PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

A regularized functional regression model enabling transcriptome-wide dosage-dependent association study of cancer drug response

Evanthia Koukouli[1]*, **Dennis Wang**[2,3], **Frank Dondelinger**[4☯], **Juhyun Park**[1☯]

1 Department of Mathematics and Statistics, Fylde College, Lancaster University, Bailrigg, Lancaster, UK, **2** Sheffield Institute for Translational Neuroscience, University of Sheffield, Sheffield, UK, **3** Department of Computer Science, University of Sheffield, Sheffield, UK, **4** Centre for Health Informatics and Statistics, Lancaster Medical School, Lancaster University, Bailrigg, Lancaster, UK

☯ These authors contributed equally to this work.
* e.koukouli@lancaster.ac.uk

Citation: Koukouli E, Wang D, Dondelinger F, Park J (2021) A regularized functional regression model enabling transcriptome-wide dosage-dependent association study of cancer drug response. PLoS Comput Biol 17(1): e1008066. https://doi.org/10.1371/journal.pcbi.1008066

Editor: Avner Schlessinger, Icahn School of Medicine at Mount Sinai, UNITED STATES

Received: June 12, 2020

Accepted: December 17, 2020

Published: January 25, 2021

Data Availability Statement: The data underlying the results presented in the study are available from the Genomics of Drug Sensitivity in Cancer website at https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Home.html.

Funding: EK is supported by the North West Social Science Doctoral Training Partnership (https://nwssdtp.ac.uk/) under grant no. 2035874. DW is supported by the National Institute of Health Research Sheffield Biomedical Research Centre (http://sheffieldbrc.nihr.ac.uk/), Rosetrees Trust

Abstract

Cancer treatments can be highly toxic and frequently only a subset of the patient population will benefit from a given treatment. Tumour genetic makeup plays an important role in cancer drug sensitivity. We suspect that gene expression markers could be used as a decision aid for treatment selection or dosage tuning. Using *in vitro* cancer cell line dose-response and gene expression data from the Genomics of Drug Sensitivity in Cancer (GDSC) project, we build a dose-varying regression model. Unlike existing approaches, this allows us to estimate dosage-dependent associations with gene expression. We include the transcriptomic profiles as dose-invariant covariates into the regression model and assume that their effect varies smoothly over the dosage levels. A two-stage variable selection algorithm (variable screening followed by penalized regression) is used to identify genetic factors that are associated with drug response over the varying dosages. We evaluate the effectiveness of our method using simulation studies focusing on the choice of tuning parameters and cross-validation for predictive accuracy assessment. We further apply the model to data from five *BRAF* targeted compounds applied to different cancer cell lines under different dosage levels. We highlight the dosage-dependent dynamics of the associations between the selected genes and drug response, and we perform pathway enrichment analysis to show that the selected genes play an important role in pathways related to tumorigenesis and DNA damage response.

Author summary

Tumour cell lines allow scientists to test anticancer drugs in a laboratory environment. Cells are exposed to the drug in increasing concentrations, and the drug response, or amount of surviving cells, is measured. Generally, drug response is summarized via a single number such as the concentration at which 50% of the cells have died (IC50). To avoid

(ref: A2501), and the Academy of Medical Sciences Springboard (https://acmedsci.ac.uk/grants-and-schemes/grant-schemes/springboard) under grant no. SBF004/1052. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

relying on such summary measures, we adopted a functional regression approach that takes the dose-response curves as inputs, and uses them to find biomarkers of drug response. One major advantage of our approach is that it describes how the effect of a biomarker on the drug response changes with the drug dosage. This is useful for determining optimal treatment dosages and predicting drug response curves for unseen drug-cell line combinations. Our method scales to large numbers of biomarkers by using regularization and, in contrast with existing literature, selects the most informative genes by accounting for drug response at untested dosages. We demonstrate its value using data from the Genomics of Drug Sensitivity in Cancer project to identify genes whose expression is associated with drug response. We show that the selected genes recapitulate prior biological knowledge, and belong to known cancer pathways.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Cancer is a heterogeneous disease, with individual tumours showing sometimes very different mutational and molecular profiles. The genetic makeup of a tumour influences how it reacts to a given anti-cancer drug. However, due to lack of predictive markers of tumour response, often patients with very different tumour genetic makeup will receive the same therapy, resulting in high rates of treatment failure [1]. Large clinical trials in rapidly lethal diseases are expensive, complex and often lead to failure due to lack of efficacy at a given dosage [2]. One major issue for some cancer treatments, e.g. chemotherapies, are cytotoxic effects that result in collateral damage of the healthy host tissue [3]. Patient remission depends not only on the selection of the right drug but also on the determination of the optimal dosage, especially when drugs with small therapeutic range, high toxicity levels or both are administered. Genetic factors can help fine-tune the dosage for individual patients, so that the minimal effective dosage can be delivered [4].

Treatment response in patients with specific cancers had been intensely examined in relation to the molecular characteristics of the tumours [5]. However, cellular heterogeneity within the tumour and the lack of standard metrics for quantifying drug response in patients can make it difficult to computationally model response as a function of molecular features. Cancer cell line drug screens can provide valuable information about the effect of genetic features on drug dose-response in a controlled setting. During the last decade, there have been several systematic studies that examined the relationship between genetic variants and drug response in cell lines [6–10]. There have also been studies that measured transcriptional profiles [11, 12] and drug response in cancer cells after administering anticancer drugs at various dosages [13, 14]. By comparing multiple genomic features of cell lines to drug response, the investigators were able to identify gene signatures for drug responsiveness in specific cancer types. However, these signatures were selected based on a single summary statistic of response, usually IC50, that may not always be the most useful metric for differentiating drugs [15], and only provides information on one dose concentration. While these existing signatures of drug response provide a way towards selecting the right drug for a patient, none of them characterize gene-dose relationships that may ultimately identify the optimal dose for a drug to use in the clinic.

With regards to the high-dimensional nature of genomic data sets, it is worth noting that highly-complex data sets with non-stationary trends are not easily amenable to analysis by classic parametric or semi-parametric mixed models. Such effects, e.g. the effect of genes on

drug response over different drug dosages (dose-varying effect), can be examined using varying coefficient models which allow for the covariate effect to be varying instead of constant [16]. Methods to estimate varying covariate effects include global and local smoothing, e.g. kernel estimators [17, 18], basis approximation [19] or penalized splines [20]. Although non-parametric techniques can reduce modeling biases [21], they often suffer from the "curse of dimensionality" [22]. Inference in these models becomes impossible as the number of predictors increases, and often selecting a smaller number of important variables for inclusion into the model is clinically beneficial. Sparse regression has enabled a more flexible and computationally "inexpensive" way of choosing the best subset of predictors [23]. However, these methods cannot handle ultra-high dimensional problems without losing statistical accuracy and algorithmic stability, since they handle all of the predictors jointly. Consequently, there is a need of prior univariate tests focused on filtering out the unimportant predictors by estimating the association of each predictor to the outcome variable separately [21, 24, 25]. The advantage of using varying coefficient models along with a variable screening algorithm on genomic data sets was first introduced to explore the effect of genetic mutations on lung function [24]. Recently, Wang et al. [26] and Tansey et al. [27] independently proposed methods for modeling drug-response curves via Gaussian processes and linking them to biomarkers. In both cases, the authors did not use their models for dosage-dependent inference of biomarker effects. Additionally, the highly non-linear neural network model in Tansey et al. [27] makes interpretation of biomarker effects challenging.

Here, we extended the methodology of Chu et al. [24] to the objective of assessing the transcriptomic effect on anti-cancer drug response, where our coefficient functions were allowed to vary with dosage. We developed a functional regression framework to study the effectiveness of multiple anticancer agents applied in different cancer cell lines under different dosage levels, adjusting for the transcriptomic profiles of the cell lines under treatment. We considered a dose-varying coefficient model, along with a two-stage variable selection method in order to detect and evaluate drug-gene relationships, and then applied this method to data extracted from the Genomics for Drug Sensitivity in Cancer (GDSC) project [7]. To compare and differentiate similar treatments, we examined a case study of five BRAF targeted compounds under different dosages to almost 1000 cancer cell lines. We used baseline gene expression measurements for the cancer cell lines to investigate gene-drug response relationships for almost 18000 genes. Gene rankings were obtained based on the estimated effects of the genes on the drug response. The resulting model describes the whole dose-response curve, rather than a summary statistic of drug response (e.g. IC50), which allowed us to identify trends in the gene-drug association at untested dose concentrations.

Materials and methods

The Genomics of drug sensitivity in cancer data

Drug sensitivity data and molecular measures derived from 951 cancer cell lines used for the screening of 138 anticancer compounds were downloaded from the GDSC database (https://www.cancerrxgene.org/). We specifically focused on cell lines of cancers of epithelial, mesenchymal and haematopoietic origin treated by five BRAF targeted inhibitors (PLX-4720, Dabrafenib, HG6-64-1, SB590885 and AZ628; GDSC1 data). The maximum screening concentration for each different drug was: 10.00 uM for PLX-4720 and Dabrafenib, 5.12 uM for HG6-64-1, 5.00 uM for SB590885 and 4.00 uM for AZ628. Additionally, we used the independently generated GDSC2 data set to validate our approach on drugs targeting *MEK1*, *MEK2* genes (Trametinib−1.00 uM; Selumetinib−10.00 uM, and; PD0325901−0.250 uM) and the PI3K/MTOR signalling pathway (Alpelisib−10.00 uM; AMG-319−10.00 uM, and; AZD8186

−10.00 uM). The drug sensitivity measurement was obtained via fluorescence-based cell viability assays 72 hours after drug administration [7]. Approximately 66% of drug sensitivity responses were measured over nine dose concentrations (2-fold dilutions) and 34% were measured over five drug concentrations (4-fold dilutions). In total, we considered 3805 cancer cell line-drug combinations (experimental units). The distribution of different tissues of origin treated were similar across the different drugs tested (for additional information see S1 Fig). Paired microarray gene expression data (17737 genes) were available together with the dose-response data.

The dose-response data also included a blank response for cells on the experimental plate that had not been seeded with cells or treated with a drug. Blank responses have been used to adjust for the magnitude of the observation error while measuring the amount of cells in each plate. We used an affine transformation to the reported responses in order to normalise them within the drug concentration interval, 0 (0% of the maximum dosage) to 1 (100% of the maximum dosage). In particular, for the normalising procedure, we have used the formula:

$$NR_{ij} = \frac{R_{ij} - BR_i}{CR_i - BR_i} \tag{1}$$

where R_{ij} is the response of the ith experimental unit at the jth dosage level, CR_i is the response under no drug administration (zero dose, $n_i = 1$), BR_i is the blank response of the ith experimental unit as described above and NR_{ij} is the new score taken from the transformation, $i = 1, \ldots, 3805$, $j = 1, \ldots, n_i$.

A two-stage algorithm for identification of gene-drug associations

Non-parametric techniques are a great tool for reducing modeling bias and producing data driven inference. However, flexible modeling techniques applied on high-dimensional genomic data sets can often cause real problems in statistical inference. Sparse regression techniques, such as the LASSO, can be used as dimensionality reduction techniques, but cannot handle ultra-high dimensional problems without introducing statistical inaccuracies, algorithmic instability and a huge computational burden [23]. Hence, the need for a feature screening algorithm which will marginally filter unimportant variables becomes essential. Below we further explain the two-stage algorithm that has been built in order to detect and explore dose-dependent gene-response associations.

