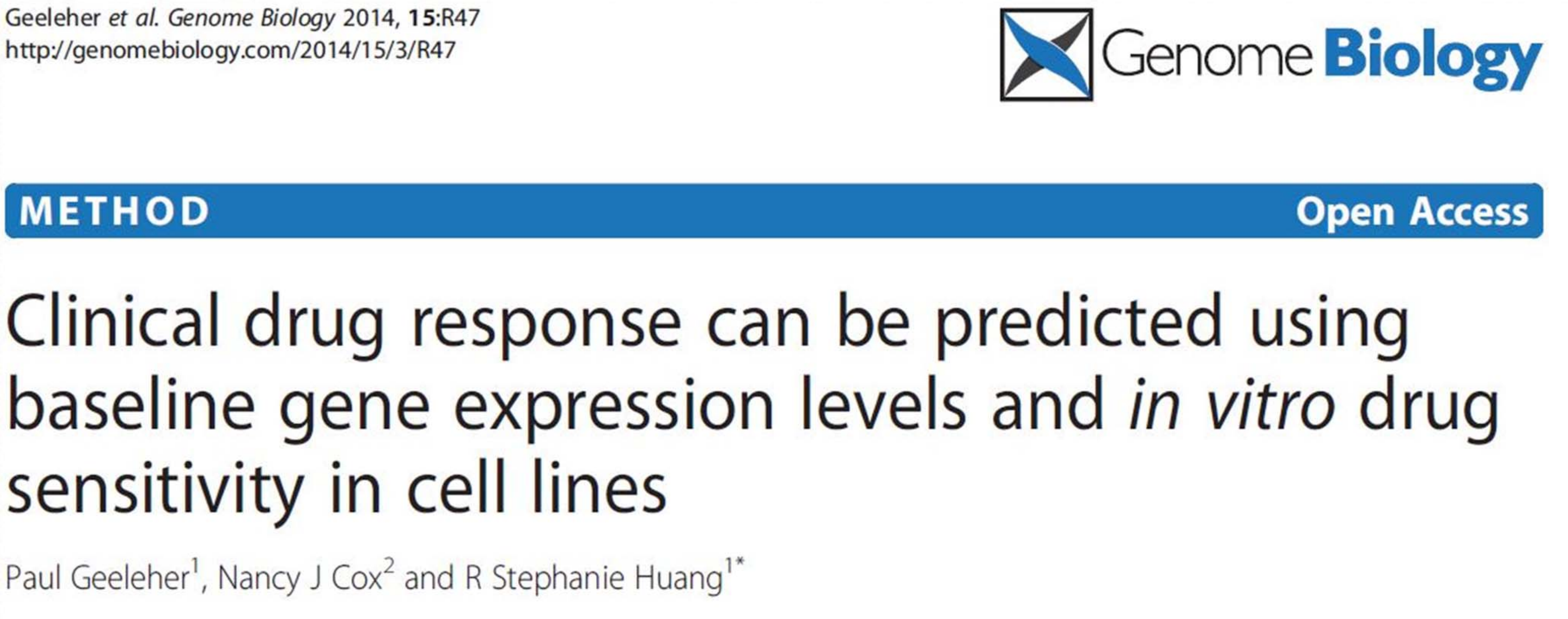
Predict drug response of cancer patients

Grace S. Shieh
Institute of Statistical Science
Academia Sinica
gshieh@stat.sinica.edu.tw

Ultimate goals

- To predict *in vivo* drug response, using models trained on cell line data.
- Geeleher et al. (2014), Genome Biology

A motivating example (2014)



Geeleher *et al. Genome Biology* 2014, **15**:R47
http://genomebiology.com/2014/15/3/R47

Genome **Biology**

METHOD **Open Access**

Clinical drug response can be predicted using baseline gene expression levels and *in vitro* drug sensitivity in cell lines

Paul Geeleher[1], Nancy J Cox[2] and R Stephanie Huang[1*]

- For model development, the approach was applied to recently released data from the Cancer Genome Project (CGP) , consisting of baseline (i.e. before drug treatment) gene expression microarray data and sensitivity to 138 drugs in a panel of almost 700 cell lines.

- Test sets:

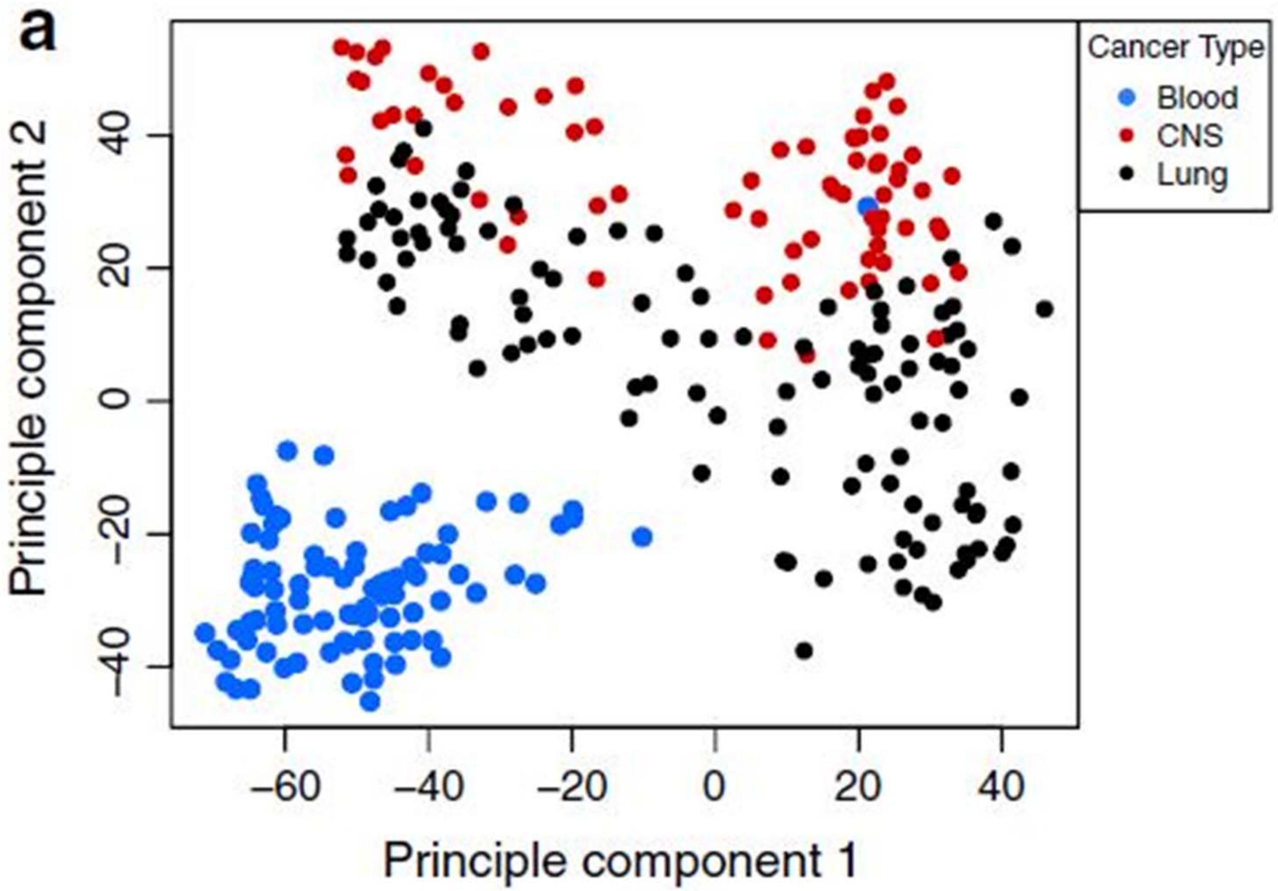
  An additional large panel of cell lines in the Cancer Cell Line Encyclopedia (CCLE), and a few sets of cancer patients' data from clinical trials.

