bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

Biomarker detection for revealing anticancer drug dynamics

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 (EK)

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 penalised regression) is used to identify genetic factors that are associated with drug response over

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 tumourgenesis 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 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 regularisation and, in contrast with existing literature, selects the most informative genes by accounting for responses 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.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 [2]. Therefore, there is a need for more effective and personalised therapeutic strategies that can improve cancer treatment decisions, and hence patient outcomes.

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 best therapeutic agent 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]. Previous work has examined the difference in transcriptional response [5] and drug response at the cell population level after administering anticancer drugs in various dosages [6, 7].

Cancer cell line drug screens provide valuable information about biomarkers that are predictive of drug response. During the last decade, there have been several systematic studies aiming to examine pharmacogenomic relationships [8–11]. These studies were conducted on human cancer cells that have been isolated from affected tissues, grown in vitro and treated with anti-cancer inhibitors. By examining the genomic profiles of these cell lines, investigators were able to identify relationships between cancer-driven genetic alterations and drug response. However, these relationships have only been modelled on the aggregate response, and hence little is known about the relationship between drug dosage and genetic factors. Recently, Tansey et al. [12] proposed a method for modelling drug-response curves via Gaussian processes and linking them to biomarkers using a neural network prediction model. The authors did not use their model for dosage-dependent inference of biomarker effects, and the highly non-linear neural network model makes interpretation of biomarker effects challenging.

Gene expression profiles can provide valuable functional information on the genetic mechanisms which determine anti-cancer drug response, offering more tailored treatments where common therapies become ineffective. However, statistical analysis for linking transcriptomic profiles with drug response becomes challenging due to the high-dimensional nature of the data. Over the last 20 years, researchers developed

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

statistical methodologies not only to mitigate the problem of high dimensionality [13–18], but also to detect markers of positive drug response to cancer treatment [19, 20] and predict patient response after drug administration [1, 5, 16, 21–25]. While these previous methods have gone some way towards solving the challenges associated with drug response modelling, none of them address all of the issues that arise in personalised medicine, namely: selecting genes associated with drug response, identifying the optimal dosage, characterising gene-dose relationships and predicting response for one or multiple drugs.

With regards to the high-dimensional nature of the dataset, 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. However, 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 [26]. Methods to estimate the covariate (e.g. gene) effect include global and local smoothing e.g. kernel estimators [27, 28], basis approximation [29] or penalised splines [30]. The most straightforward and computationally efficient method is through basis approximation where each coefficient function is approximated through some basis functions and the varying coefficient model can be written as a linear regression model. Then, estimation for repeated measurements data (e.g. drug response over different dosages) can be incorporated through minimising a weighted least squares criterion based on a specified weighting scheme (repeated measurements covariance structure) [29]. However, inference becomes impossible as the number of predictors increases and when 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. When combining sparse regression with the varying coefficient model framework, predictors are handled jointly under the assumption that the majority are irrelevant to the outcome variable. Penalties from group versions of the least absolute shrinkage and selection operator (LASSO), smoothly clipped absolute deviation (SCAD), bridge etc. have been used for fitting the varying coefficient model [31]. Because these methods handle all of the predictors jointly, their implementation becomes extremely challenging and impractical when the number of predictors (e.g. thousands) is much larger than the

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

number of samples (e.g. hundreds). Consequently, attempts to develop prior univariate tests focused on filtering out the unimportant predictors by simply estimating the association of each predictor to the outcome variable separately [32–34]. Often, these screening methods are conservative, and still return many more predictors than those which are truly associated to the response. To overcome this issue, regularisation or alternative variable selection methods have been used after screening to further fine-tune the set of predictors [32, 34].

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 [32]. Here, we extended their methodology to a completely different objective of assessing the transcriptomic effect on anti-cancer drug response, where our coefficient functions were allowed to vary with dosage. Note that unlike in Tansey et al. [12], biomarker effects will be a linear function function of dosage, allowing for straight-forward interpretation of the coefficient functions.

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. We applied this method to data extracted from the Genomics for Drug Sensitivity in Cancer (GDSC) project [9]. To compare and differentiate similar treatments, we examined the effect 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 gene effect on the drug response. Consequently, in contrast to past studies, we managed to model 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Materials and Methods

The Genomic 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). 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. The drug sensitivity measurement was obtained via fluorescence-based cell viability assays 72 hours after drug administration [9]. 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 S2 Fig.). Paired microarray gene expression data (17737 genes) was available together with the drug response dataset (`https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Home.html`).