Let the repeated measures data $\{(d_{ij}, y_{ij}, \boldsymbol{z}_i, \boldsymbol{x}_i): j = 1, \ldots, n_i, i = 1, \ldots, n\}$, where y_{ij} is the response of the ith experimental unit (corresponds to a drug sensitivity assay of a specific drug on a specific cell line) at the jth drug dosage level d_{ij} and $\boldsymbol{z}_i$ along with $\boldsymbol{x}_i$ are the corresponding vectors of scalar (dose-invariant) covariates. The covariate vector $\boldsymbol{z}_i = (1, z_{i1}, \ldots, z_{ip})^T$ is a low-dimensional vector of predictors that should be included in the model, whereas $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iG})^T$ is a high-dimensional vector, i.e. 17737 gene expression measurements, that needs to be screened. We assumed that only a small number of x-variables (in our case, genes) are truly associated with the response while most of them are expected to be irrelevant (sparsity assumption).

To explore potential dose-varying effects between the covariates and the drug response, we consider the following varying coefficient model:

$$y_{ij} = \sum_{k=0}^{p} z_{ik}\beta_k(d_{ij}) + \sum_{g=1}^{G} x_{ig}\gamma_g(d_{ij}) + \varepsilon_{ij} \tag{2}$$

where $\{\beta_k(\cdot), k = 0, \ldots, p\}$ and $\{\gamma_g(\cdot), g = 1, \ldots, G\}$ are smooth functions of dosage level $d \in \mathscr{D}$, where $\mathscr{D}$ is a closed and bounded interval of $\mathbb{R}$. The errors ε_{ij} were assumed to be independent

across subjects and potentially dependent within the same subject with conditional mean equal to zero and variance $\mathrm{Var}(\varepsilon) = \sigma^2(d) = V(d)$.

Methods for estimating the coefficient functions in Eq (2) include local and global smoothing methods, such as kernel smoothing, local polynomial smoothing, basis approximation smoothing etc. For computational convenience, in this application we used basis approximation smoothing via B-splines.

Let the sets of basis functions $\{B_{lk}(\cdot): l = 1, \ldots, L_k\}$ and $\{B'_{lg}(\cdot) : l = 1, \ldots, L_g\}$ and constants $\{\zeta_{lk}: l = 1, \ldots, L_k\}$ and $\{\eta_{lg}: l = 1, \ldots, L_g\}$ where $k = 0, \ldots, p$ and $g = 1, \ldots, G$ such that, $\forall d \in \mathscr{D}$, $\beta_k(d)$ and $\gamma_g(d)$ can be approximated by the expansion

$$\beta_k(\cdot) \approx \sum_{l=1}^{L_k} \zeta_{lk} B_{lk}(\cdot) \quad \text{for} \quad k = 0, \ldots, p \tag{3}$$

$$\gamma_g(\cdot) \approx \sum_{l=1}^{L_g} \eta_{lg} B'_{lg}(\cdot) \quad \text{for} \quad g = 1, \ldots, G. \tag{4}$$

Substituting $\beta_k(\cdot)$ and $\gamma_g(\cdot)$ of Eq (2) with Eqs (3) and (4), we approximated Eq (2) by

$$y_{ij} \approx \sum_{k=0}^{p} z_{ik} \sum_{l=1}^{L_k} \zeta_{lk} B_{lk}(d_{ij}) + \sum_{g=1}^{G} x_{ig} \sum_{l=1}^{L_g} \eta_{lg} B'_{lg}(d_{ij}) + \boldsymbol{\varepsilon}_{ij} \tag{5}$$

If $B_k(\cdot)$ and $B'_g(\cdot)$ are groups of B-spline basis functions of degree q_k and q_g respectively, and $\delta_0 < \delta_1 < \ldots < \delta_{K_k} < \delta_{K_k+1}$ and $\delta_0 < \delta_1 < \ldots < \delta_{K_g} < \delta_{K_g+1}$ are the corresponding knots, then $L_k = K_k + q_k$ and $L_g = K_g + q_g$.

Using the approximation Eq (5), the coefficients $\boldsymbol{\zeta} = (\zeta_0, \zeta_1, \ldots, \zeta_p)^T$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_G)^T$ can be estimated by minimizing the squared error

$$\ell_w((\boldsymbol{\zeta}, \boldsymbol{\eta})^T) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} w_{ij} \left[y_{ij} - \sum_{k=0}^{p} z_{ik} \sum_{l=1}^{K_k} \zeta_{lk} B_{lk}(d_{ij}) - \sum_{g=1}^{G} x_{ig} \sum_{l=1}^{L_g} \eta_{lg} B'_{lg}(d_{ij}) \right]^2 \tag{6}$$

where w_{ij} are known non-negative weights.

In cases where $p + G >> n$ though, minimisation of Eq (6) is infeasible. Our aim was to identify factors of the covariate vector $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_G)^T$ (genes) that are truly associated with the response (cancer cell line sensitivity to the drug). In addition, we wanted to explore potential dose-varying effects on the drug response.

We make the following sparsity assumption: any valid solution $\hat{\boldsymbol{\gamma}}(d)$ will have $\hat{\gamma_g}(d) = 0, \forall d \in \mathscr{D}$ for the majority of components g. To detect non-zero coefficient functions, we applied a two-stage approach which incorporated a variable screening step and a further variable selection step.

Screening. The sparsity assumption applies only to components of $\boldsymbol{x}$, the high-dimensional covariate vector in Eq (2).

Let the set of indices

$$\mathscr{M}_0 = \{1 \leq g \leq G : \|\gamma_g(\cdot)\|_2 > 0\} \tag{7}$$

where $\|\cdot\|_2$ is the L_2-norm. In order to rank the different components of x, we fitted the

marginal non-parametric regression model for the gth x-predictor:

$$y_{ij} \approx \sum_{k=0}^{p} z_{ik} \sum_{l=1}^{K_k} \zeta_{lk}^{(g)} B_{lk}^{(g)}(d_{ij}) + x_{ig} \sum_{l=1}^{L_g} \eta_{lg}^{(g)} B_{lg}^{(g)\prime}(d_{ij}) + \boldsymbol{\varepsilon}_{ij}^{(g)} \tag{8}$$

where: $\{B_{lk}^{(g)}(\cdot) : l = 1, \ldots, L_k\}$ and $\{B_{lg}^{(g)\prime}(\cdot) : l = 1, \ldots, L_g\}$ are sets of coefficient functions; $\{\zeta_{lk}^{(g)} : l = 1, \ldots, L_k\}$ and $\{\eta_{lg}^{(g)} : l = 1, \ldots, L_g\}$ are constants to be estimated, $k = 0, \ldots, p$; and, $\boldsymbol{\varepsilon}^{(g)}$ is the error term similar to Eq (5). We then computed the following weighted mean squared error for each $g \in \{1, \ldots, G\}$,

$$\hat{u}_g = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{y}_i - \hat{\mathbf{y}}_i^{(g)})^T \mathbf{W}_i (\mathbf{y}_i - \hat{\mathbf{y}}_i^{(g)}) \tag{9}$$

to quantify the importance of the gth variable. Here,

$$\boldsymbol{W}_i = \frac{1}{n_i} \hat{\boldsymbol{V}}_i^{-\frac{1}{2}} \boldsymbol{R}_i^{-1}(\hat{\boldsymbol{\phi}}) \hat{\boldsymbol{V}}_i^{-\frac{1}{2}} \tag{10}$$

where $\hat{\boldsymbol{V}}_i$ is the $n_i \times n_i$ diagonal matrix consisting of the dose-varying variance

$$\hat{\boldsymbol{V}}_i = \begin{bmatrix} \hat{V}(d_{i1}) & 0 & \cdots & 0 \\ 0 & \hat{V}(d_{i2}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{V}(d_{in_i}) \end{bmatrix} \tag{11}$$

and $\boldsymbol{R}_i(\phi) = (R_{jk})$ the $n_i \times n_i$ working correlation matrix for the ith subject. By ϕ, we denoted the $s \times 1$ vector that fully characterizes the correlation structure. The estimate of ϕ, $\hat{\boldsymbol{\phi}}$, was obtained by taking the moment estimators for the parameters ϕ in the correlation structure based on the residuals obtained from fitting the following model

$$y_{ij} = \sum_{k=0}^{p} z_{ik} \beta_k(d_{ij}) + \boldsymbol{\varepsilon}_{ij} \quad \text{where} \quad i = 1, \ldots, n, j = 1, \ldots, n_i. \tag{12}$$

The variance function $V(d)$ in Eq (11) was estimated using techniques described in [24].

After having obtained $\{\hat{u}_g : g = 1, \ldots, G\}$, we sorted gene utilities in an increasing order, where smaller $\hat{u}_g$ values indicate stronger marginal associations. The x-predictors included in the screened submodel are, then, given by

$$\widehat{\mathscr{M}}_{\tau_n} = \{1 \leq g \leq G : \ \hat{u}_g \text{ ranks among the first } \tau_n(v)\} \tag{13}$$

where $\tau_n(v)$ corresponds to the size of the submodel which is chosen to be smaller than the sample size n.

Variable selection using a group SCAD (gSCAD) penalty. Screening algorithms aim to discard all unimportant variables but tend to be conservative. In order to preserve only the most important x-predictors in the final model, we considered a model including the first

$\tau_n(v)$ outranked genes and we applied a gSCAD penalty by minimising the following criterion:

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}\left\{ y_{ij} - \sum_{k=0}^{p} z_{ik}\sum_{l=1}^{L_k}\zeta_{lk}B_{lk}(d_{ij}) - \right. \tag{14}$$

$$\left. \sum_{g\in\widehat{\mathscr{M}}_{\tau_n}} x_{ig}\sum_{l=1}^{L_g}\eta_{lg}B'_{lg}(d_{ij}) \right\}^2 + \sum_{g\in\widehat{\mathscr{M}}_{\tau_n}} p_{\lambda,\alpha}(\|\eta_g\|) \tag{15}$$

where

$$p_{\lambda,\alpha}(u) = \begin{cases} \lambda u & \text{if } 0 \le u \le \lambda \\ -\frac{(u^2 - 2\alpha\lambda u + \lambda^2)}{2(\alpha-1)} & \text{if } \lambda \le u \le \alpha\lambda \\ \frac{(\alpha+1)\lambda^2}{2} & \text{if } u \ge \alpha\lambda \end{cases},$$

α is a scale parameter, λ controls for the penalty size and $\|\cdot\|$ is the Euclidean ℓ_2-norm. At this point, note that grouping is applied for the coefficients η_g that correspond to the same coefficient function. In addition, in order to reduce the bias introduced when applying a LASSO penalty, we alternatively chose the SCAD, which coincides with the LASSO until $u = \lambda$, then transits to a quadratic function until $u = \alpha\lambda$ and then it remains constant $\forall u > \alpha\lambda$, meaning that it retains the penalization and bias rates of the LASSO for small coefficients but at the same time relaxes the rate of penalization as the absolute value of the coefficients increases. In Fig 1 the reader can find a brief overview of the employed methodology.

Tuning parameter selection

We used knots placed at the median of the observed data values along with cubic B-splines with 1 interior knot. The suitable number of interior knots was calculated using the formula $N_n = \left[n^{\frac{1}{2p+3}}\right]$ proposed and applied by [19, 28, 29]. Due to the computational burden this would add, we did not apply cross-validation.

