Figure 1 Plot of the first two principal components from the NCI-60 training set for docetaxel, into which the validation set from Chang *et al.*[2] has been projected. The first principal component completely separates sensitive from resistant cell lines. Test samples from breast cancer patients treated with docetaxel project into the center of the space, with responders (Resp) and nonresponders (NR) overlapping.

nonresponders) and point 5 (combining training and testing data when choosing gene weights) can produce 'better than chance' predictions in the wrong direction. This appears to have happened with "another, independent dataset of samples cultured from adriamycin-treated individuals (GEO accession numbers GSE650 and GSE651)"[1]. These GEO datasets, from Holleman *et al.*[3], include samples from pediatric patients with acute lymphocytic leukemia. There are 28 samples that are resistant to adriamycin (daunorubicin) and 94 that are sensitive. Figure 2c of Potti *et al.*[1] shows 99 resistant and 23 sensitive samples, suggesting that

most labels are reversed. If the labels are reversed, the model suggests administering the drug only to the patients it would not benefit.

7. When we apply the same methods but maintain the separation of training and test sets, predictions are poor (**Fig. 1** and **Supplementary Report 7**). Simulations show that the results are no better than those obtained with randomly selected cell lines (**Supplementary Report 8**).

We do not believe that any of the errors we found were intentional. We believe that the paper demonstrates a breakdown that results from the complexity of many bioinformatics analyses. This complexity requires extensive double-checking and documentation to ensure both data validity and analysis reproducibility. We believe that this situation may be improved by an approach that allows a complete, auditable trail of data handling and statistical analysis. We use Sweave[4,5], a package that allows analysts to combine source code (in R)[6] and documentation (in LaTeX)[7] in the same file. Our Sweave files are available at (http://bioinformatics.mdanderson.org/Supplements/ReproRsch-Chemo/). Running them reproduces our results and generates figures, tables and a complete PDF manuscript.

The idea of using the NCI-60 cell lines to predict patient response to chemotherapy is exciting. Our analysis, however, suggests that it did not work here.

Kevin R Coombes, Jing Wang & Keith A Baggerly

Department of Bioinformatics and Computational Biology, University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030, USA.
e-mail:kcoombes@mdanderson.org

1. Potti, A. *et al. Nat. Med.* **12**, 1294–1300 (2006).
2. Chang, J.C. *et al. Lancet* **362**, 362–369 (2003).
3. Holleman, A. *et al. N. Engl. J. Med.* **351**, 533–542 (2004).
4. Leisch, F. & Rossini, A.J. *Chance* **16**, 46–50 (2003).
5. Gentleman, R. *Stat. Appl. Genet. Mol. Biol.* **4**, Article 2 (2005).
6. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2006).
7. Lamport, L. *LaTeX: A Document Preparation System* (Addison Wesley, Boston, 1994).

Potti *et al.* reply:

We appreciate the interest that Coombes *et al.*[1] have shown in our reported results and agree that an algorithm that tracks the results of the various steps would be useful in such complex analyses. Unfortunately, they have not followed our methods in several crucial contexts and have made unjustified conclusions in others, and as a result their interpretation of our process is flawed.

Coombes *et al.*[1] raise three main issues. First, they cannot reproduce our methods for cell selection. This process involves using not only GI_{50} (the concentration needed to reduce the growth of treated cells to half that of untreated cells) concentrations but also LC_{50} (the concentration that kills 50% of treated cells) and TGI (the concentration required to completely halt the growth of treated cells) concentrations for each drug, as well as raw data available at the National Cancer Institute website in cases in which the –log concentrations are truncated. We have provided details describing these steps on our web page (http://data.cgt.duke.edu/NatureMedicine.php). Because Coombes *et al.*[1] did not follow these methods precisely and excluded cell lines and experiments with truncated –log concentrations, they have made assumptions inconsistent with our procedures.

Second, they point to inaccuracies in the gene lists we reported. As they note, software problems resulted in an off-by-one error in the matching of probe IDs with gene names. Additional inaccuracies resulted from

errors made when we assembled the gene lists. We have corrected these errors, and accurate gene lists were posted on the *Nature Medicine* website on 10 October. We regret any inconvenience this may have caused for other investigators but emphasize that these errors in no way influence the primary results of our study, as the models are defined by the training set, not by gene lists.