- In preliminarily studies, we assessed several of the plethora of available machine learning algorithms, including random forests, PAM, principal component regression, Lasso and ElasticNet regression. Among them, ridge regression was consistently the best performer, with the added advantage of being highly computationally efficient.

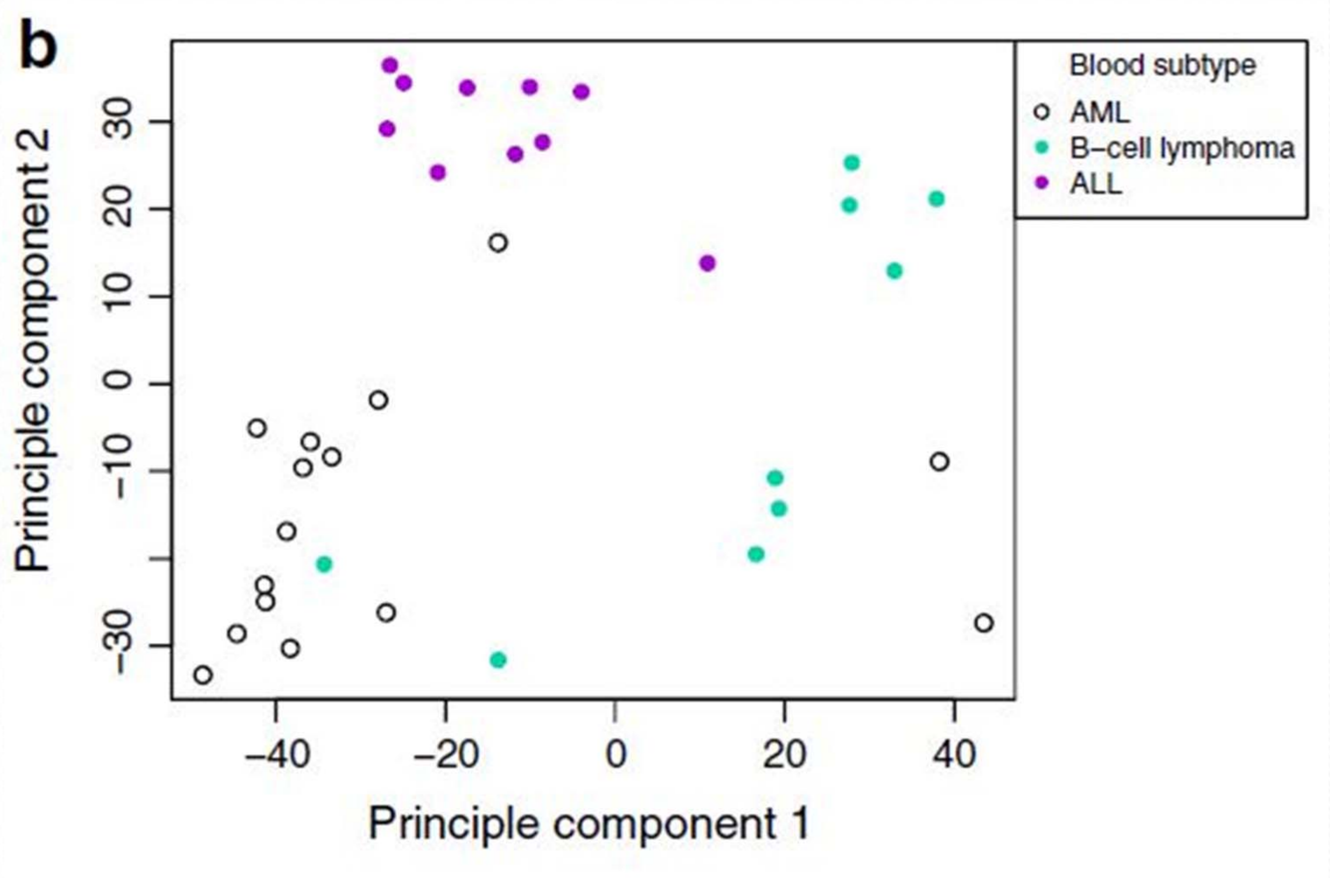
- Furthermore, principal component analysis (PCA) demonstrates that whole-genome gene expression can capture far more information about cancer biology, than may have been previously appreciated. As illustrated in Figure 2.

- This suggests that whole-genome gene expression acts as a surrogate for unmeasured genetic and non-genetic phenotypes.

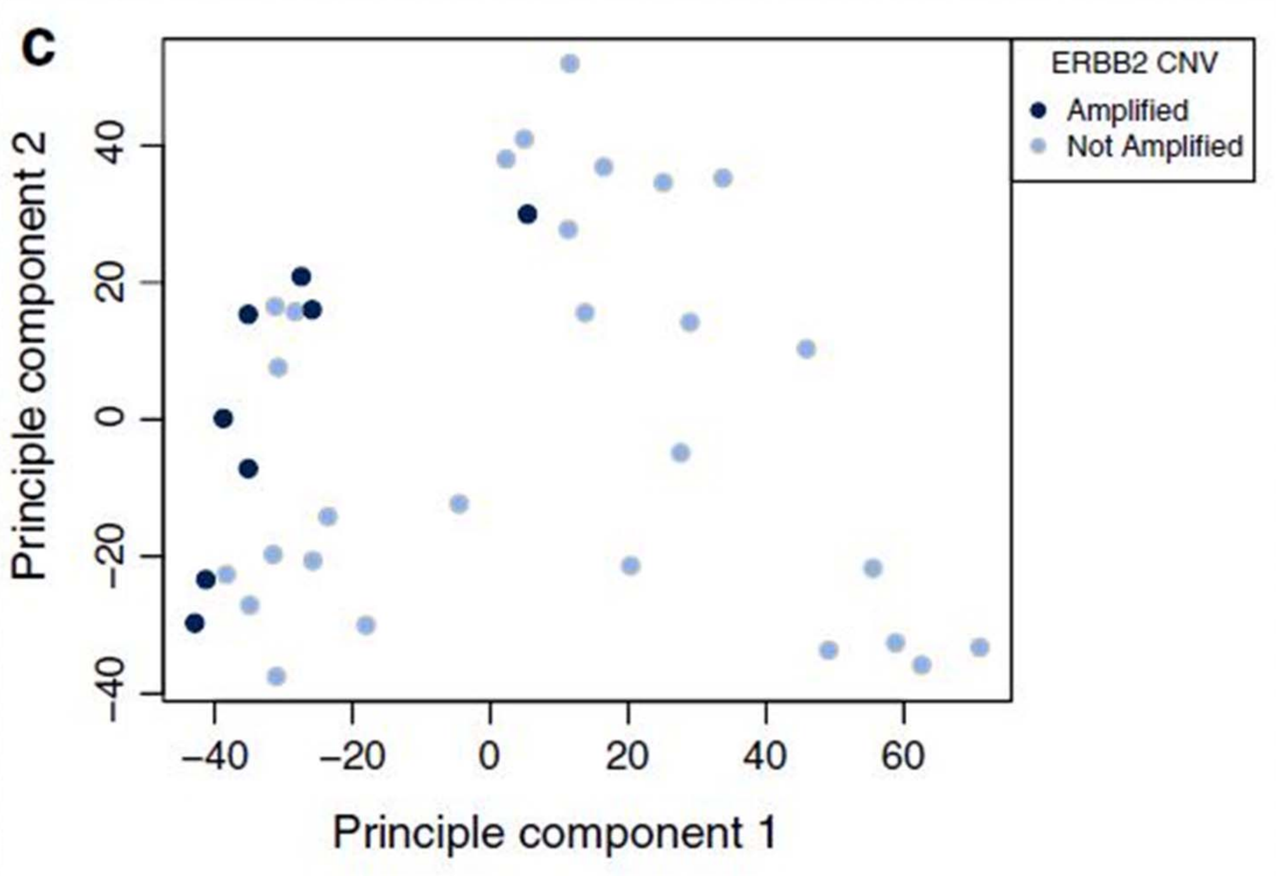
Clustering of cancer types on PC1 & PC2 of a gene expression matrix from the CGP cell lines.