As for the screening threshold τ_n, its magnitude could be determined by the fraction $v\left[\frac{n}{log(n)}\right]$, $v \in \{1, 2, 3, \ldots\}$. We conducted a pilot simulation study in order to decide the most appropriate size (for further details see S1 Text). We also considered an automated algorithm for its selection (Greedy Iterative Non-parametric Independence Screening-Greedy INIS, [21]). Finally, the penalty size for the gSCAD step λ was determined using a 5-fold cross-validation.

Simulation study

Monte Carlo simulations were conducted to examine the ability of our model to detect the genes that are truly associated with the drug response. This had a key role in tuning model parameters and simultaneously assessing model goodness-of-fit using a fraction of the original data set in order to reduce the computational burden of conducting a simulation study under the original dimensions of the data. Responses over different dosage levels were generated based on a subset of genes, the corresponding low-dimensional GDSC data covariates (drug and cancer type) and some prespecified smooth coefficient functions (see S1 Text). In particular, we repeatedly sampled without replacement 190 experimental units and 886 genes based

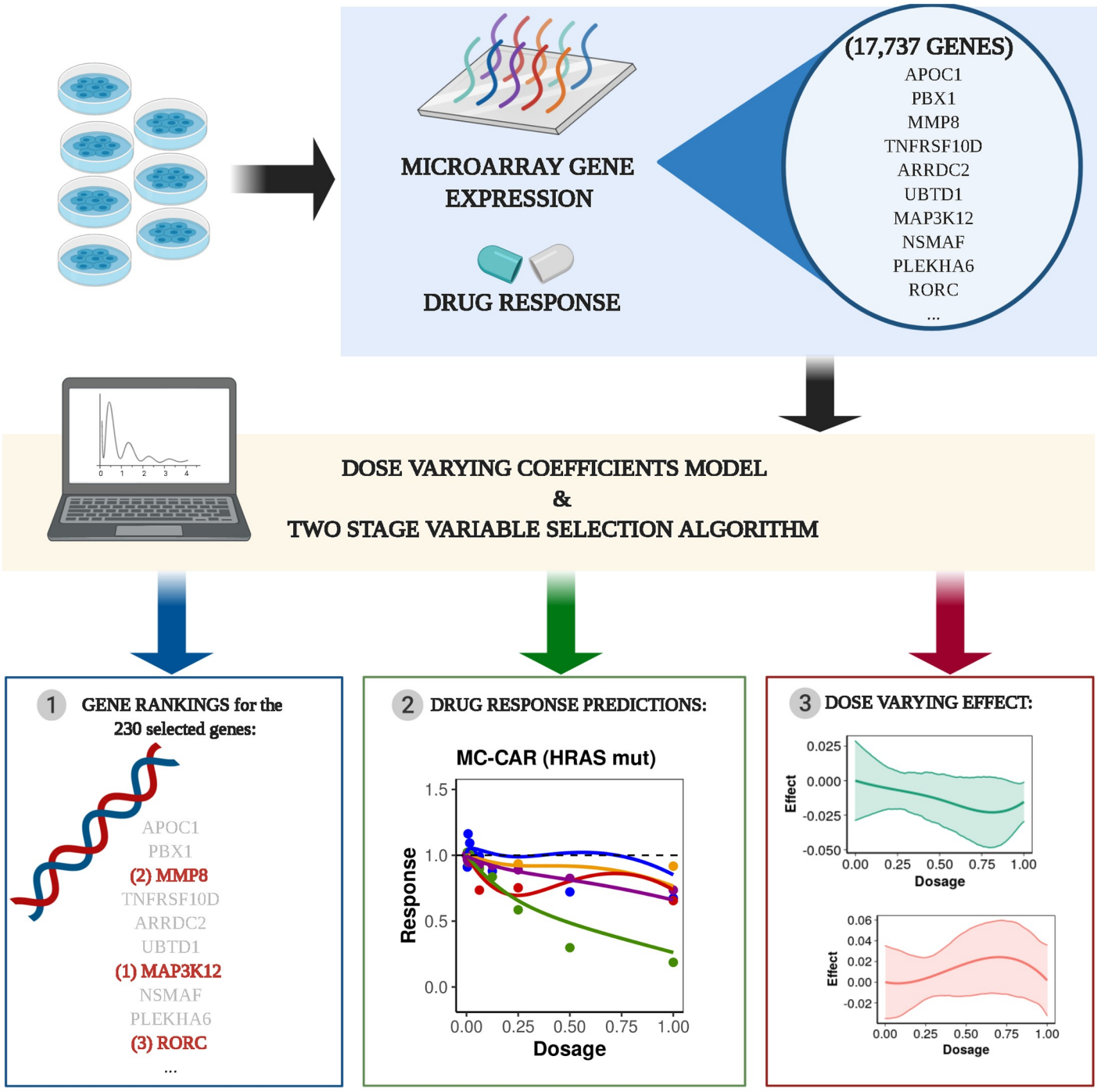

Fig 1. The two-stage algorithm for identifying dose-dependent associations between genes and drugs. Gene expression and drug response data from a drug screening study (e.g. GDSC) are used to fit our dose-varying coefficients model to estimate the dose-varying effect between covariates and drug response. A two-stage variable screening and selection algorithm is applied to rank gene-drug associations. The selected genes can then be used to predict dose-dependent response for the drugs of interest.

https://doi.org/10.1371/journal.pcbi.1008066.g001

on which the simulated responses have been generated. The performance of the employed methodology has been assessed based on 1000 simulations using three screening thresholds ($\tau_n(v) = \left[\frac{n}{log(n)}\right]$, $\tau_n(v) = \left[\frac{2n}{log(n)}\right]$ and $\tau_n(v)$ chosen using the greedy-INIS algorithm [21]) and two estimated covariance structure scenarios (independence and rational quadratic covariance structure). Cubic B-splines and knots placed at the median of the observed data values have been used for estimating the coefficient functions.

To evaluate the performance of the proposed procedure we used the following summary measures: TP–number of genes correctly identified as active; FP–number of the genes incorrectly identified as active; TN–number of the genes correctly identified as inactive; FN–number of the genes incorrectly identified as inactive.

Simulation results suggested that our method accurately detects the drug associated genes from the simulated responses under most of the examined scenarios (S1 Text). A screening threshold of size $\left[\frac{2n}{log(n)}\right]$ and regression weights adjusted for the covariance structure of the data were identified as the scenario where our method reached its maximum accuracy. Consequently, for the GDSC application, we chose the screening threshold to be the maximum possible, i.e. 923 genes derived from the formula $\left[\frac{2n}{log(n)}\right]$, and weights derived by assuming a rational quadratic covariance structure for the repeated measures.

Software availability

The analysis has been conducted using R version 3.6.3. Code for applying the two-stage variable selection algorithm is available online as an R package at https://github.com/koukoulEv/fbioSelect.

Results and discussion

Dose-dependent associations with gene expression in a large-scale drug sensitivity assay

We applied the two stage variable selection algorithm under the dose-varying coefficient model framework described above. Gene rankings and predicted mean drug effects over different dosage levels were obtained. Our algorithm identified 230 candidate genes associated with drug response. The effect of each of those genes was assessed with respect to:

1. the area under the estimated coefficient curve (AUC) and its corresponding standard deviation (estimated using bootstrapping);
2. the effect on cell survival (overall positive, overall negative, mixed);
3. Spearman correlation between the coefficient function value and the dosage level;
4. the mean fold change of the expression of cell lines carrying *BRAF* mutations with respect to wild type; and,
5. the protein-protein interaction network distance between the *BRAF* gene and the selected genes using the Omnipath database [30].

The 230 genes were ranked based on the estimated AUC value (S1 Table), and the top 30 genes were highlighted for further analysis (Table 1). The higher the AUC, the larger the effect of the gene on the drug response. The overall effect on cell survival can be either positive, negative or vary over the different dosage levels as determined by the range of the estimated

Table 1. Top 30 gene rankings based on the estimated area under the coefficient function curve.

Gene Name	Area	SD	Sign	Spearman's Correlation	Mean fold change in *BRAF* mutant vs wild-type cell lines	Protein-protein interaction network distance to *BRAF*
KIR3DL1	0.370	0.107	-	-0.874	0.978	3
CHST11	0.257	0.092	-	-0.817	0.899	NI
APOC1P1	0.247	0.09	-	-0.918	1.190	NI
PLEKHA6	0.239	0.086	-	-0.908	1.037	3
PPM1F	0.223	0.068	+	0.910	0.883	3
BFSP1	0.222	0.074	-	-0.800	1.217	NI
PPP1R3A	0.217	0.082	+	0.774	1.078	3
C16orf87	0.207	0.087	+	0.851	0.977	NI
PARVA	0.203	0.081	+	0.890	0.984	2
SLC39A13	0.202	0.079	-	-0.461	1.055	NI
UCN2	0.198	0.07	-	-0.928	0.979	NI
STMN3	0.198	0.087	+	0.834	1.201	2
RNF130	0.197	0.083	-	-0.927	1.153	NI
C3orf58	0.196	0.076	+	0.922	1.133	NI
CXXC4	0.188	0.079	+	0.866	0.995	NI
THBD	0.179	0.093	0	-0.967	1.231	4
SIRT3	0.173	0.066	-	-0.760	1.013	3
PLAT	0.172	0.092	-	-0.878	1.322	4
MPPED1	0.168	0.066	+	0.430	0.978	NI
INSL3	0.162	0.068	-	-0.973	0.965	NI
FAM163A	0.159	0.078	-	-0.983	1.106	NI
CNIH3	0.153	0.08	-	-0.918	0.938	NI
GJA3	0.153	0.067	0	-0.940	0.933	NI
BTG2	0.152	0.078	+	0.959	1.035	2
DLX6	0.152	0.059	0	0.686	0.987	NI
DLC1	0.151	0.053	-	-0.928	0.974	3
GAPDHS	0.150	0.077	+	0.886	1.232	NI
JAG2	0.149	0.069	-	-0.994	0.981	3
SMOX	0.146	0.057	0	0.816	1.070	NI
ZMYND8	0.145	0.091	+	0.907	1.020	3

Gene rankings of the top 30 selected genes based on the magnitude of the genetic effect on drug response. A positive (+) sign translates to a positive effect on cell survival after drug administration, a negative (-) sign translates to a negative effect on cell survival and a mixed (0) effect translates to a varying effect on cell survival which depends on drug dosage. Spearman correlation is calculated between drug dosage and gene estimated coefficient function values as an indicator of the magnitude change of the gene effect over the increasing dosage. Area corresponds to the area under the estimated coefficient curve and the SD corresponds to the standard deviation of the area based on bootstrapping. Mean fold change is calculated between the selected gene expression values of the cell lines carrying *BRAF* mutations with respect to wild type. Protein-protein interaction network distance is computed based on the shortest interaction path between the *BRAF* gene and each of the selected genes. Here, NI denotes absence of any interaction.

https://doi.org/10.1371/journal.pcbi.1008066.t001

coefficient function. Spearman's rank correlation was used as an indicator of the coefficient function's monotonicity by characterising the progress of the genetic effect over different dosage levels. For instance, high expression of the *C3orf58* gene at baseline has a positive effect on cancer cell survival, which becomes stronger as the dosage increases (Spearman correlation = 0.922). In other words, high expression of this gene can be an indicator of drug resistant cell lines. On the other hand, the *DLC1* gene has an overall decreasing and negative effect on cancer cell survival (Spearman correlation = -0.928) which suggested that as the dosage

increases, higher baseline expression of this gene can indicate higher drug sensitivity at higher dosage. Elevated expression of DLC1 has been observed in melanoma and is a well known tumour suppressor that could be a novel marker of BRAF inhibition [31]. Finally, in cases where the overall effect varies (changes between positive and negative), the effect of gene expression on the drug response depends on the drug dosage. In particular, the effect of *DLX6* increases and then decreases at higher dosages (Fig 2). Given the biological and technical variation in drug screens, we should treat the mean effect estimates with caution and consider the confidence intervals of the coefficient functions in order to derive conclusions about the exact effect of the selected genes on the dose response (Fig 2).