The dose-response dataset also included a blank response for wells on the experimental plate that have not been seeded with cells or treated with a drug. Blank responses have been used to adjust for the magnitude of the error while measuring the amount of cells in each well. 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 subject 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 subject as described above and NR_{ij} is the new score taken from the transformation,

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$i = 1, \ldots, 3805, j = 1, \ldots, n_i.$

A two-stage algorithm for identification of gene-drug 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 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, z_{iG})^T$ 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; i.e. we make a 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. Due to computational convenience, for 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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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

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

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

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

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 $\mathbf{x}$, the high-dimensional covariate vector in Eq (2).

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 $\boldsymbol{x}$, we fitted the marginal non-parametric regression model for the gth x-predictor:

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

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, $\varepsilon^{(g)}$ 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 x-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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The variance function $V(d)$ in Eq (11) was estimated using techniques similar to [32].

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

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

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

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

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

where

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

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 $\boldsymbol{\eta}_g$ that correspond to the same coefficient function. In addition, in order to reduce the bias introduced when applied a LASSO penalty, we alternatively chose the SCAD, which coincides with the LASSO until $u = \lambda$, then transits to a quadratic function until

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$u = \alpha\lambda$ and then it remains constant $\forall u > \alpha\lambda$, meaning that retains the penalisation and bias rates of the LASSO for small coefficients but at the same time relaxes the rate of penalisation as the absolute value of the coefficients increases. In Fig 1 the reader can find a brief overview of the employed methodology.