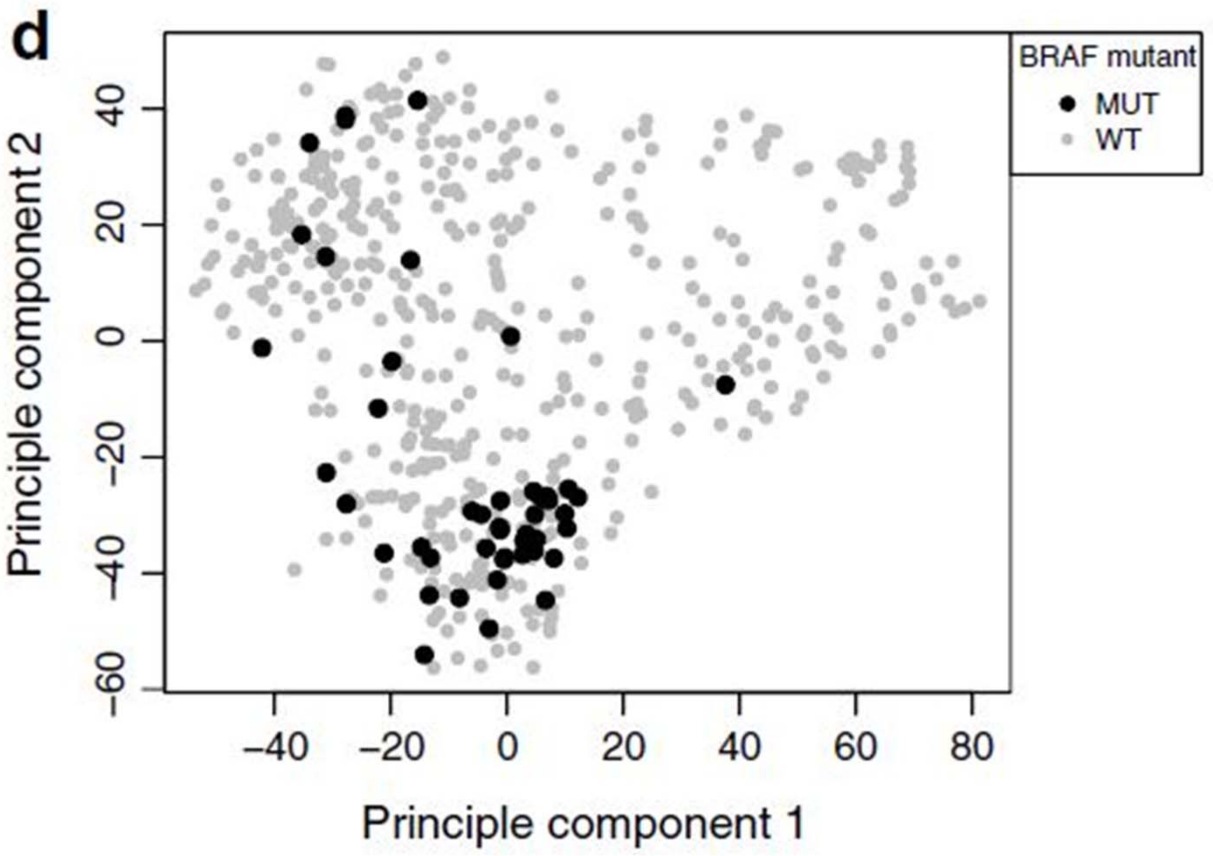
Clustering of subtypes of hematological cancers on PC1 & PC2 of a gene expression.



Clustering of ERBB2 amplified breast cancers on PC1 & PC2 of a gene expression matrix of CGP cell lines.



Clustering of BRAF mutant cancers on PC1 & PC2 of a gene expression matrix from all CGP cell lines.



Docetaxel (紫杉醇) and cisplatin treatment of breast cancer

- We first applied our method to gene expression microarray data obtained from 24 breast cancer tumor biopsies through a clinical trial.

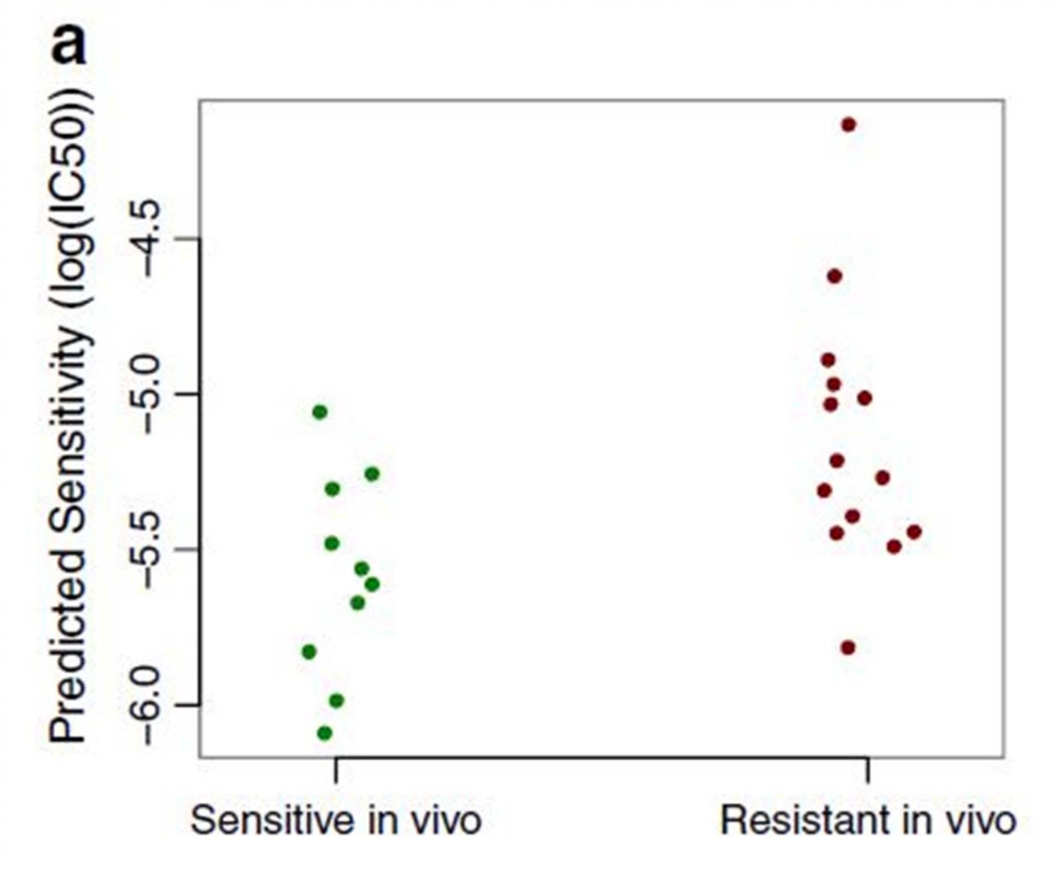
- The authors designated individuals as ‘sensitive’ or ‘resistant’ to docetaxel, depending on whether there was ≤25% or >25% of the tumor remaining.