Coefficient function estimates provide a lot of information about the dosage, cancer type and genetic effects on drug response. Fig 2 illustrates the estimated coefficient functions for different drugs, cancer types and three genes in relation to the model intercept, Dabrafenib response in *BRAF* mutant cell lines originating from the skin (melanoma). Except from HG6-64-1, all other BRAF inhibitors (AZ628, SB590885 and PLX4720) showed no additional effect compared to the intercept. Similar patterns can be observed for cancer cell lines coming from most of the tissues examined. This result indicates that the examined drugs may have similar or worse behaviour over the different dosages for most of the examined cancer types. We observed greater efficacy (negative values of the coefficient function) for cell lines originating from the endocrine system, autonomic ganglia and heamatopoietic and lymphoid tissues at lower dosages. The observed effect in endocrine system cell lines reflects the Dabrafenib responses observed in anaplastic thyroid cancer patients [32]. Interestingly, the drug Trametinib, taken in combination with Dabrafenib is a MEK inhibitor, and genes interacting with MEK (*MAP2K1*) were selected features from our model (Fig 3A). Together these results provide important insights into the effectiveness of the five *BRAF* targeted drugs examined on different cancer types, highlighting the potential for effective treatment of a wide range of cancers given the tumour genetic characteristics.

Since the *BRAF* gene is the target of the drugs, mean fold change and protein-protein interaction network distance were used to examine whether and how the selected genes are related to the target of the inhibition. From the selected genes, 120 genes had a mean fold change greater than 1 whereas the rest had a mean fold change between 1 and 0.792. Some of the genes with the highest mean fold change of *BRAF* mutation were *PSMC3IP*, *KIF3C*, *UBE2Q2*, *SERPIND1* and *PLAT*, however only *PLAT* is displayed in Table 1. From the genes identified through the two-stage algorithm, 35% of them encode proteins interacting with the *BRAF* gene, though none of them directly. Most of the selected genes interact with the *BRAF* gene via pathways mediated by *HRAS*, *MAPK1* (*ERK*), *MAP2K1* (*MEK*) and *BAD* (Fig 3A).

Since *HRAS* mutations are frequent in patients receiving *BRAF* targeted therapies [33], we examined the mean estimated trajectory over different dosages under treatment with *BRAF* inhibitors tested in six cancer cell lines with and without *BRAF* and *HRAS* mutations (Fig 4). As stated previously, we observed that in most cases HG6-64-1 seems to be the most effective drug. The estimated coefficient functions facilitate drug examination and response prediction under the different dosages. In some instances, we observed different drugs having similar behaviour for lower drug dosages and larger divergence for higher dosages. In most cases, regardless of the cell line origin, our method successfully estimates the expected survival rates of the cancer cell lines for the different drugs given their gene expression information.

For validation purposes, we performed the same analysis using the independently-generated GDSC2 data set, with a different set of drugs. Note that the drug set in GDSC2 only partially overlaps with the one used in GDSC1. The results are reported in supplementary S2 Fig, and show similar properties to the analysis of the BRAF drugs in Fig 4. As the GDSC2 dose-responses are produced from independently-generated experiments, the measured drug

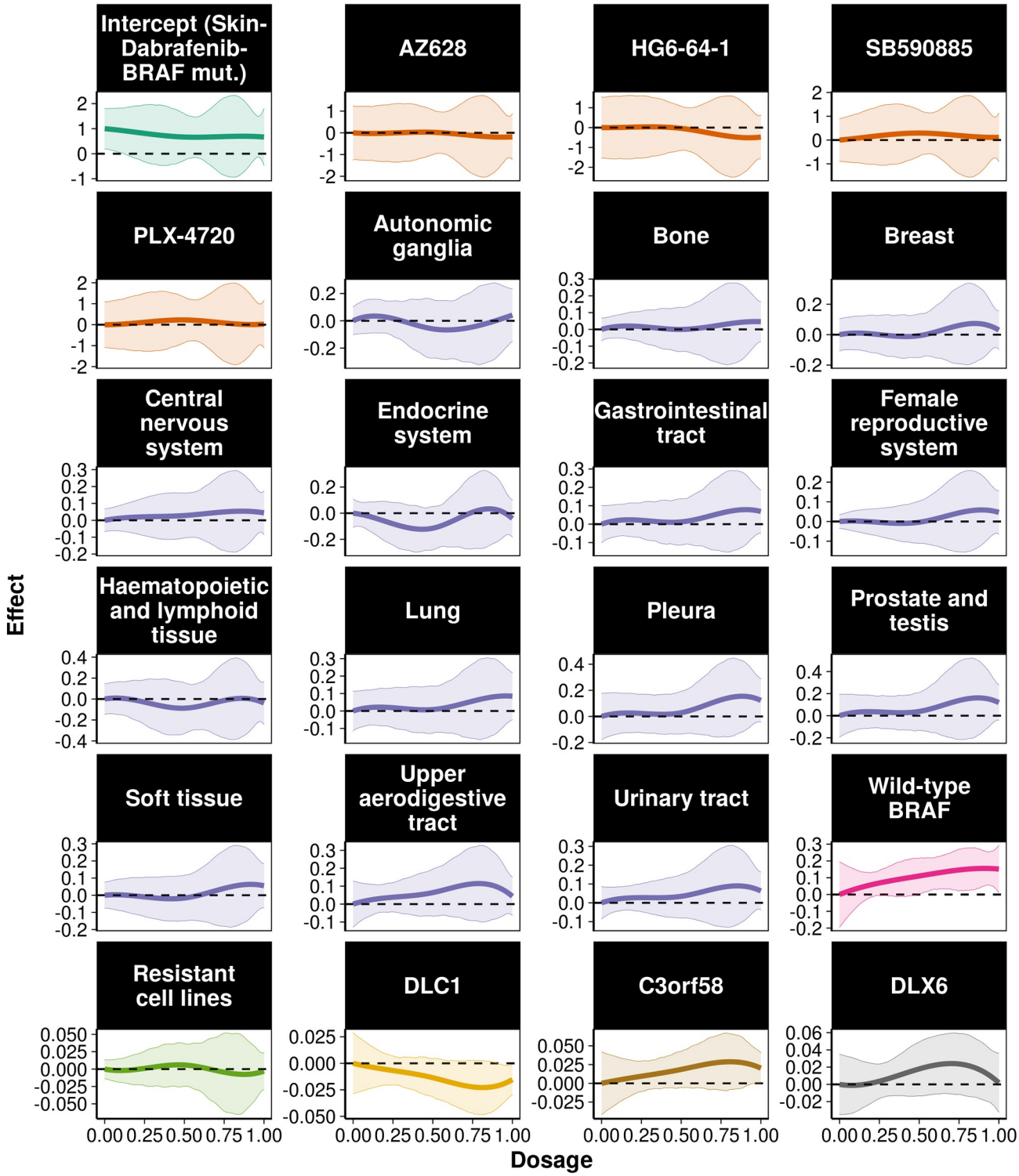

Fig 2. Estimated coefficient functions for the low-dimensional predictors and three of the selected genes. Estimated coefficient functions for the intercept, different drugs, tissue of origin and three of the selected genes along with 95% bootstrap confidence intervals. Baseline corresponds to *BRAF* mutant cell lines treated with Dabrafenib in skin tumours.

https://doi.org/10.1371/journal.pcbi.1008066.g002

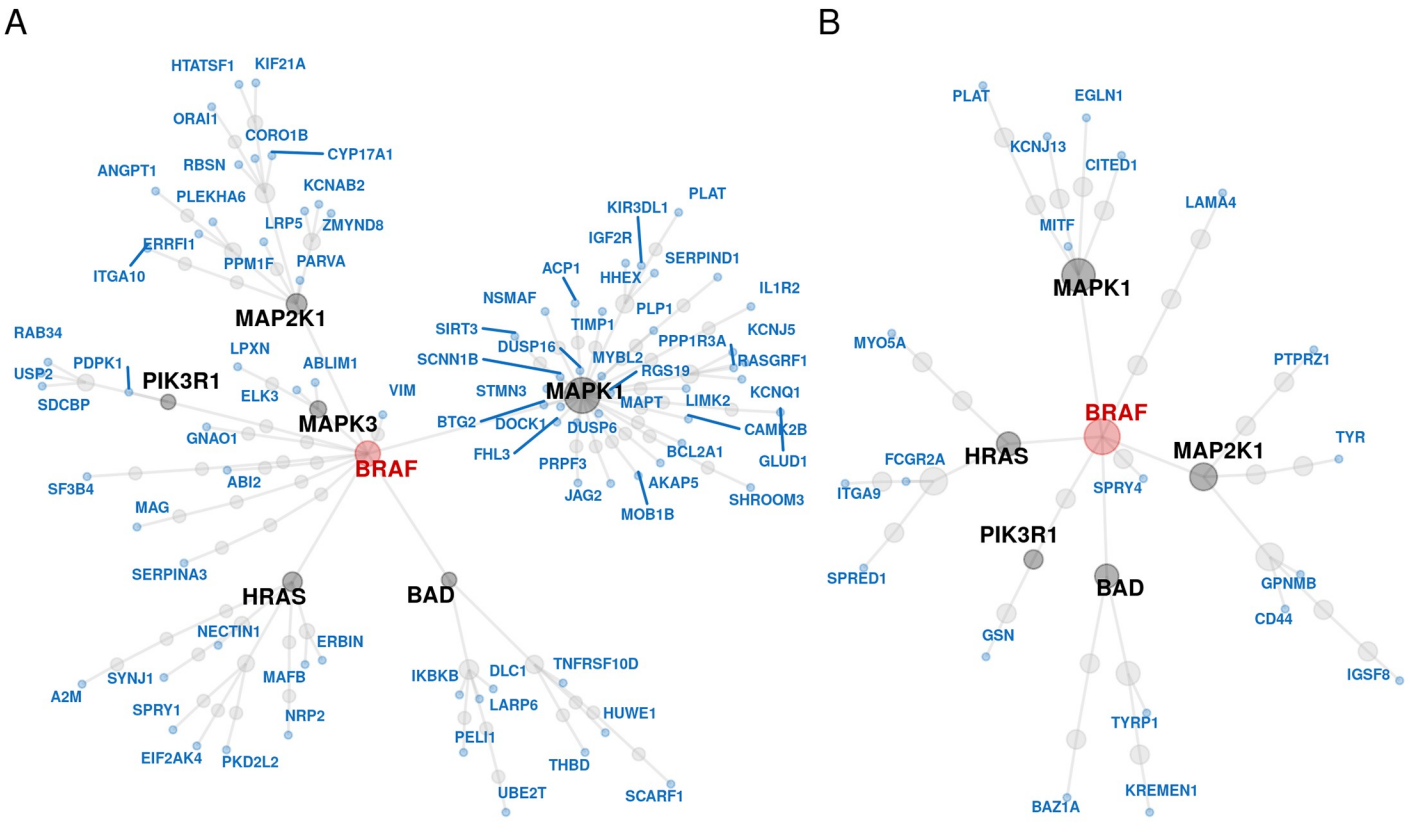

Fig 3. Protein-protein interaction network for the genes selected from the two-stage variable selection algorithm. (A) Undirected protein-protein interaction network between the 230 selected (blue) and the *BRAF* (red) genes (full scale analysis). (B) Undirected protein-protein interaction network between the 65 genes selected from the two-stage variable selection algorithm for the cell lines resistant to BRAF inhibitors (blue) and the *BRAF* (red) gene. In both panels genes depicted with black are the interaction mediators. Common mediators include the *HRAS*, *MAPK1*, *MAP2K1* and *BAD* genes.

https://doi.org/10.1371/journal.pcbi.1008066.g003

response is different for some of the drug-cell line combinations. We observe some divergences between GDSC1 and GDCS2 estimated trajectories for Dabrafenib and PLX-4720, which can be explained due to measurement error in the GDSC experiments themselves.