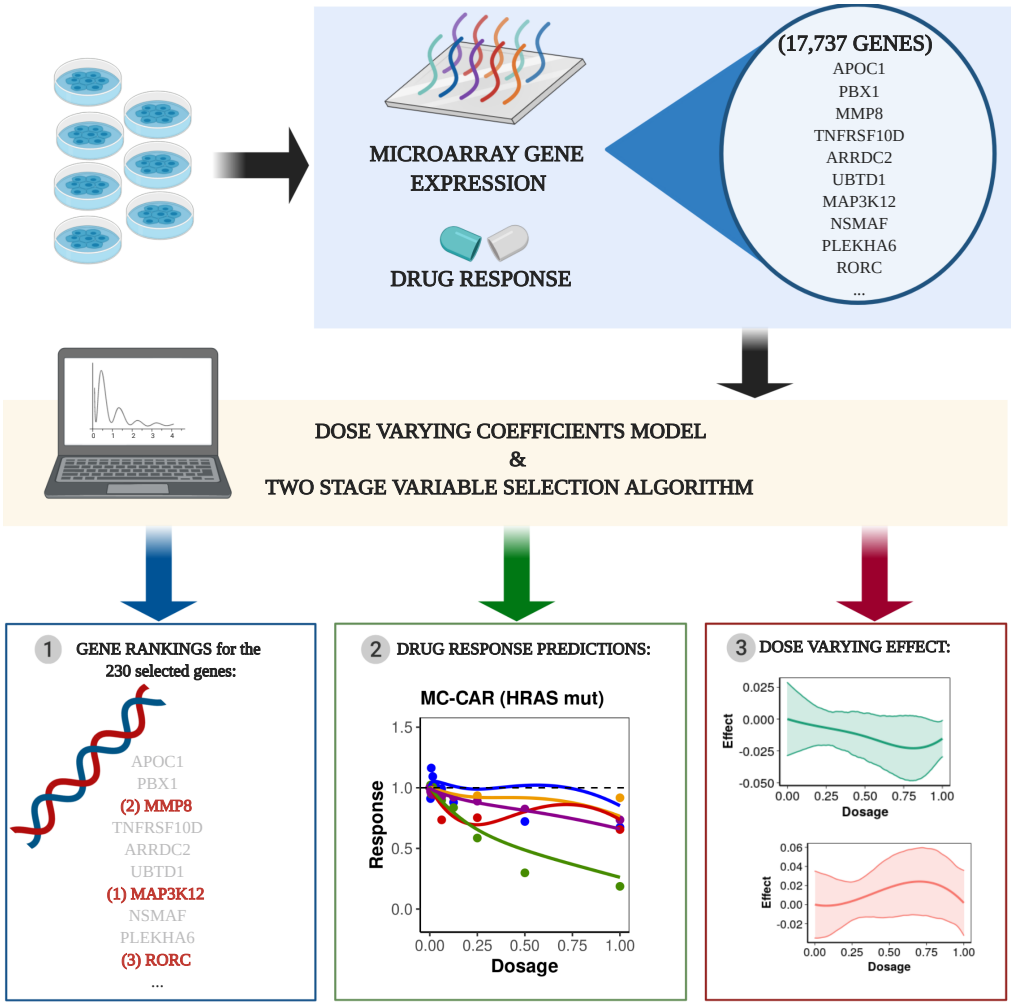

Fig 1. The two-stage algorithm for identifying dose-dependent associations between genes and drugs. Gene expression and drug response data from a drug screening study (e.g. GDSC) are used to fit our dose-varying coefficients model to estimate the dose-varying effect between covariates and drug response. 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 drugs of interest.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Tuning parameters selection

We used knots placed at the median of the observed data values along with cubic B-splines with 1 interior knot, resulting from calculating the number of interior knots suitable using the formula $N_n = [n^{\frac{1}{2p+3}}]$ proposed and applied by [29,35,36]. 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 $\nu[\frac{n}{log(n)}]$, $\nu \in \{1, 2, 3, \ldots\}$. 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, [37]). 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. 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 specified smooth coefficient functions (see S1 Text). Due to the computational burden associated with a simulation of the same scale as the data set, we conducted a simulation study using smaller random fragments of the original GDSC data set. In particular, we repeatedly sampled without replacement 190 experimental units and 886 genes based 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 ($\tau_n(\nu) = [\frac{n}{log(n)}]$, $\tau_n(\nu) = [\frac{2n}{log(n)}]$ and $\tau_n(\nu)$ chosen using the greedy-INIS algorithm [37]) 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (Fig. A in S1 Text). A screening threshold of size $[\frac{2n}{log(n)}]$ and regression weights adjusted for the covariance structure of the data have been 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 $[\frac{2n}{log(n)}]$, and weights derived by assuming a rational quadratic covariance structure for the repeated measures.

Data and 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`. The data is available online at the Genomics of Drug Sensitivity in Cancer website `https://www.cancerrxgene.org/`.

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;

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 [38].

The 230 genes were ranked based on the estimated AUC value (S4 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 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's 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 a decreasing (Spearman's correlation=-0.928) and negative effect on cancer cell survival 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 [39]. 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's 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 addition effect compared to this intercept. Similar patterns can be

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 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.

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. Interestingly, 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