- We used the CGP cell lines to build a ridge regression model, which related whole-genome gene expression to docetaxel sensitivity. We applied the model to *the in vivo* pretreatment breast cancer tumor expression data.

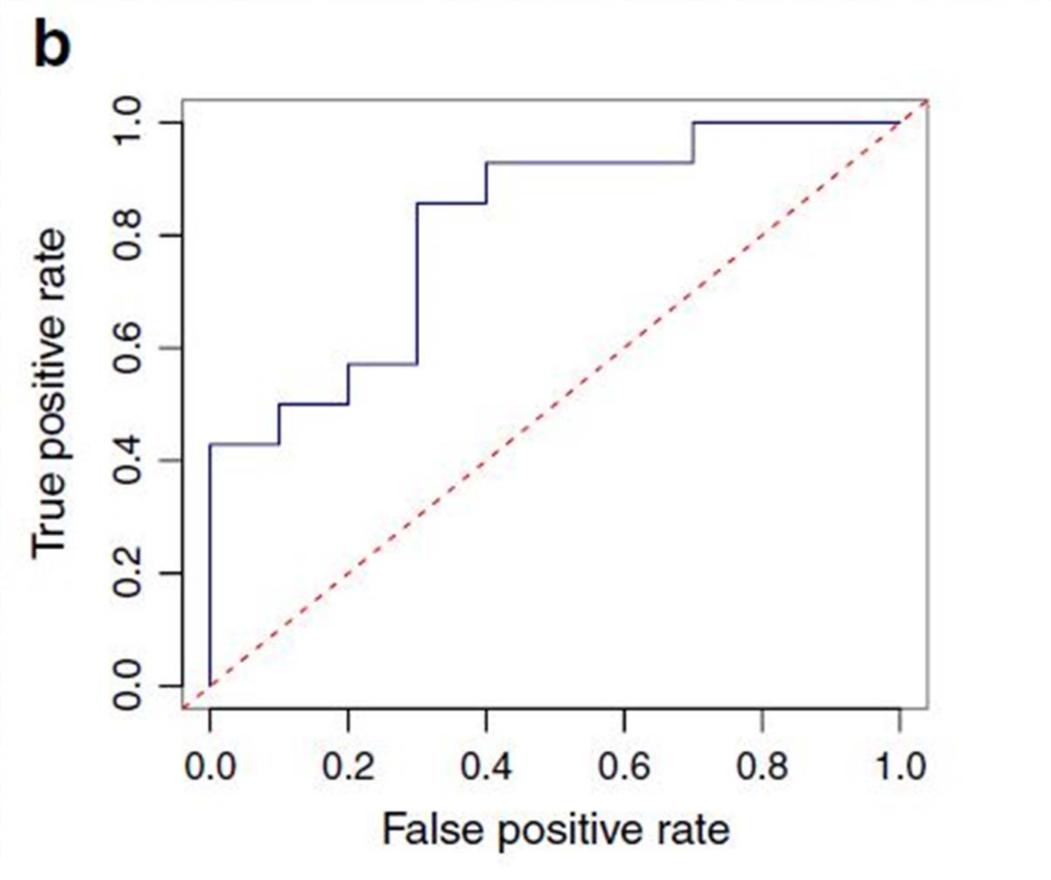
- Of the seven individuals who were predicted to be most sensitive, six are in the trial-defined sensitive group. ROC curve analysis revealed an AUC of 0.81

(Figure 3b; P = 5.0 × 10−3).

Prediction of docetaxel sensitivity in breast cancer patients.



Prediction of docetaxel sensitivity in breast cancer patients.



- For comparison, ElasticNet and Lasso regression models were also applied to this data, but both underperformed when compared to ridge regression (*P* = 0.01 from t-tests for both models)

- Next, we applied our method to a second breast cancer dataset, which assessed the response of 24 triple negative patients to neoadjuvant cisplatin therapy. We downloaded the raw data from ArrayExpress (accession number E-GEOD-18864).

This time, our models did not capture variability in clinical response.

- LOOCV indicated that, for the cell line panel, our models captured approximately the same proportion of variability in cellular response to cisplatin as they had for docetaxel (r = 0.35, P = 2.6 × 10^{−15} for docetaxel and r = 0.32, P = 1.4 × 10^{−13} for cisplatin from Pearson’s correlation test between LOOCV estimated log IC50 and measured log IC50 values).

- Notably, the authors of the original trial could not generate a gene signature from their data.
- Furthermore, they found that no genes were significantly correlated with response, following correction for multiple testing.

- Because of the lack of variability in drug response among a small group of patients, as cisplatin is not routinely used to treat breast cancer.

- Encouragingly, patients showing a ‘complete response’ or ‘progressive disease’ had the lowest and highest median predicted drug sensitivity values, respectively.

- We showed that the impressive performance may (at least in part) stem from the fact that whole-genome gene expression may act as a surrogate for unmeasured phenotypes that are directly relevant to chemotherapeutic sensitivity (Figure 2).

- Our method attained classification accuracy approaching, or even surpassing that of the gene signatures derived directly from clinical trials.

Bortezomib in myeloma

- Next, ridge regression was applied to a larger publicly available clinical phase II/III trial dataset, which assessed response to bortezomib in relapsed multiple myeloma patients.

- 168 patients had a clinically evaluable bortezomib response, which was classified as complete response (CR), partial response (PR), minimal response (MR), no change (NC) or progressive disease (PD).
- CR, PR and MR patients were defined as responders and NC and PD patients as non-responders.

- Expression in tumor cells was measured using either Affymetrix Human Genome U133A or U133B arrays in triplicate.
- This clinical dataset presents some obstacles. Firstly, only the preprocessed data is publicly available (GEO accession number [GEO:GSE9782]).