Variable selection algorithm identifies cancer pathways associated with BRAF inhibitor response

Using our functional regression approach, we identified 230 genes that were selected via the SCAD step (observed gene set). We used the Enrichr [34, 35] and WikiPathways [36] databases to see if the selected genes can be grouped into common functional classes or pathways. In total, 183 were identified, of which 11 were statistically significant at 5% level, including apoptosis modulation, NOTCH1 regulation, and MAPK signaling (S2 Table). The model identified genes (*IKBKB*, *RASGRF1*, *DUSP16*, *DUSP8*, *DUSP6*, *MAPT* and *IL1R2*) downstream of the MAPK signaling pathway targeted by BRAF inhibitors.

Previous studies of these pathways have found associations with tumorigenesis and cancer treatment [37–40]. Genes in more than one of these pathways include *IKBKB*, *PLAT*, *IL1R2* and *PDPK1*. The IKB kinase composed of IKBKB had previously been suggested as a marker of sensitivity for combination therapy with BRAF inhibitors [41]. Taken together, these results suggest that the identified associations between the drug response and the observed genes may reveal new predictive markers of tumour response to the examined BRAF inhibitors.

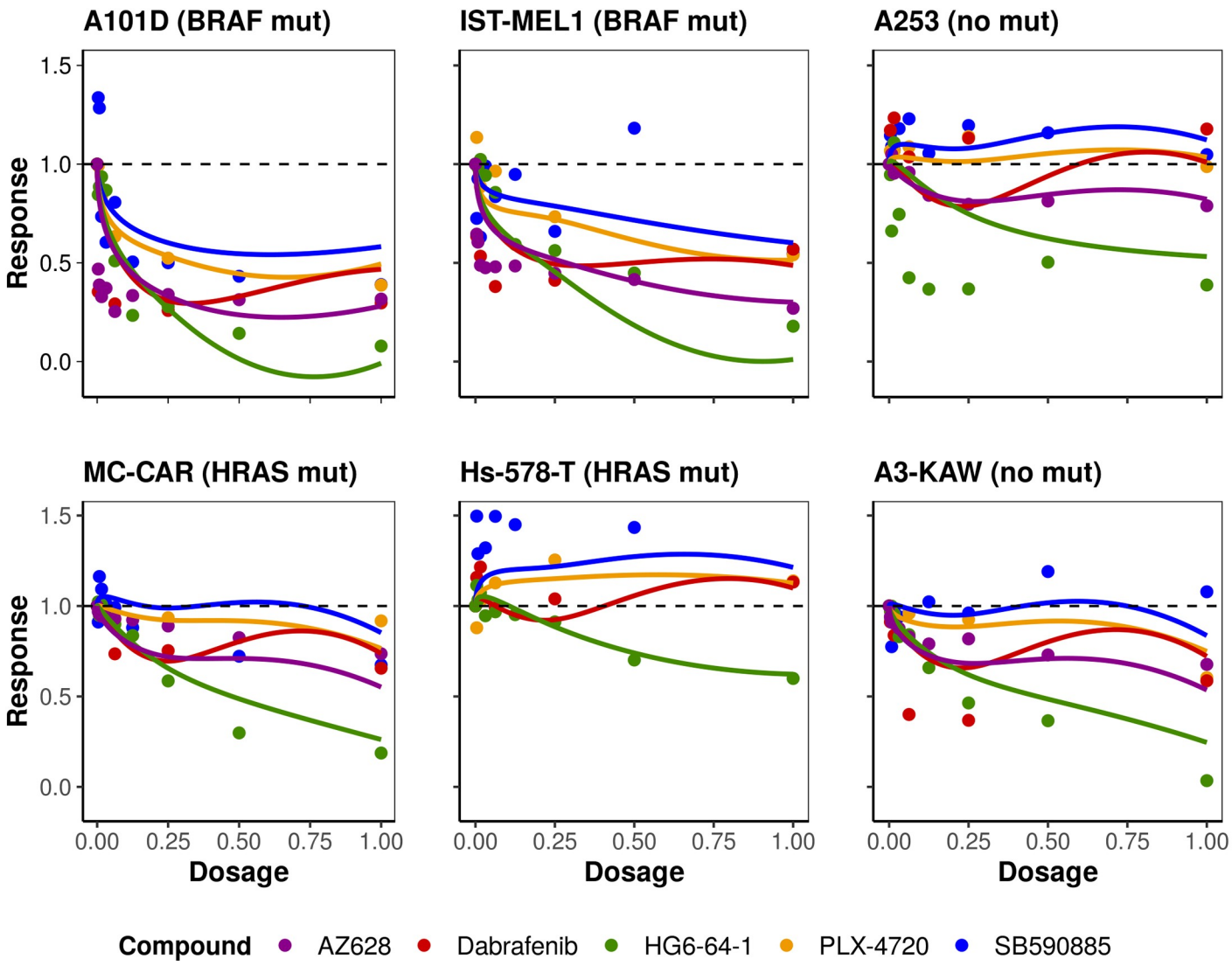

Fig 4. Estimated mean drug response trajectories for six cancer cell lines with *BRAF* and *HRAS* mutations. Observed responses (points) and estimated mean trajectory (lines) of cell concentration for cancer cell lines with and without *BRAF* and *HRAS* mutations after treatment with the five anticancer compounds examined.

https://doi.org/10.1371/journal.pcbi.1008066.g004

In addition to the pathway enrichment analysis, we used the Molecular Signatures Database (MSigDB database v7.0 updated August 2019: [42]) to compute overlaps between the observed gene set and known oncogenic gene sets. Fig 5A and S3 Table display the 29 overlaps found. Interestingly, we identified three instances where the observed gene set significantly overlapped with gene sets over-expressing an oncogenic form of the *KRAS* gene.

We further explored potential biologically relevant pathways using the Reactome database [44, 45]. More than 40 enriched pathways were identified at a 5% significance level. The top 40 pathways are depicted in Fig 6 along with the pathway-gene network of the top 5 pathways (for the full list, see S4 Table). Interestingly, axon guidance and VEGF signalling were among the enriched pathways, confirming relevance of the selected genes to the intended role of the examined compounds, since BRAF kinase activity drives axon growth in the central nervous system [46] and VEGF blockade has potential anti-tumour effects when combined with BRAF inhibitors [47]. Note that axon growth has not, thus far, been directly implicated in BRAF

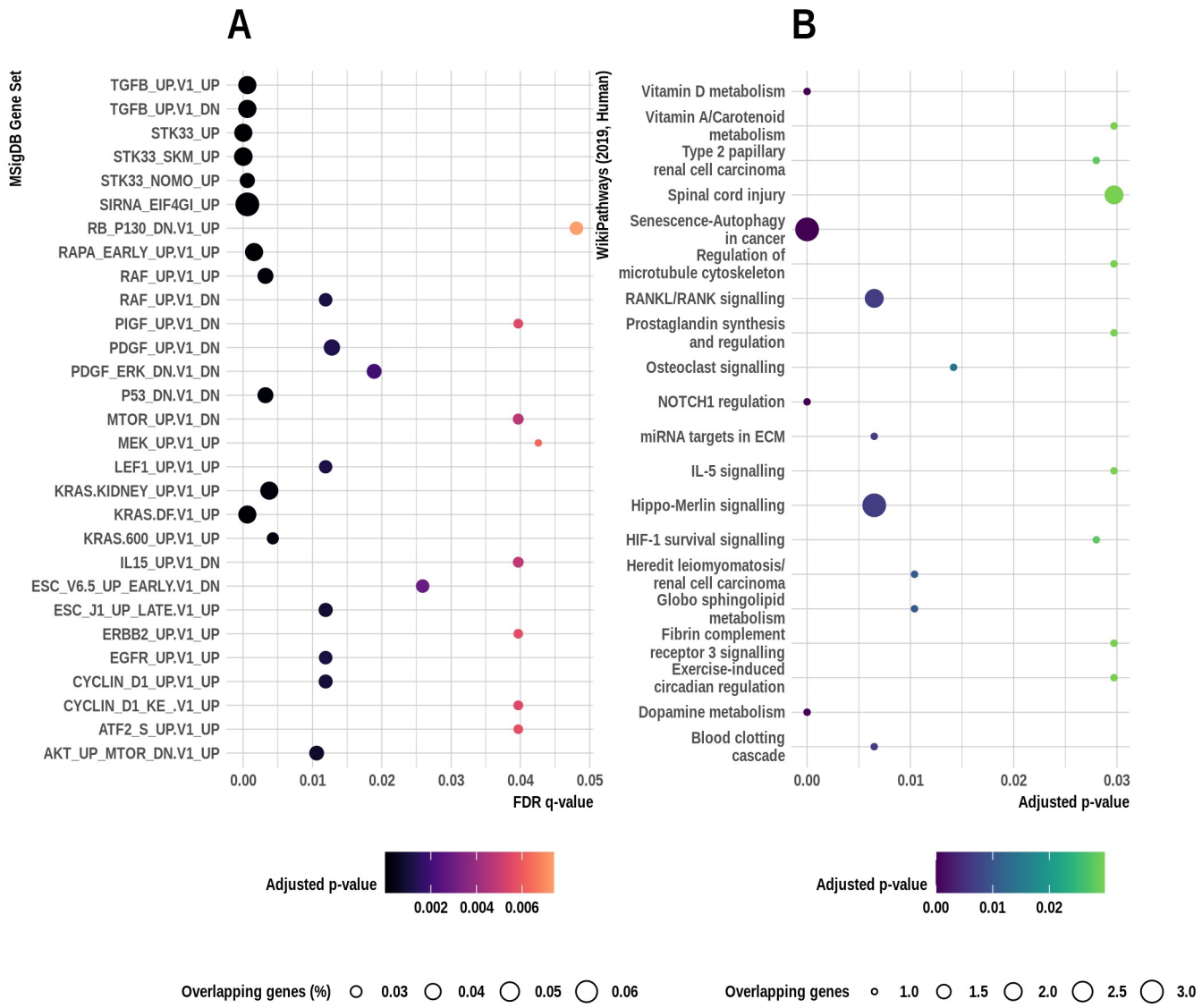

Fig 5. Overlaps between the observed gene set and oncogenic signatures in the Molecular Signatures Database (full data analysis); signalling pathways enriched for genes predictive of *BRAF* inhibitor response (resistant cell lines). (A) Full gene set names can be found in S3 Table. Overlaps have been detected using gene set enrichment analysis performed using a hypergeometric distribution. The false discovery rate analog of the hypergeometric p-value is displayed after correction for multiple hypothesis testing according to Benjamini and Hochberg [43]. (B) Top 20 enriched signalling pathways along with the adjusted p-values and the number of overlapping genes obtained after pathway enrichment analysis to the resistant cell line analysis results (for the full list of the pathways identified see S6 Table).

https://doi.org/10.1371/journal.pcbi.1008066.g005

inhibitor activity, and, consequently, our analysis provides important evidence towards this theory.