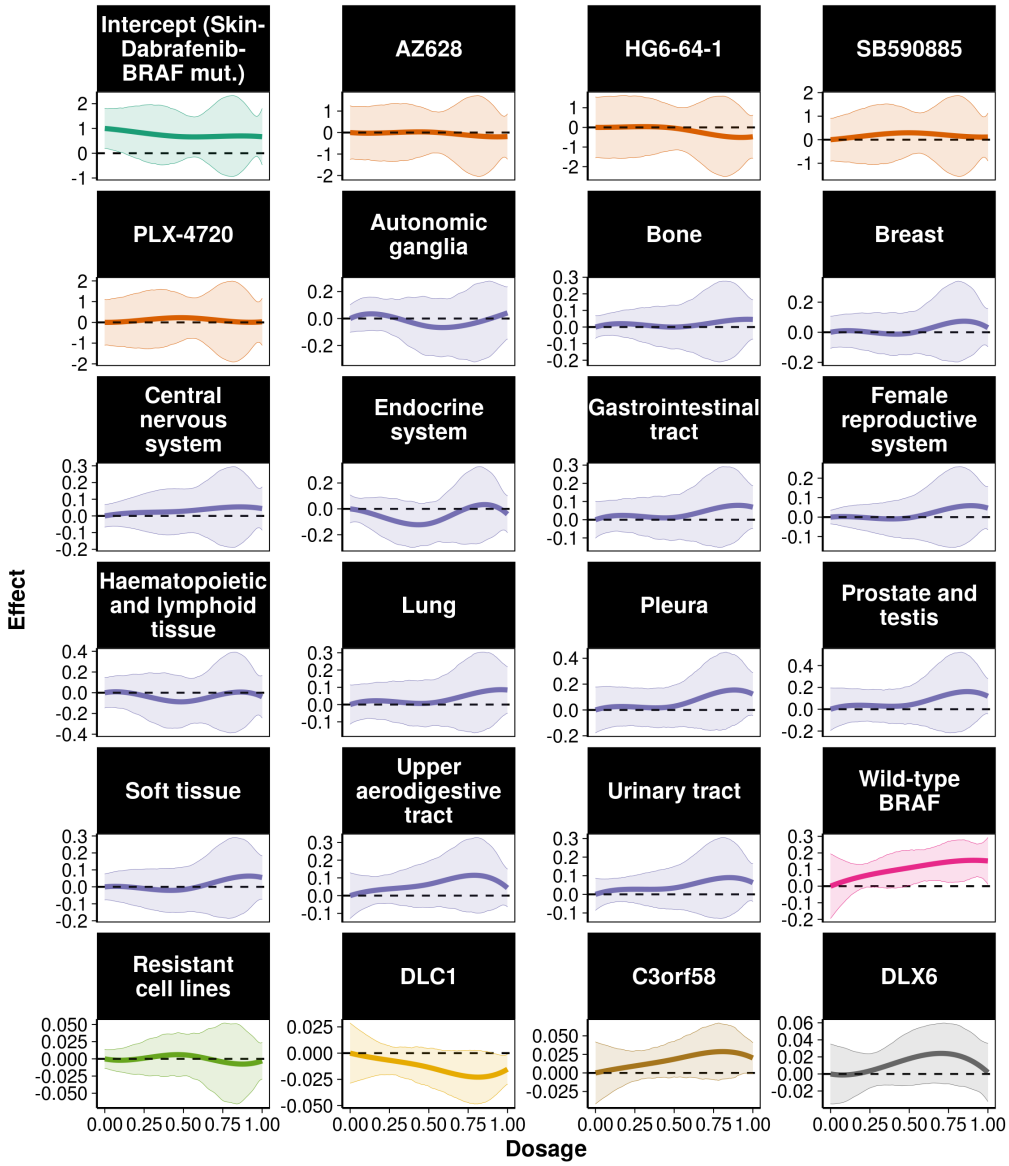

Fig 2. Estimated coefficient functions for the low-dimensional predictors and three of the selected genes. Estimated coefficient functions for intercept, different drugs, tissue of origin and three of the selected genes along with 95% bootstrap confidence intervals. Baseline corresponds to *BRAF* mutant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

lower dosages. The observed effect in endocrine system cell lines reflects the Dabrafenib responses observed in anaplastic thyroid cancer patients [40]. 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 3). Together these results provide important insights into the effectiveness of the five

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

BRAF targeted drugs examined on different cancer types, highlighting the potential for effective treatment of a wide range of cancers given cancers' genetic characteristics.

