article, we have shown that a specific analytical approach does not work. Other approaches, which may use a more sophisticated method for feature selection or an alternative algorithm for training models, might conceivably work. Critically, we have also employed a method for testing other approaches. The method of comparing cell lines chosen based on the dose response data to random cell lines can be used to compute empirical p-values for any statistical method that claims to build predictive models.

We have not shown that it is impossible to take signatures from any cell lines and apply them to human samples. We do, admittedly, find it biologically implausible that a signature derived from a relatively small set of cell lines (like the NCI60) that spans numerous tissue types could produce a robust chemosensitivity signature that would be visible above the variability arising from the heterogeneity of tissue origins. We do think it possible, however, that a signature derived from a large panel of non-small-cell lung cancer cell lines could be relevant for predicting response in lung cancer patients, for example.

METHODS

Public data sources. Table 3 lists the data sets that were used by Potti and colleagues, along with links to the web sites where they could be located as of December, 2006. Note that the web site at `http://data.cgt.duke.edu/Combo1.php`, which is referenced in the supplementary material on the Nature Medicine web site, has since been removed, and different files have been posted at `http://data.cgt.duke.edu/NatureMedicine.php`. In this article, we used (i) individual array data from replicated Novartis experiments on the NCI60 cell lines using Affymetrix U95Av2 microarrays as training data, and (ii) the neoadjuvant breast tumor data set from the *Lancet* article by Chang and colleagues[4] as test data. Note

that the Novartis data set, as of December 2006, contains an error: the data for probe set "100_g_at" is duplicated; we removed the duplicate before starting our analysis (SR1).

We also used the summary table data on 50% growth inhibition (GI50), total growth inhibition (TGI), and 50% lethal concentration (LC50) from the website of the Developmental Therapeutics Program (DTP) at the National Cancer Institute (NCI). The drug response data from this source is indexed by NSC number, not by the name of the compound. Table 4 lists the NSC numbers and the names of the drugs studied by Potti and colleagues. In this report, we focus on docetaxel (taxotere), whose NSC number is 628503.

Statistical analysis. All analysis was performed using version 2.4.0 of the statistical programming environment R[6] on a machine with four Xeon 2.80 GHz CPUs and 3.5 GB of RAM, running Windows XP with Service Pack 2. We also used the R packages `xtable` (version 1.4-2), `cluster` (version 1.11.2), and `colorspace` (version 0.9), which are available from the Comprehensive R Archive Network (http://cran.r-project.org/). We used three R packages from BioConductor (http://www.bioconductor.org/); these were `Biobase` (version 1.12.2), `affyio` (version 1.2.0), and `affy` (version 1.12.0). Finally, we used version 1.3 of the packages `oompaBase`, `PreProcess`, `ClassDiscovery` and `ClassComparison` from the Object-Oriented Microarray and Proteomic Analysis project (available from our web site, http://bioinformatics.mdanderson.org/software.html).

The `binreg` software from the Duke web site was run using version 7.0.1 of MATLAB (The Mathworks Inc., Natick MA) on a 1GHz PowerPC G4 Mac PowerBook laptop running Mac OS X version 10.3.9. The Chang breast cancer data was processed using the default settings for the PM-only model in the DNA Chip Analyzer (`dchip2006.exe`; available from http://biosun1.harvard.edu/complab/dchip/).[5]

The complete Sweave source code for this analysis is available at the web site

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

All additional parameter settings for the software used in the analysis are specified in the source code.

ACKNOWLEDGEMENTS

We sincerely thank Dr. Anil Potti and Dr. Joseph Nevins of Duke University for their

patient cooperation while we asked them numerous questions about their analysis. They were consistently open and forthcoming, making every effort to supply us with the details we requested. Our failure to reproduce their results, in the face of their good faith effort to help us, points out the inherent difficulties in describing these kinds of analyses without supplying source code.

We thank Zoltan Szallasi, Jane Fridlyand, Lajos Pusztai, Gordon Mills, and David Stivers for helpful discussions during this work. We also thank Sarah Edmonson for her detailed comments on an early draft of the manuscript.

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

1. Potti, A., *et al.* Genomic signatures to guide the use of chemotherapeutics. *Nat Med* **12**, 1294–300 (2006).

2. Pounds, S., Morris, S.W. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* **19**, 1236–42 (2003).

3. Benjamini, Y., Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JRSS (B)* **57**, 289–300 (1995).

4. Chang, J.C., *et al.* Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet* **362**, 362–9 (2003).

5. Li, C., Wong, W.H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A* **98**, 31–6 (2001).

6. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2006).