Identifying dose-dependent genes in drug-resistance conditions

Acquired resistance to BRAF inhibitors is often observed in the clinic [48]. To further examine the utility of the employed methodology, we applied the variable selection algorithm to a data subset containing only cell lines with mutations activating resistance mechanisms to BRAF inhibitors [49]. Out of the 951 cell lines in the data, 191 had some mutation in any of the following: *RAC1* gene, *NRAS* gene, *cnaPANCAN44* or *cnaPANCAN315*. We identified 65 genes

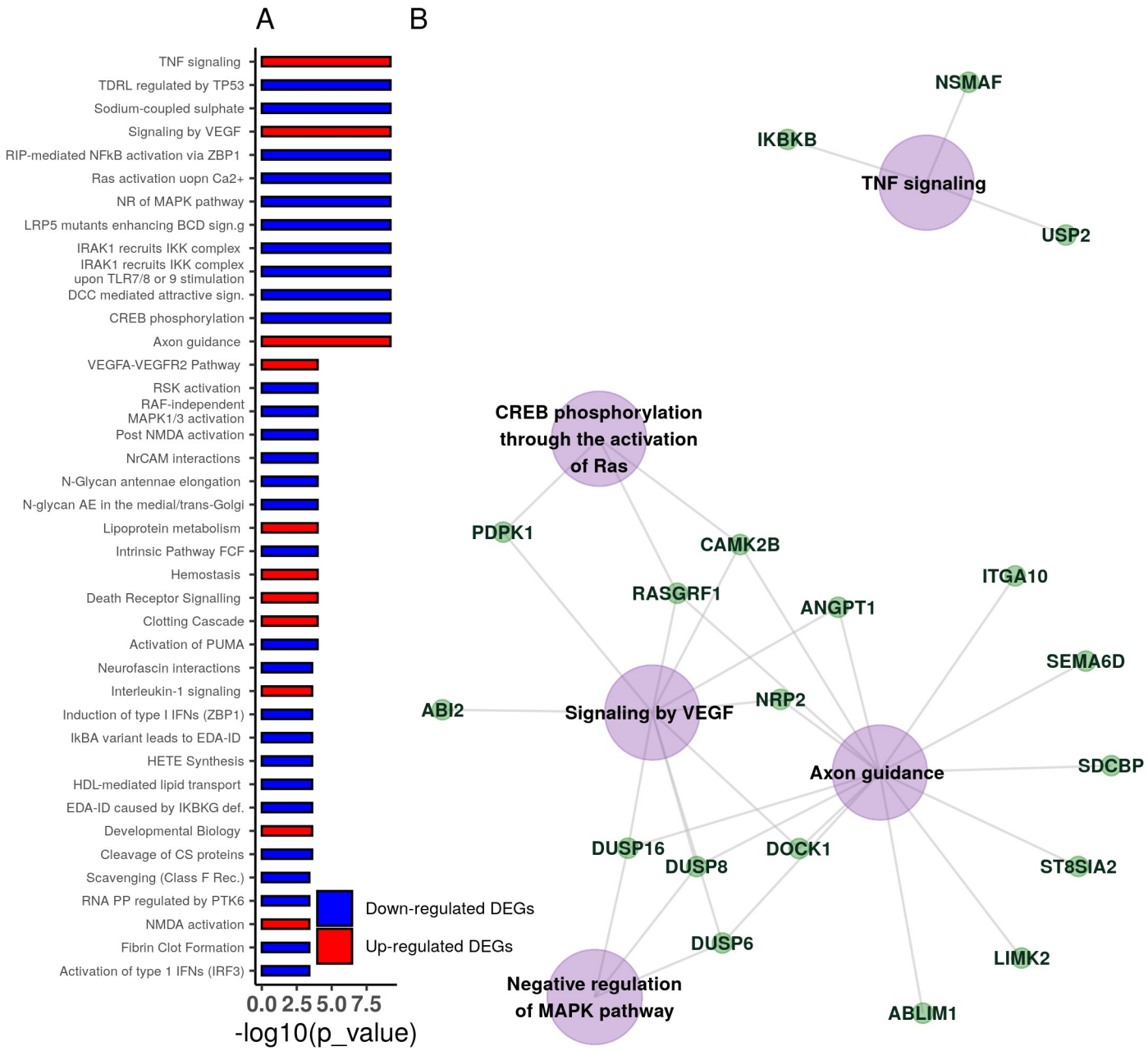

Fig 6. Pathway enrichment analysis using the Reactome database. (A) Top 40 enriched signalling pathways along with the adjusted p-values. For the full list, see S4 Table. (B) Pathway-gene network of the top 5 enriched signalling pathways as found using Reactome [44, 45].

https://doi.org/10.1371/journal.pcbi.1008066.g006

associated with dose-response, though none of them were directly associated with the MAPK/ERK pathway. However, from these, 25 genes have been found to indirectly interact with the *BRAF* gene (Fig 3B) and 21 to overlap with three oncogenic gene sets in the Molecular Signatures Database (genes down-regulated in NCI-60 panel of cell lines with mutated *TP53*; genes up-regulated in Sez-4 cells (T lymphocyte) that were first starved of *IL2* and then stimulated with *IL21*, and; genes down-regulated in mouse fibroblasts over-expressing *E2F1* gene; S5 Table). Finally, we found 34 pathways enriched for genes predicting drug response of the

mutated cell lines to the examined BRAF inhibitors, of which the top 20 are depicted in Fig 5B (for the full list of the pathways identified see S6 Table).

Table 2 presents gene rankings based on the AUC and the overall coefficient function effect (sign) for the 42 genes in either the enriched pathways, the three oncogenic gene sets discussed above or the protein-protein interaction network with the *BRAF* gene (full list available in S7 Table). Eight of the selected genes in the current implementation were also selected from the algorithm implemented on the full data: *ASB9*, *PRSS33*, *GJA3*, *PLAT*, *KLF9*, *BFSP1*, *MTARC1* and *UCN2*.

Predictive performance of dose-dependent models

As discussed above, the employed methodology gives a good overview of the baseline genetic effect on drug response. We assessed the overall predictive performance of our method using 10-fold cross validation under two different scenarios. For the first, we split the data into training and test set holding out the experimental units (cancer cell line-drug combinations) and for the second, holding out cancer cell lines. The absolute mean error for both cases was around 0.12. Our analysis showed robust cross-validated performance when it comes to predicting sensitivity to the administered drugs, as shown in S3 Fig which displays the correlation between predicted and true response. This result was further validated by repeating the analysis on the independently-generated GDSC2 data set, using a different set of drugs (S4 Fig), which demonstrated comparable predictive performance.

Predictive accuracy for the dose-response curves was evaluated under four different subscenarios: prediction of the most effective drug-dosage combination for the 951 cell lines in the data set; prediction of the most effective drug given a cell line; prediction of the most effective dosage given treatment with a particular drug and prediction of the most effective dosage range given treatment with a drug (Table 3). The proposed model performs well when it comes to predicting the most effective drug or dosage range (≈79% in both scenarios). Results are less reliable when it comes to prediction of the exact dosage or drug-dosage combination (≈48-49% and ≈57-58% in both scenarios) but this can be due to either the large variability observed in the observed responses or due to the small number of cell lines for some predictor level combinations. Results were similar for both cross-validation scenarios (differences range from 0 to <2%, Table 3), meaning that as long as a cell line has similar genetic characteristics to those observed, the model can be reliable in predicting the outcome after anticancer drug administration.

We additionally compared the performance of our two-stage algorithm approach to a penalized linear (LASSO) regression for predicting the IC50 and area under the dose-response curve (AUC). Note that our functional regression model is not directly predicting either of these values, but rather predicts the full drug-response curve. As this is a harder problem, we would expect the LASSO to have a natural advantage; however, our method has the added benefit of being able to detect dose-dependent associations, which is not possible when predicting summary statistics of the dose-response curve directly. We employed 10-fold cross-validation to evaluate the predictive error in terms of root mean squared error (RMSE), and we used a sigmoid curve fit for estimating the IC50 values from the predicted dose-response curves with our two-stage method. Our method outperformed standard LASSO in terms of predicting the AUC ($RMSE_{2-stage} = 0.176$; $RMSE_{LASSO} = 0.347$) and performed well on predicting the IC50, although the LASSO performed better ($RMSE_{2-stage} = 1.969$; $RMSE_{LASSO} = 1.134$). This could be expected, as estimating the IC50 from the predicted dose-response curve adds a further level of complexity, compared to directly predicting this value using the LASSO.

Table 2. Rankings of the genes identified from the pathway and oncogenic gene set enrichment analysis.

Gene Name	Area	SD	Sign	Spearman Correlation	Mean fold change in *BRAF* mutant vs wild-type cell lines	Protein-protein interaction network distance to BRAF
MYO5A	0.531	0.261	+	0.955	1.358	4
S100A1	0.488	0.189	+	0.812	1.263	NI
GPNMB	0.424	0.196	+	1	1.169	3
ACP5	0.359	0.149	-	-0.998	1.039	NI
FCGR2A	0.341	0.158	-	-0.588	1.25	3
CITED1	0.28	0.348	0	-0.603	1.63	3
SPRY4	0.274	0.127	-	-0.611	1.228	2
CD44	0.239	0.164	+	0.868	1.413	3
RAP2B	0.236	0.179	0	0.927	1.254	NI
KCNJ13	0.205	0.094	0	-0.604	1.101	3
ALX1	0.202	0.099	-	-1	1.104	NI
PLAT	0.201	0.121	-	-0.405	1.312	4
RETSAT	0.201	0.142	0	0.689	1.127	NI
GSN	0.196	0.109	+	0.588	1.079	4
CDH19	0.185	0.102	0	0.943	0.933	NI
ATP1B3	0.178	0.115	-	-1	1.063	NI
BAZ1A	0.173	0.105	+	-0.29	1.109	4
SLC16A4	0.166	0.117	-	-0.298	1.234	NI
ST6GALNAC2	0.164	0.102	0	-0.815	1.264	NI
MFSD12	0.16	0.148	0	-0.788	1.13	NI
GJA3	0.157	0.075	0	-0.85	1.071	NI
CYP27A1	0.156	0.09	-	-0.743	1.373	NI
EGLN1	0.15	0.119	-	-0.442	1.053	3
TRPV2	0.147	0.118	0	0.769	1.074	NI
MITF	0.146	0.106	+	1	0.743	2
TBC1D7	0.146	0.118	0	-0.603	1.304	NI
SLC6A8	0.144	0.111	0	-0.263	0.941	NI
PTPRZ1	0.139	0.138	-	-0.808	1.074	4
PLOD3	0.132	0.135	0	0.696	1.166	NI
ANKRD7	0.131	0.12	+	0.92	1.241	NI
KANK1	0.107	0.113	0	-0.493	1.345	NI
GYPC	0.105	0.092	+	-0.3	1.072	NI
TYR	0.1	0.098	-	0.467	1.11	4
TYRP1	0.1	0.097	0	0.457	1.326	3
IGSF8	0.09	0.129	0	-0.668	1.313	5
SPRED1	0.067	0.116	0	-0.556	1.239	4
ITGA9	0.056	0.111	0	0.785	1.154	4
KREMEN1	0.053	0.086	0	-0.555	1.123	4
LAMA4	0.038	0.083	-	0.344	1.151	4
MLANA	0.037	0.097	0	0.534	1.147	NI
KLF9	0.011	0.074	0	0.932	1.064	NI

Table notes rankings of the genes found to have some biological importance. A positive (+) sign translates to a positive effect on cell survival after drug administration, a negative (-) sign translates to a negative effect on cell survival and a neutral (0) effect translates to a varying effect on cell survival which depends on drug dosage. Spearman correlation is calculated between drug dosage and gene estimated coefficient function values as an indicator of the magnitude change of the gene effect over the increasing dosage. Area corresponds to the area under the estimated coefficient curve and the SD corresponds to the standard deviation of the area based on bootstrapping. Pearson's correlation is calculated between the selected gene microarray expression values and the *BRAF* expression across all the cell lines. Protein-protein interaction network distance is computed based on the shortest interaction path between the *BRAF* gene and each of the selected genes. Here, NI denotes absence of interaction.

https://doi.org/10.1371/journal.pcbi.1008066.t002

Table 3. Predictive performance of the employed model (mean absolute error = 0.121).