Third, they suggest that our method of including both training and test data in the generation of metagenes (principal components) is flawed. We feel this approach is entirely appropriate, as it does not include any information regarding the actual patient response and thus does not influence the generation of the signature with respect to predicting patient outcome. The aim of generating metagenes from test and validation data is to accommodate differences among the characteristics of the data from cancer cell lines and human tumors, and is similar to the use of methods of 'standardization' that are intended to correct for intrinsic differences in data, including batch effects, before analysis. Indeed, we find that the predictions are equally robust if the data are first standardized and the predictions are then carried out on independent validation cohorts with metagenes generated from only the training data (A.P. and J.N., unpublished data). Additionally, there was no accidental inclusion of genes from the validation data distinguishing responders from non-responders and this is not an explanation

for the generation of 'better than chance' predictions (including those within the acute lymphocytic leukemia dataset in which the labels are accurate—full details are provided on our web page), as the models and predictions depend solely on the training samples, not on gene lists. Moreover, when Coombes et al.[1] compared the results of models that create metagenes from training data alone to the more extensive model that creates metagenes with both training and test data, they obtained a very similar result to ours (Fig. 8 in **Supplementary Report 9**). In short, they reproduce our result when they use our methods. Coombes et al.[1] may disagree with us about the logic of creating metagenes, but clearly the models are not influenced by inaccurate gene lists.

Finally, we also note that we have applied our methods, as well as several of the original signatures, to predict patient response in additional datasets, some blinded to us, yielding accuracies consistent with our initial results[2,3]. We do see reproducible prediction of patient response with the previously reported methods and continue to believe that these methods are appropriate and robust.

Anil Potti & Joseph R Nevins

Duke Institute for Genome Sciences and Policy, Durham, North Carolina 27708, USA.
e-mail: j.nevins@duke.edu

1. Coombes, C.R., Wang, J. & Baggerly, K.A. *Nat. Med.* **13**, 1276–1277 (2007).
2. Hsu, D.S. *et al. J. Clin. Oncol.* **25**, 4350–4357 (2007).
3. Bonnefoi, H. *et al. Lancet Oncol.* (in the press).

Reply to 'Arsenic patent keeps drug for rare cancer out of reach for many'

To the editor:
In his news article "Arsenic patent keeps drug for rare cancer out of reach of many"[1], your reporter got his facts right, but I suggest that his editor muffed the title. Here's an accurate headline: "Patent got drug into the hands of many quickly."

The 1990s saw extraordinary changes for individuals with acute promyelocytic leukemia (APL)—changes that transformed a disease from 80% lethal to 80% curable in less than a decade. The central clinical discoveries, including Wang Zheng-yi's work with all-*trans* retinoic acid that built on earlier observations from Laurent Degos in Paris, came from China. Arsenicals have long medicinal histories in both East and West, but observations of the beneficial activity of arsenic trioxide in APL began in Harbin, China and then rapidly spread via an 'APL club' of scientific and clinical collaborations in Shanghai, Paris, Lyon, New York and Tokyo.

Nonetheless, the failure to disclose the medicinal formulation of this drug—undoubtedly because a Chinese patent had not yet been filed—effectively prevented others from replicating the work. Setting aside the question of whether this failure was helpful for patients or congruent with science, it is ironic that this effective nondisclosure actually enabled the US patent.

Your readers may guess how many companies in 1997 were interested in licensing intravenous arsenic for a disease that affected fewer than 1,000 patients annually in the US, especially when the manufacturing process described in the article by Hugues de Thé was given as "boiling"[1]. Nonetheless, that aroused interest, albeit minimal interest, in precisely specifying a pharmaceutical-grade formulation, developing manufacturing processes to make a lethally toxic compound in commercial quantities, running clinical trials to develop and replicate dosing schedules that are employed today, collecting and analyzing the clinical data, and collating nonclinical and clinical information into a drug application that could be reviewed by global regulatory agencies.

The US Food and Drug Administration is an unsung hero in this saga, both for funding our original trial through its Orphan Drug Grants program and for dropping any requirement for animal testing. Enormous credit belongs to the medical reviewer, Dr. Steven Hirschfeld, as the usual requirement for animal toxicology would have brought the project to a screeching halt. With extensive collaboration between academic, industrial and regulatory groups, a carefully manufactured and specified product went into the veins of patients without the necessity for killing a single animal.

The patenting process—and the system—worked exceptionally well: a paradigm for how the above-mentioned groups can work together. This process rapidly put the most effective drug for APL within reach of most individuals with the disease. And, importantly, this drug is safe—because although the solution used in Harbin provided for some seminal observations, it was simply not pharmaceutical grade, and when another 'boiled' drug was administered to patients, several people in the US died.

High prices for drugs—not to mention all other products—reduce their availability in developing countries. But, inevitably, these prices come down and patents expire. I have no idea what the 'right' price for arsenic should be, but I know that both companies mentioned in your article have had catastrophic failures with other drugs. The pharmaceutical industry survives only if its many failures can be amortized over a few successes.

A staggering proportion of mostly young adults with APL who are alive today would be dead if they had been diagnosed in 1990. They have the rest of their lives to complain about the price. Those of us in the APL club witnessed and enabled a remarkable period in human medicine. We are fortunate to enjoy the gratitude and camaraderie of the many patients who benefited from that success.

Raymond P Warrell, Jr

Genta Incorporated, 200 Connell Drive, Berkeley Heights, New Jersey 07092, USA.
e-mail: warrell@genta.com

1. Cyranoski, D. *Nat. Med.* **13**, 1005 (2007).