7. Lamport, L. *LaTeX: A document preparation system* (Addison Wesley, Boston, 1994).

8. Laine, C., Goodman, S.N., Griswold, M.E., Sox, H.C. Reproducible research: moving toward research the public can really trust. *Ann Intern Med* **146**, 450–3 (2007).

9. Leisch, F., Rossini, A.J. Reproducible statistical research. *Chance* **16**, 46–50 (2003).

10. Buckheit, J., Donoho, D.L. Wavelab and reproducible research. In: A. Antoniadis, ed., *Wavelets and Statistics* (Springer-Verlag, Berlin,. New York, 1995).

11. Schwab, M., Karrenbach, M., Claerbout, J. Making scientific computations reproducible. *Computing in Science and Engineering* **2**, 61–67 (2000).

12. Peng, R.D., Dominici, F., Zeger, S.L. Reproducible epidemiologic research. *American Journal of Epidemiology* **163**, 783–789 (2006).

13. Gentleman, R. Reproducible research: a bioinformatics case study. *Stat Appl Genet Mol Biol* **4**, Article 2 (2005).

14. Mansmann, U., Ruschhaupt, M., Huber, W. Reproducible statistical analysis in microarray profiling studies. *Methods Inf Med* **45**, 139–45 (2006).

15. Holleman, A., *et al.* Gene-expression patterns in drug-resistant acute lymphoblastic leukemia cells and response to treatment. *N Engl J Med* **351**, 533–42 (2004).

16. Gyorffy, B., *et al.* Gene expression profiling of 30 cancer cell lines predicts resistance towards 11 anticancer drugs at clinically achieved concentrations. *Int J Cancer* **118**, 1699–712 (2006).

17. Gemma, A., *et al.* Anticancer drug clustering in lung cancer based on gene expression profiles and sensitivity database. *BMC Cancer* **6**, 174 (2006).

18. Rouzier, R., *et al.* Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clin Cancer Res* **11**, 5678–85 (2005).

19. Rouzier, R., *et al.* Microtubule-associated protein tau: a marker of paclitaxel sensitivity in breast cancer. *Proc Natl Acad Sci U S A* **102**, 8315–20 (2005).

Table 1: Size of the overlap in the top 50 genes using different sets of replicate microarrays.

	Average	A	B	C
Average	50	12	17	10
A	12	50	7	4
B	17	7	50	7
C	10	4	7	50

Table 2: Predictions of sensitivity or resistance on the test samples using the optimal set of principal components (Resp = responder, NR = non-responder).

	NR	Resp
Resistant	3	1
Sensitive	10	10

Table 3: Data sets and sources used in the paper by Potti et al.

Data Set	Platform	Web Site
NCI60 Drug Response		http://dtp.nci.nih.gov/docs/cancer/cancer_data.html (Oct 2006 release)
NCI60 expression, Novartis	U95Av2	http://dtp.nci.nih.gov/mtargets/download.html
24 breast tumors, docetaxel[4]	U95Av2	GSE349, GSE350, GDS360
17 lung, 13 ovarian cell lines[16]	U133A	http://www.mrw.interscience.wiley.com/jpages /0020-7136/suppmat/ijc.21570.html
29 lung cancer cell lines[17]	U133A	GSE4127
Adriamycin treated ALL [15]	U133A	GSE650, GSE651
51 breast tumor, TFAC[18,19]	U133A	http://data.cgt.duke.edu/Combo1.php
45 breast tumor, FAC	U95Av2	subset of GSE3143
171 breast tumor	U95Av2	GSE3143
91 lung tumor	U133Plus2.0	GSE3141
119 ovarian tumor	U133A	GSE3149
binreg (MATLAB) software		http://data.cgt.duke.edu/Combo1.php

Table 4: NSC numbers of drugs studied by Potti and colleagues.

NSC Number	Drug
628503	Docetaxel (Taxotere)
123127	Adriamycin (Doxorubicin)
26271	Cytoxan (Cyclophosphamide)
141540	Etoposide
125973	Paclitaxel (Taxol)
19893	5-Fluorouracil
609699	Topotecan