Scenario	Accuracy EU	Accuracy CL
Model predicts the more effective drug-dosage combination	57.85%	57.42%
Model predicts the more effective drug given a cell-line	78.21%	78.21%
Model predicts the more effective dosage given a drug	48.44%	48.65%
Model predicts the most effective dosage range ($>$ or $\leq$ 31.25% of the maximum dosage)	79.47%	79.28%

Table notes the predictive performance of the model based on the percentages for correctly identifying the most effective drug, dosage or drug-dosage combinations. Results obtained based on 10-fold cross-validation of the final model (based on holding out either experimental units–EU– or cancer cell lines–CL–).

https://doi.org/10.1371/journal.pcbi.1008066.t003

Conclusion

Genetic alternations and gene expression in tumours are known to affect disease progression and response to treatment. Here, we studied dosage-dependent associations between gene expression and drug response, using a functional regression approach which adjusts for genetic factors. We analysed data from the Genomics of Drug Sensitivity in Cancer project relating to drug effectiveness for suspending cancer cell proliferation under different dosages, and examined five *BRAF* targeted inhibitors, each applied in a number of common and rare types of cancer cell lines. Our implementation of a two-stage screening algorithm revealed a number of genes that are potentially associated with drug response. Gene, drug and cancer type trajectories have been modeled using a varying coefficient modeling framework. The proposed methodology allows for dose-dependent analysis of genetic associations with drug response data. It enables us to study the effect of different drugs simultaneously, which results in high accuracy of drug response prediction. Drug comparisons using the proposed methodology could support drug repositioning, especially in diseases where existing treatment options are limited. In addition, our methodology can help to reveal unknown potential relationships between genetic characteristics and drug efficacy. Hence, the good predictive performance of our method could be due to the fact that some genes may act as proxies for unmeasured phenotypes that are directly relevant to drug sensitivity.

Our work relies on two major assumptions. First, that out of tens of thousands genes regulating protein composition only a small proportion is actually associated with cancer cell survival in a dosage-dependent manner. In other words, transcriptomic profiles exert influence on disease progress after drug administration in a sparse and dynamic way. However, if a large number of genes is associated with the drug response, our method may produce biased results, and some important information about the biological mechanisms can be lost. Secondly, we assume that the different drugs are comparable on the scale of maximum dosage percentage level for our joint model. We acknowledge that different drugs have different chemical structure and maximum screening concentrations. Our focus is to identify genetic components that could be informative for dose response given drugs that belong to a particular family, for example *BRAF* targeted therapies. However, our methodology is flexible enough to allow each drug to be examined separately if it appears to be clinically appropriate.

Drug response prediction from gene expression data has been widely studied in the literature. Sparse regression methods, gene selection algorithms such as the Ping-pong algorithm [50], or a combination of network analysis and penalized regression, e.g. the sparse network-regularized partial least squares method [51], have all been employed to simultaneously predict drug response and select genetic factors that seem to be associated with the drug response.

However, none of these methods are able to quantify the effect of drug dosage on the response, which is one of the main contributions of this work. Employing the proposed dose-varying model gives a detailed picture of different drug effects and can be extremely valuable in predicting drug response for agents with small therapeutic range and high toxicity levels. Our algorithm showed moderate predictive performance due to the complexity of predicting whole drug-response curves. Methods for further enhancing the performance of the proposed methodology, such as judicious use of prior information and leveraging information sharing across multiple data sources should be explored in the future in order to overcome this issue and make good use of its full potentials.

To conclude, the main purpose of this paper is to examine the dose-dependent associations between genes and drugs. The proposed methodology, by using the raw data to infer the effects of interest, allows to obtain a more comprehensive picture of the biological mechanisms that undergo cancer treatment and the role of drug dose on that. In addition, due to its simple structure, it allows extension to different types of molecular data (e.g. RNA-seq gene expression, methylation or mutational profiles) and enrichment with further information, such as drug chemical composition.

Supporting information

S1 Text. Accurate detection of drug associated genes from simulated responses. Simulated responses have been generated to examine the accuracy of the employed method in detecting the genes that are truly associated to drug response. Three screening thresholds, three active gene sets and two covariance structure scenarios for the repeated measurements simulation have been considered. This text includes all the details of the simulation study that we conducted.
(PDF)

S1 Fig. Distribution of tissue of origin across the five BRAF compounds used for cell line screening in the Genomics of Drug Sensitivity in Cancer data. Overall, similar proportion of cell lines have been treated with all of the compounds examined with smaller number of cell lines been treated with AZ628, Dabrafenib and PLX-4720. Larger number of cell lines in the data set were originated from the lungs, the gastrointestinal tract and the haematopoietic and lymphoid tissues.
(TIF)

S2 Fig. Estimated mean drug response trajectories for *BRAF* and *HRAS* mutated and non-mutated cancer cell lines: analysis performed on GDSC2 data. Observed responses (points) and estimated mean trajectory (lines) of cell concentration for cancer cell lines with and without *BRAF* and *HRAS* mutations after treatment with the eight anticancer compounds examined using data from GDSC2.
(TIF)

S3 Fig. Prediction accuracy for each different drug and scenario. Pearson correlation was estimated across observed and predicted AUC values. AUC values have been computed by calculating the area under the coefficient function curve (both observed and predicted). Training and test sets have been considered based on either the experimental units or on cancer cell lines only.
(TIF)

S4 Fig. Prediction accuracy for each different drug and scenario: analysis performed on GDSC2 data. Pearson correlation across observed and predicted AUC values. AUC values

have been computed by calculating the area under the coefficient function curve (both observed and predicted) using the GDSC2 data. Training and test sets have been considered based on either the experimental units or on cancer cell lines only.
(TIF)

S1 Table. Full gene rankings based on the estimated area under the coefficient function curve (analysis on the full data set). Gene rankings of all selected genes based on the magnitude of the genetic effect on drug response. A positive (+) sign translates to a positive effect on cells survival after drug administration, a negative (-) sign translates to a negative effect on cells survival and a mixed (0) effect translates to a varying effect on cells survival which depends on drug dosage. Spearman's correlation is calculated between drug dosage and gene estimated coefficient function values as an indicator of the magnitude change of the gene effect over the increasing dosage. Area corresponds to the area under the estimated coefficient curve and the SD corresponds to the standard deviation of the area based on bootstrapping. Mean fold change is calculated between the selected gene expression values of the cell lines carrying BRAF mutations with respect to wild type. Protein-protein interaction network distance is computed based on the shortest interaction path between the BRAF gene and each of the selected genes. Here, NI denotes absence of any interaction.
(XLSX)

S2 Table. Signalling pathways linked to genes predictive of BRAF inhibitor response (analysis on the full data set). Signalling pathways along with the adjusted p-values and the number of overlapping genes obtained after pathway enrichment analysis applied to the full scale analysis results.
(XLSX)

S3 Table. Overlaps between the observed gene set and oncogenic signatures in the Molecular Signatures database (analysis on the full data set). Overlaps have been detected using gene set enrichment analysis performed using a hypergeometric distribution. The false discovery rate analog of the hypergeometric p-value is displayed after correction for multiple hypothesis testing according to Benjamini and Hochberg.
(XLSX)

S4 Table. Signalling pathways linked to genes predictive of BRAF inhibitor response (analysis on the full data set)-Reactome. Signalling pathways along with the adjusted p-values and the number of overlapping genes obtained after pathway enrichment analysis applied to the full scale analysis results using the Reactome database.
(XLSX)

S5 Table. Overlaps between the observed gene set and oncogenic signatures in the Molecular Signatures database (resistant cell lines analysis). Overlaps have been detected using gene set enrichment analysis performed using a hypergeometric distribution. The false discovery rate analog of the hypergeometric p-value is displayed after correction for multiple hypothesis testing according to Benjamini and Hochberg.
(XLSX)

S6 Table. Signalling pathways linked to genes predictive of BRAF inhibitor response (analysis on resistant cell lines). Signalling pathways along with the adjusted p-values and the number of overlapping genes obtained after pathway enrichment analysis applied to the resistant cell line analysis results.
(XLSX)

S7 Table. Full gene rankings based on the estimated area under the coefficient function curve (analysis on resistant cell lines). Gene rankings of all selected genes based on the magnitude of the genetic effect on drug response. A positive (+) sign translates to a positive effect on cells survival after drug administration, a negative (-) sign translates to a negative effect on cells survival and a mixed (0) effect translates to a varying effect on cells survival which depends on drug dosage. Spearman's correlation is calculated between drug dosage and gene estimated coefficient function values as an indicator of the magnitude change of the gene effect over the increasing dosage. Area corresponds to the area under the estimated coefficient curve and the SD corresponds to the standard deviation of the area based on bootstrapping. Mean fold change is calculated between the selected gene expression values of the cell lines carrying *BRAF* mutations with respect to wild type. Protein-protein interaction network distance is computed based on the shortest interaction path between the BRAF gene and each of the selected genes. Here, NI denotes absence of any interaction.
(XLSX)

Acknowledgments

We acknowledge Emily Chambers and the Sheffield Bioinformatics Core Facility for their assistance in GDSC2 data pre-processing. Their input to this study is greatly appreciated.

Author Contributions

Conceptualization: Evanthia Koukouli, Frank Dondelinger, Juhyun Park.

Formal analysis: Evanthia Koukouli.

Investigation: Evanthia Koukouli, Dennis Wang, Frank Dondelinger, Juhyun Park.

Methodology: Evanthia Koukouli, Frank Dondelinger, Juhyun Park.

Project administration: Frank Dondelinger, Juhyun Park.

Software: Evanthia Koukouli.

Supervision: Frank Dondelinger, Juhyun Park.

Validation: Dennis Wang.

Visualization: Evanthia Koukouli, Dennis Wang.

Writing – original draft: Evanthia Koukouli.

Writing – review & editing: Evanthia Koukouli, Dennis Wang, Frank Dondelinger, Juhyun Park.