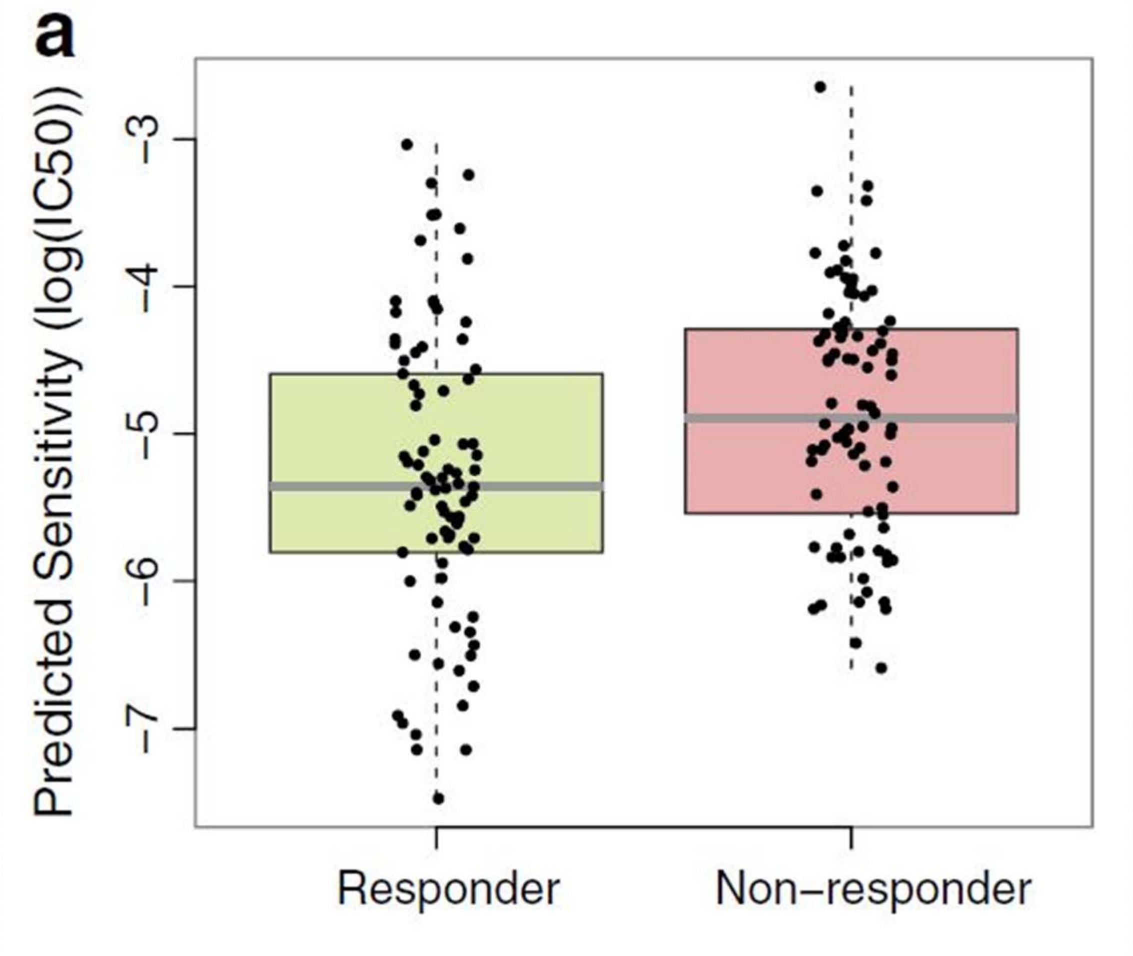
- the lack of standardized raw data processing likely lowers the performance of our model. Also, clinical samples were collected as part of three different clinical trials from various sites, and they were also hybridized in different batches.

- Furthermore, there is only a single myeloma cell line in the training panel.
- LOOCV in the cell line (training) set revealed similar correlations to other drugs ($r = 0.45$; $P = 2.6 \times 10-15$). Despite the suboptimal clinical data, our method captured substantial variability in bortezomib response.

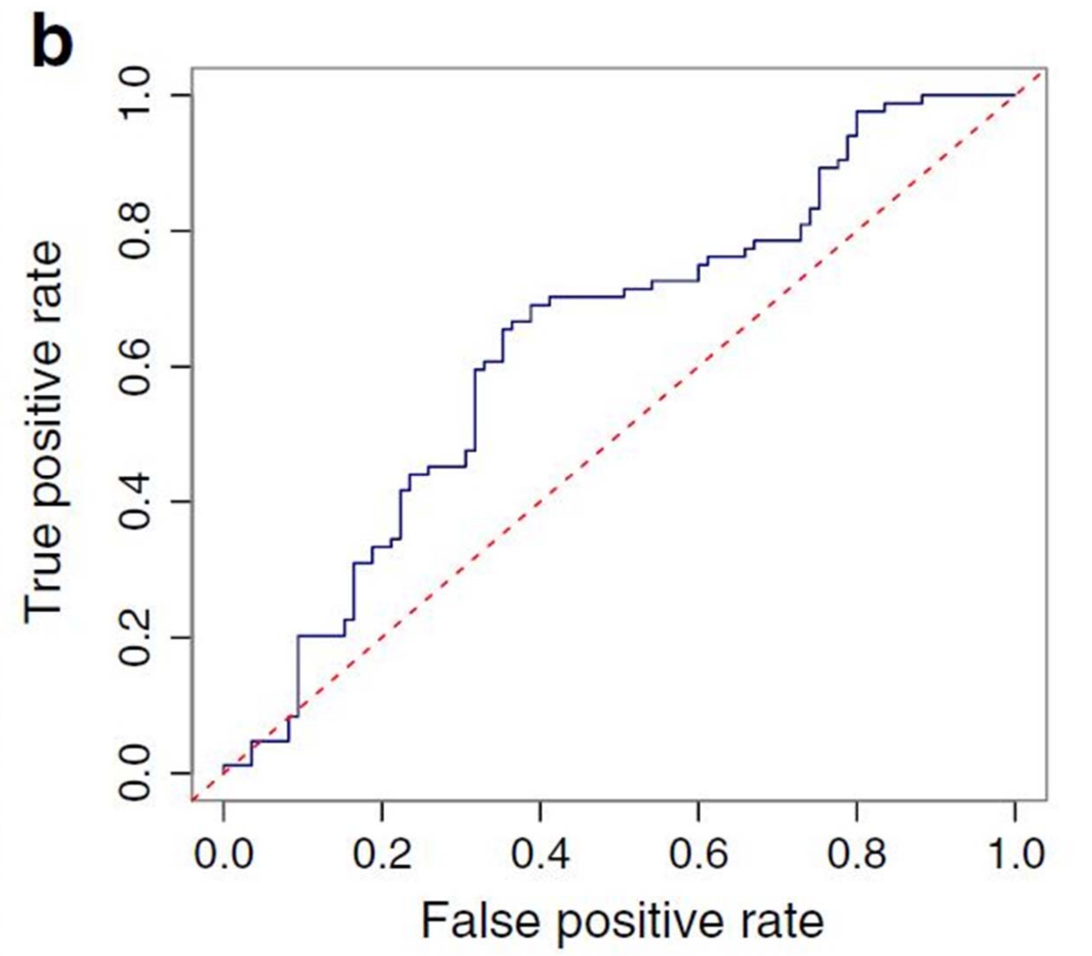
- There was a statistically significant difference between the predicted drug sensitivity in patients between the trial-defined responder and non-responder groups (Figure 4a; P = 8.9 × 10−4 for samples quantified using U133A.

- In the U133A dataset, the nine patients who were predicted to be most sensitive were all drug responders (Figure 4a). The AUCs from ROC curve analysis are 0.63 and 0.71 for U133A and U133B measurements(Figure 4b).

Prediction of bortezomib sensitivity in multiple myeloma patients.