References

1. Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, Elledge R, et al. Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. The Lancet. 2003; 362(9381):362–369. https://doi.org/10.1016/S0140-6736(03)14023-8 PMID: 12907009
2. Cook D, Brown D, Alexander R, March R, Morgan P, Satterthwaite G, et al. Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. Nature Reviews Drug Discovery. 2014; 13(6):419–431. https://doi.org/10.1038/nrd4309 PMID: 24833294
3. Corrie PG. Cytotoxic chemotherapy: clinical aspects. Medicine. 2008; 36(1):24–28. https://doi.org/10.1016/j.mpmed.2007.10.012
4. Relling MV, Dervieux T. Pharmacogenetics and cancer therapy. Nature Reviews Cancer. 2001; 1 (2):99. https://doi.org/10.1038/35101056 PMID: 11905809

5. Zhang W, Liu Y, Sun N, Wang D, Boyd-Kirkup J, Dou X, et al. Integrating genomic, epigenomic, and transcriptomic features reveals modular signatures underlying poor prognosis in ovarian cancer. Cell Reports. 2013; 4(3):542–553. https://doi.org/10.1016/j.celrep.2013.07.010 PMID: 23933257

6. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A landscape of pharmacogenomic interactions in cancer. Cell. 2016; 166(3):740–754. https://doi.org/10.1016/j.cell.2016.06.017 PMID: 27397505

7. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Research. 2012; 41(D1):D955–D961. https://doi.org/10.1093/nar/gks1111 PMID: 23180760

8. Hyman DM, Taylor BS, Baselga J. Implementing genome-driven oncology. Cell. 2017; 168(4):584–599. https://doi.org/10.1016/j.cell.2016.12.015 PMID: 28187282

9. Ben-David U, Siranosian B, Ha G, Tang H, Oren Y, Hinohara K, et al. Genetic and transcriptional evolution alters cancer cell line drug response. Nature. 2018; 560:325–330. https://doi.org/10.1038/s41586-018-0409-3 PMID: 30089904

10. Tavassoly I, Hu Y, Zhao S, Mariottini C, Boran A, Chen Y, et al. Genomic signatures defining responsiveness to allopurinol and combination therapy for lung cancer identified by systems therapeutics analyses. Molecular Oncology. 2019; 13(8):1725–1743. https://doi.org/10.1002/1878-0261.12521 PMID: 31116490

11. Ji RR, de Silva H, Jin Y, Bruccoleri RE, Cao J, He A, et al. Transcriptional profiling of the dose response: a more powerful approach for characterizing drug activities. PLoS Computational Biology. 2009; 5(9). https://doi.org/10.1371/journal.pcbi.1000512 PMID: 19763178

12. Delpuech O, Rooney C, Mooney L, Baker D, Shaw R, Dymond M, et al. Identification of pharmacodynamic transcript biomarkers in response to FGFR inhibition by AZD4547. Molecular Cancer Therapeutics. 2016; 15(11):2802–2813. https://doi.org/10.1158/1535-7163.MCT-16-0297 PMID: 27550940

13. Falcetta F, Lupi M, Colombo V, Ubezio P. Dynamic rendering of the heterogeneous cell response to anticancer treatments. PLoS Computational Biology. 2013; 9(10):e1003293. https://doi.org/10.1371/journal.pcbi.1003293 PMID: 24146610

14. Silverbush D, Grosskurth S, Wang D, Powell F, Gottgens B, Dry J, et al. Cell-specific computational modeling of the PIM pathway in acute myeloid leukemia. Cancer Research. 2017; 77(4):827–838. https://doi.org/10.1158/0008-5472.CAN-16-1578 PMID: 27965317

15. Keshava N, Toh TS, Yuan H, Yang B, Menden MP, Wang D. Defining subpopulations of differential drug response to reveal novel target populations. NPJ Systems Biology and Applications. 2019; 5:36. https://doi.org/10.1038/s41540-019-0113-4 PMID: 31602313

16. Hastie T, Tibshirani R. Varying-coefficient models. Journal of the Royal Statistical Society: Series B (Methodological). 1993; 55(4):757–779.

17. Wu CO, Chiang CT, Hoover DR. Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. Journal of the American Statistical Association. 1998; 93 (444):1388–1402. https://doi.org/10.1080/01621459.1998.10473800

18. Wu CO, Chiang CT. Kernel smoothing on varying coefficient models with longitudinal dependent variable. Statistica Sinica. 2000; p. 433–456.

19. Huang JZ, Wu CO, Zhou L. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. Statistica Sinica. 2004; p. 763–788.

20. Qu A, Li R. Quadratic inference functions for varying-coefficient models with longitudinal data. Biometrics. 2006; 62(2):379–391. https://doi.org/10.1111/j.1541-0420.2005.00490.x PMID: 16918902

21. Fan J, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high-dimensional additive models. Journal of the American Statistical Association. 2011; 106(494):544–557. https://doi.org/10.1198/jasa.2011.tm09779 PMID: 22279246

22. Geenens G, et al. Curse of dimensionality and related issues in nonparametric functional regression. Statistics Surveys. 2011; 5:30–43. https://doi.org/10.1214/09-SS049

23. Song R, Yi F, Zou H. On varying-coefficient independence screening for high-dimensional varying-coefficient models. Statistica Sinica. 2014; 24(4):1735. PMID: 25484548

24. Chu W, Li R, Reimherr M. Feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data. The Annals of Applied Statistics. 2016; 10(2):596. https://doi.org/10.1214/16-AOAS912 PMID: 27630755

25. Fan J, Ma Y, Dai W. Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. Journal of the American Statistical Association. 2014; 109(507):1270–1284. https://doi.org/10.1080/01621459.2013.879828 PMID: 25309009

26. Wang D, Hensman J, Kukaite G, Toh TS, Dry JR, Saez-Rodriguez J, et al. A statistical framework for assessing pharmacological response and biomarkers with confidence. BioRxiv. 2020.

27. Tansey W, Li K, Zhang H, Linderman SW, Rabadan R, Blei DM, et al. Dose-response modeling in high-throughput cancer drug screenings: An end-to-end approach. arXiv preprint arXiv:181205691. 2018.

28. Xue L, Qu A, Zhou J. Consistent model selection for marginal generalized additive model for correlated data. Journal of the American Statistical Association. 2010; 105(492):1518–1530. https://doi.org/10.1198/jasa.2010.tm10128

29. Xue L, Qu A. Variable selection in high-dimensional varying-coefficient models with global optimality. Journal of Machine Learning Research. 2012; 13(Jun):1973–1998.

30. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nature Methods. 2016; 13(12):966. https://doi.org/10.1038/nmeth.4077 PMID: 27898060

31. Yang X, Hu F, Liu JA, Yu S, Cheung MPL, Liu X, et al. Nuclear DLC1 exerts oncogenic function through association with FOXK1 for cooperative activation of MMP9 expression in melanoma. Oncogene. 2020; 39(20):4061–4076. https://doi.org/10.1038/s41388-020-1274-8 PMID: 32214200

32. Subbiah V, Kreitman RJ, Wainberg ZA, Cho JY, Schellens JH, Soria JC, et al. Dabrafenib and trametinib treatment in patients with locally advanced or metastatic BRAF V600–mutant anaplastic thyroid cancer. Journal of Clinical Oncology. 2018; 36(1):7. https://doi.org/10.1200/JCO.2017.73.6785 PMID: 29072975

33. Sharma SP. RAS mutations and the development of secondary tumours in patients given BRAF inhibitors. The Lancet Oncology. 2012; 13(3):e91. https://doi.org/10.1016/S1470-2045(12)70046-3

34. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013; 14(1):128. https://doi.org/10.1186/1471-2105-14-128 PMID: 23586463

35. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Research. 2016; 44(W1): W90–W97. https://doi.org/10.1093/nar/gkw377 PMID: 27141961

36. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Research. 2017; 46 (D1):D661–D667. https://doi.org/10.1093/nar/gkx1064

37. Rangaswami H, Bulbule A, Kundu GC. Osteopontin: role in cell signaling and cancer progression. Trends in Cell Biology. 2006; 16(2):79–87. https://doi.org/10.1016/j.tcb.2005.12.005 PMID: 16406521

38. Sharma N, Jha S. NLR-regulated pathways in cancer: opportunities and obstacles for therapeutic interventions. Cellular and Molecular Life Sciences. 2016; 73(9):1741–1764. https://doi.org/10.1007/s00018-015-2123-8 PMID: 26708292

39. Whyte J, Bergin O, Bianchi A, McNally S, Martin F. Key signalling nodes in mammary gland development and cancer. Mitogen-activated protein kinase signalling in experimental models of breast cancer progression and in mammary gland development. Breast Cancer Research. 2009; 11(5):209. https://doi.org/10.1186/bcr2361 PMID: 19818165

40. Mortezaee K, Salehi E, Mirtavoos-mahyari H, Motevaseli E, Najafi M, Farhood B, et al. Mechanisms of apoptosis modulation by curcumin: Implications for cancer therapy. Journal of Cellular Physiology. 2019; 234(8):12537–12550. https://doi.org/10.1002/jcp.28122 PMID: 30623450

41. Colomer C, Margalef P, Villanueva A, Vert A, Pecharroman I, Solé L, et al. IKKα kinase regulates the DNA damage response and drives chemo-resistance in cancer. Molecular Cell. 2019; 75(4):669–682. https://doi.org/10.1016/j.molcel.2019.05.036 PMID: 31302002

42. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005; 102(43):15545–15550. https://doi.org/10.1073/pnas.0506580102 PMID: 16199517

43. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1995; 57 (1):289–300.

44. Wilke CO. Bringing molecules back into molecular evolution. PLoS Computational Biology. 2012; 8(6): e1002572. https://doi.org/10.1371/journal.pcbi.1002572 PMID: 22761562

45. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Research. 2020; 48(D1):D498–D503. https://doi.org/10.1093/nar/gkz1031 PMID: 31691815

46. O'Donovan KJ, Ma K, Guo H, Wang C, Sun F, Han SB, et al. B-RAF kinase drives developmental axon growth and promotes axon regeneration in the injured mature CNS. Journal of Experimental Medicine. 2014; 211(5):801–814. https://doi.org/10.1084/jem.20131780 PMID: 24733831

47. Comunanza V, Corà D, Orso F, Consonni FM, Middonti E, Di Nicolantonio F, et al. VEGF blockade enhances the antitumor effect of BRAFV 600E inhibition. EMBO Molecular Medicine. 2017; 9(2):219–237. https://doi.org/10.15252/emmm.201505774 PMID: 27974353

48. Solit DB, Rosen N. Resistance to BRAF inhibition in melanomas. New England Journal of Medicine. 2011; 364(8):772–774. https://doi.org/10.1056/NEJMcibr1013704 PMID: 21345109

49. Manzano JL, Layos L, Bugés C, de los Llanos Gil M, Vila L, Martinez-Balibrea E, et al. Resistant mechanisms to BRAF inhibitors in melanoma. Annals of Translational Medicine. 2016; 4(12). https://doi.org/10.21037/atm.2016.06.07 PMID: 27429963

50. Kutalik Z, Beckmann JS, Bergmann S. A modular approach for integrative analysis of large-scale gene-expression and drug-response data. Nature Biotechnology. 2008; 26(5):531. https://doi.org/10.1038/nbt1397 PMID: 18464786

51. Chen J, Zhang S. Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. Bioinformatics. 2016; 32(11):1724–1732. https://doi.org/10.1093/bioinformatics/btw059 PMID: 26833341