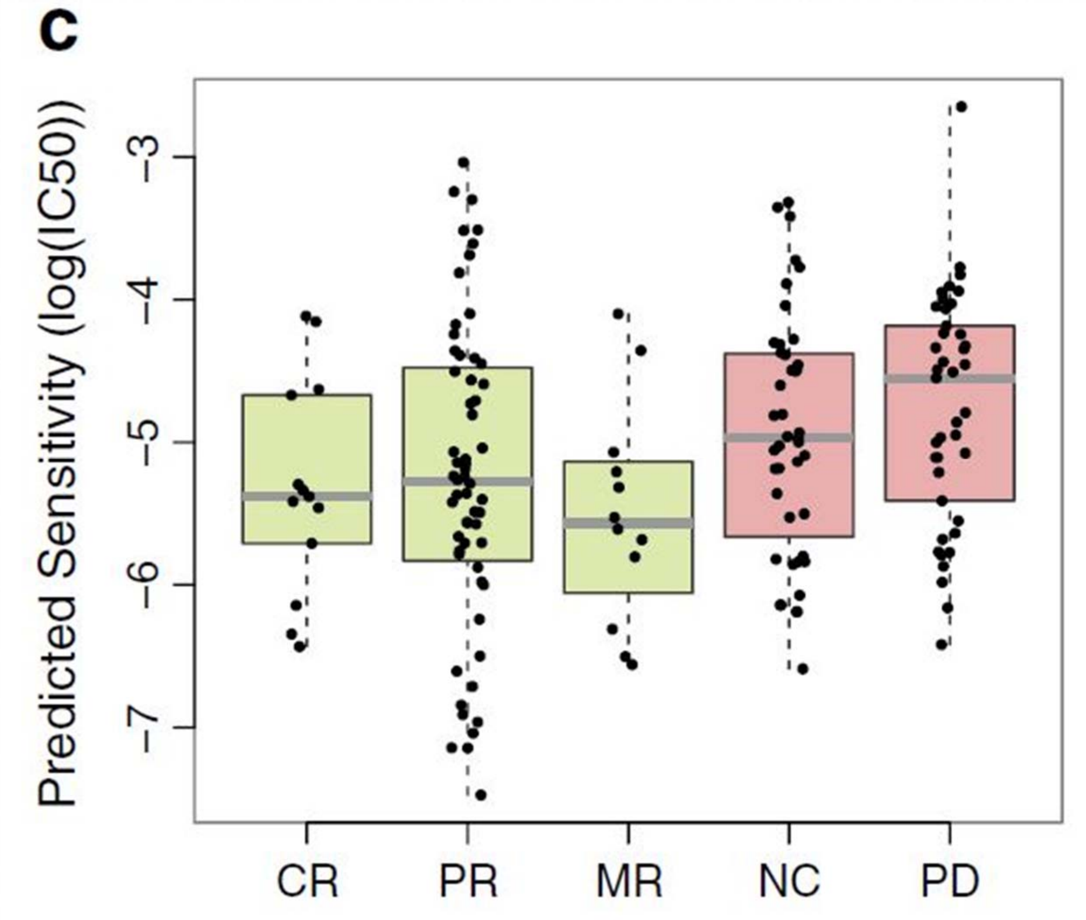


Prediction of bortezomib sensitivity in multiple myeloma patients.



- When the response was further subdivided (as CR, PR, MR, NC and PD), the median predicted drug sensitivity in each of these five groups was in exactly the correct order (Figure 4c and Additional file 1: Figure S6) in the U133B samples.

Prediction of bortezomib sensitivity in multiple myeloma patients.

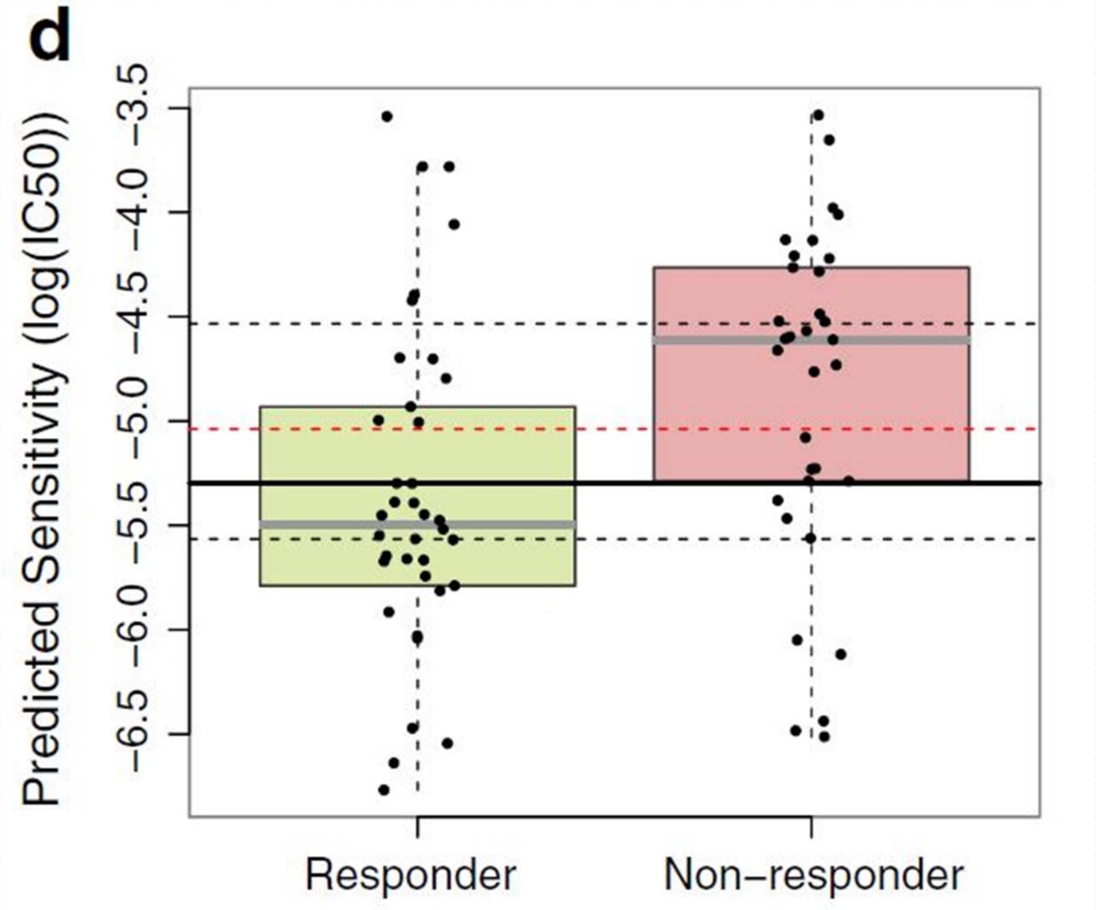


- The authors of the original clinical study reported that a 100-gene signature model, built on two arms of the trial (025 and 040), could predict bortezomib response in the third (039) arm of the trial with 63% accuracy.

- Our models generate a continuous variable and to compare the results previously reported directly, we must dichotomize this variable (i.e. split the data into ‘sensitive’ and ‘resistant’ at an arbitrary cut-point). At the optimal cut-point (−5.29), 51 of 71 patients were correctly classified, meaning that our method achieved a classification accuracy of 72%.

- For a large range (−5.57 to −4.53) of possible cut-points, our accuracy was greater than the 63% achieved by the trialderived gene signature (Figure 4d).
- The results indicate that our models offer a substantial performance improvement.

Prediction of bortezomib sensitivity in multiple myeloma patients.



Erlotinib in non-small cell lung cancer (NSCLC)

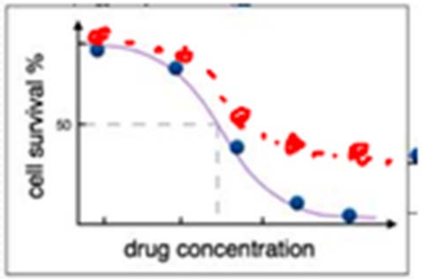
- Finally, we applied our approach to a dataset from a cancer study. A subset of patients (NSCLC) were treated with either erlotinib (n=25), an EGFR inhibitor, or sorafenib (n = 37), a VEGFR inhibitor, in a second-line setting.

- Inspection of the training data revealed that only a very small proportion of the cell lines assessed for sensitivity to these drugs were within the drug screening concentration used by the CGP. In contrast, most cell lines treated with cytotoxic agents, for example docetaxel, have more accurately quantified IC50 values.

- This can be rigorously demonstrated by segmenting all drugs into two groups (cytotoxic or targeted標靶藥) and comparing the median size of the confidence intervals associated with the IC50 values of each drug. Unsurprisingly, the confidence intervals are larger for targeted agents (average of 1.9 for cytotoxic compared to 4.5 for targeted agents; P = 1.4 × 10^{-5} from a Wilcoxon rank sum test).

Ridge vs logistic ridge regression

- In light of this, it is not reasonable to fit a linear ridge regression for most targeted agents, because IC50 values for most cell lines were derived using extrapolated data, and thus have very large associated confidence intervals.



- Consequently, we fitted logistic ridge regression models for the 15 most sensitive (which had reliably measured IC50 values) versus the 55 most resistant CGP cell lines.

LOOCV in trainging set

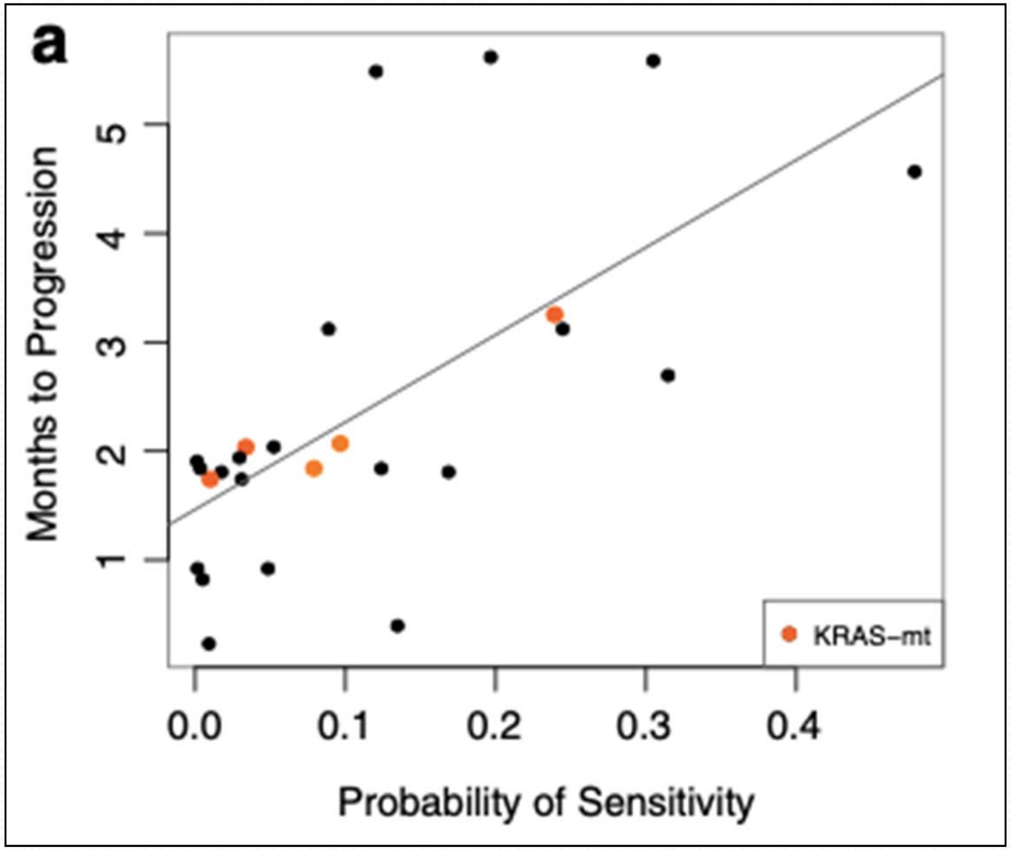
- In LOOCV, this method provided 89% classification accuracy on the training set and separated sensitive and resistant groups with $P = 9.3 \times 10^{-5}$, providing additional support for applying this approach to clinical samples.

Test set result

- When applied to the clinical trial data, this modified approach captured a large proportion of variability in the *in vivo* erlotinib response (Figure 5a; rho = 0.64 and P = 5.3×10^{-4} from a Spearman's correlation test).

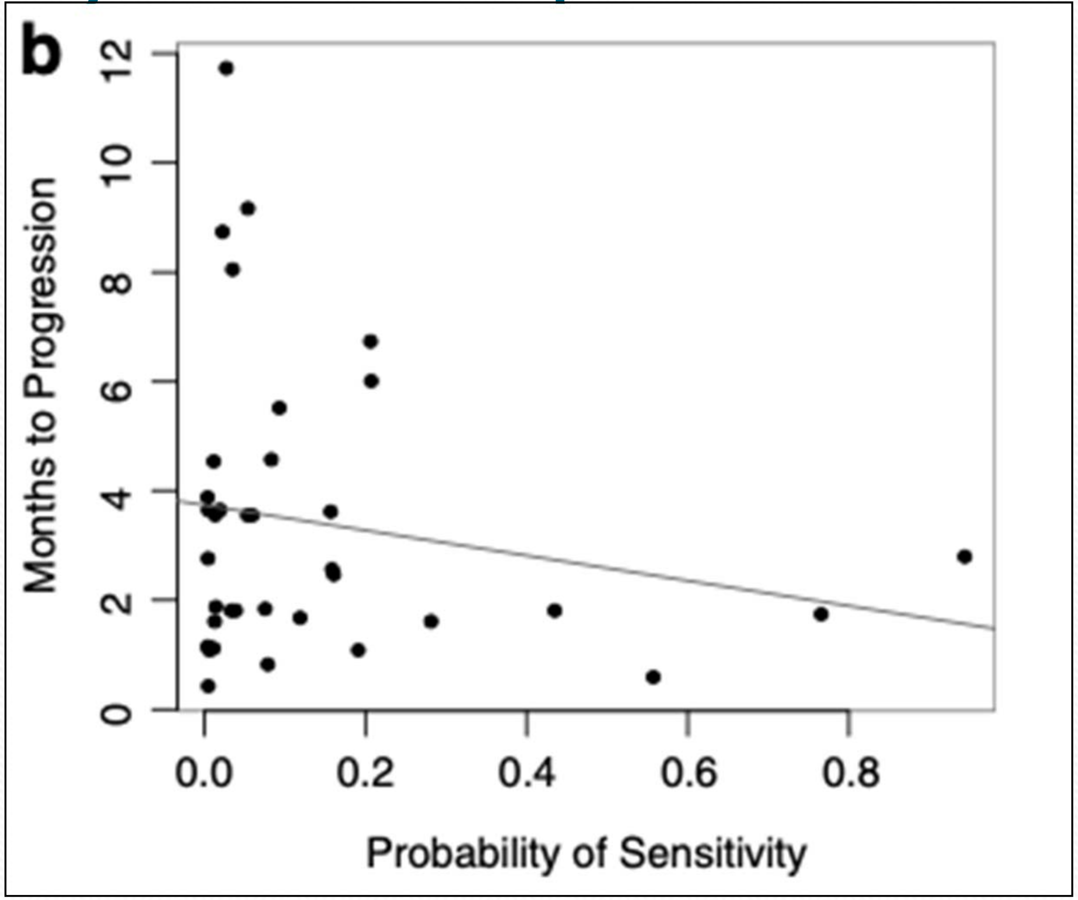
- We assessed the difference in predicted probability of erlotinib sensitivity (from the logistic ridge regression model) between individuals with disease progression (Resistant) and those without disease progression at two months (sensitive). In our case, the difference was highly statistically significant ($P = 4.9 \times 10^{-4}$ from a t-test).

Figure 5 Prediction of erlotinib sensitivity in NSCLC patients



(a) Months-to-progression plotted against predicted probability of erlotinib sensitivity. KRAS mutations are highlighted and a linear regression line is shown.

Figure 5 Prediction of sorafenib sensitivity in NSCLC patients



(b) Months-to-progression for individuals treated with sorafenib plotted against the predicted probability of sensitivity.

- This suggests that whole-genome gene expression models, derived from a large panel of cell lines (pancancer), have superior power to predict erlotinib sensitivity, compared to the 76-gene EMT signature (generated using both cell lines and patients).

- We modified the original algorithm (to use logistic ridge instead of linear ridge regression) in the analysis of erlotinib data. This is justifiable given the severe noise associated with the IC50 values for these types of targeted agents in the CGP cell lines.

- These results are congruent with the emerging view that –omics characterization of tumors may rival traditional tissue-of-origin and pathological descriptors for a variety of clinically important classifications.
- However, a different machine learning algorithm may be better suited to the distribution of RNA-seq data.

Discussion

- We found that, in lymphoblastoid (類淋巴母細胞) cell lines, far more of the variability in growth rate (between cell lines isolated from different individuals) can be explained by the transcriptome (GED) than by genome-wide SNPs (DNA).

- We have demonstrated that models derived from a very large panel of cell lines achieve equal or better performance for clinical drug sensitivity prediction than those derived directly from patients, and these findings can have a profound impact on patient care.

Methods

- These data were preprocessed using the robust multi-array average algorithm (implemented by the rma() function in the affy library in R).

- Both datasets were generated on different microarray platforms, thus we used a subset of genes represented on both platforms. This typically left approximately 10,000 gene symbols remaining.
- We removed the 20% of genes with lowest variability in expression across all samples. Critique: Using 8K genes, too many!

Predicting in vivo drug sensitivity using linear ridge regression for docetaxel, cisplatin and bortezomib clinical trials

- We used the linearRidge() function from the ridge package in R.

- The predict.linearRidge() function from the ridge package in R, thus yielding a drug sensitivity estimate for each patient.

Leave-one-out cross-validation

- These predicted IC50 values were then compared to the measured IC50 values, using a Pearson's correlation test, giving an estimate of prediction accuracy.

But for dichotomized response (S or R), we should compute the Cohen's kappa (in R), measuring corr between training & test set.

ElasticNet and Lasso models

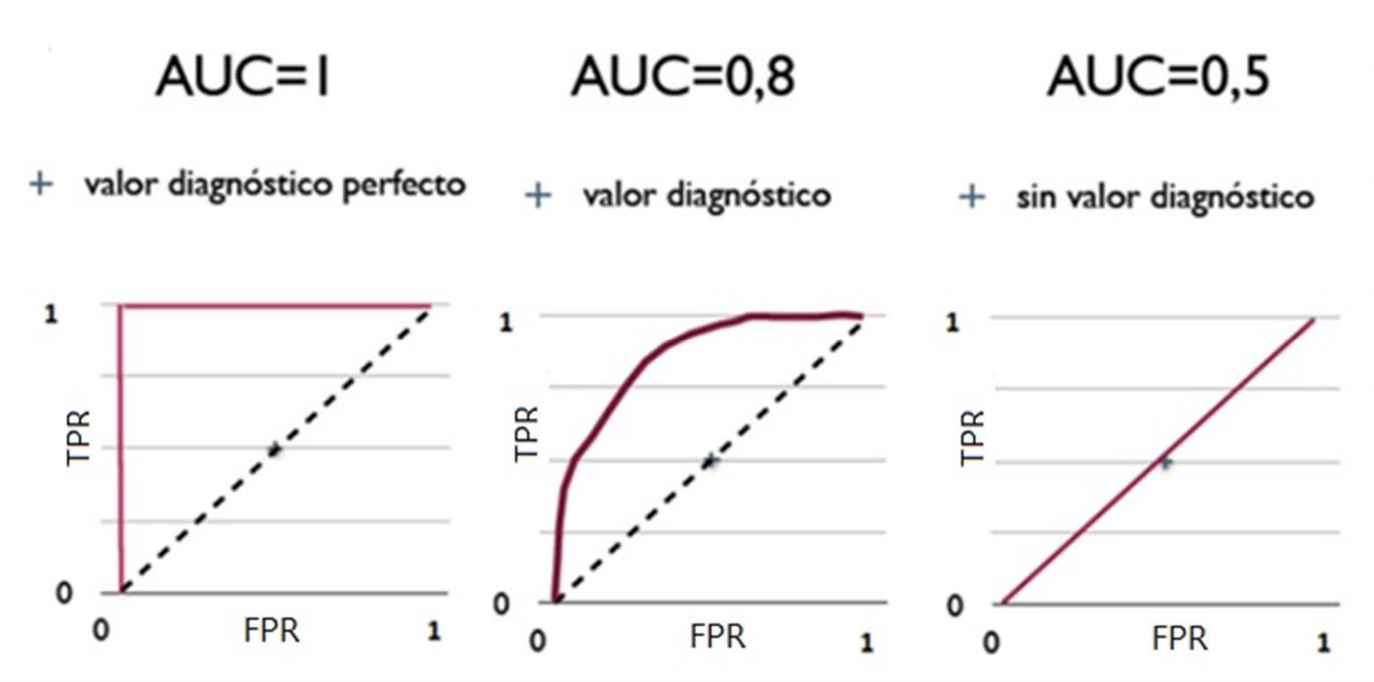
- ElasticNet and Lasso regression models were fitted using the glmnet package in R. The Lasso penalty parameter was selected using the automatic cross-validation feature (i.e. the cv.glmnet() function). ElasticNet penalty parameters (alpha and lambda) were selected using the caret package in R.

Statistical analysis of results

- For logistic ridge regression (a classifier), ROC curve analysis was performed using the ROCR package in R.

ROC curve

- Area under curve: (AUC):



- Thank you!

ROC curve

- Area under curve: (AUC):

ROC曲線下方的面積（Area under the Curve of ROC (AUC ROC)），其意義是：

1. 因為是在 1 x 1 的方格裡求面積，AUC必在0~1之間。
2. 假設閾值以上是陽性，以下是陰性；
3. 若隨機抽取一個陽性樣本和一個陰性樣本，分類器正確判斷陽性樣本的值高於陰性樣本之機率 = AUC。
4. 簡單說：AUC值越大的分類器，正確率越高。