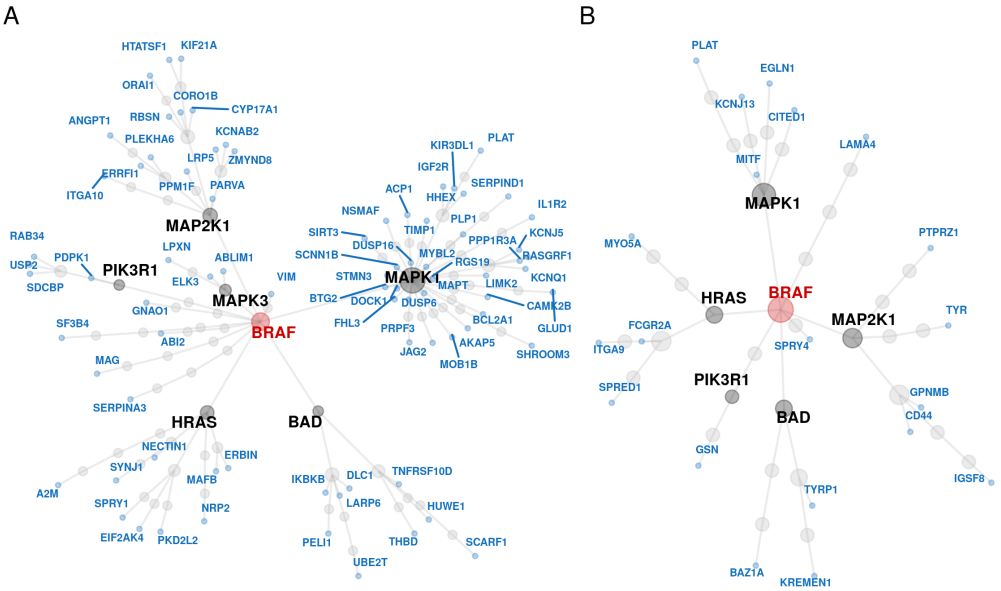

Fig 3. Protein-protein interaction network for the genes selected from the two-stage variable selection algorithm. (A) Undirected protein-protein interaction network between the 230 selected (blue) and the *BRAF* (red) genes (full scale analysis). (B) Undirected protein-protein interaction network between 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.

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 inhibitors' target. 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 3).

Since *HRAS* mutations are frequent in patients receiving *BRAF* targeted therapies [41], we examined the mean estimated trajectory over different dosages under

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

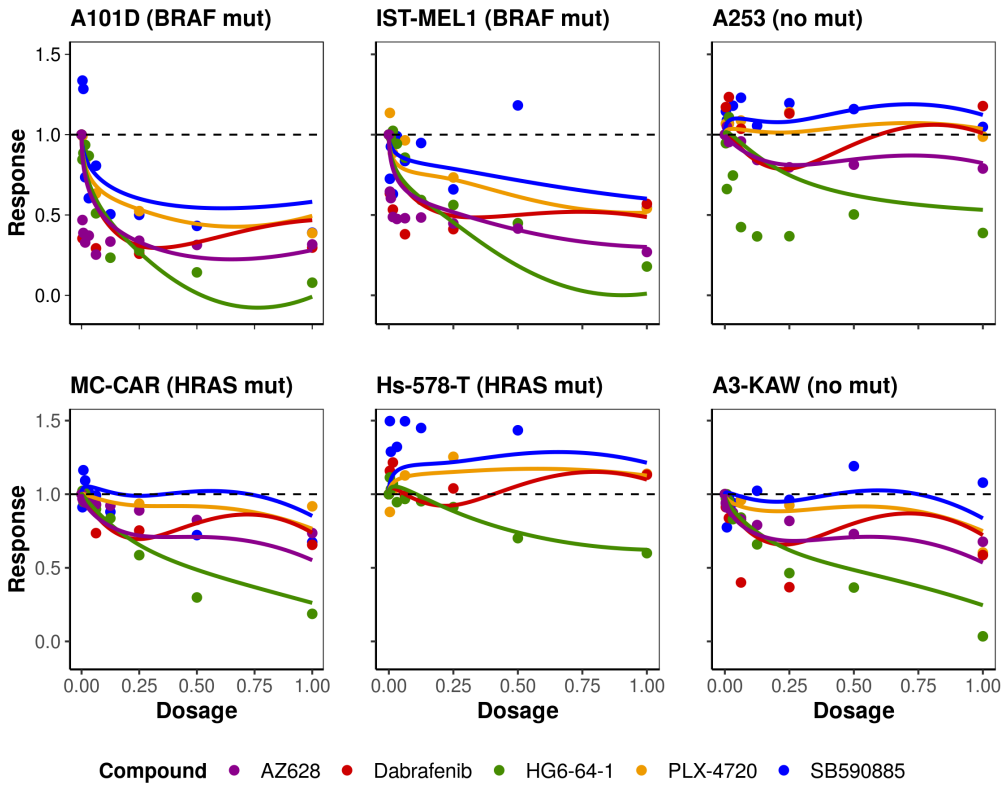

Fig 4. Estimated mean drug response trajectories for four cancer cell lines with *BRAF* and *HRAS* mutations. Observed responses (points) and estimated mean trajectory (lines) of cells' concentration for cancer cell lines with and without *BRAF* and *HRAS* mutations after treatment with the five anticancer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

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 [42, 43] and WikiPathways [44] databases to see if the selected genes can be grouped into common functional classes or pathways. In total, 183 pathways identified, from which 11 were statistically significant at 5% level, including apoptosis modulation, NOTCH1 regulation, and MAPK signaling (S6 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 tumourgenesis and cancer treatment [45–48]. 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 [49]. 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.

In addition to the pathway enrichment analysis, we used the Molecular Signatures Database (MSigDB database v7.0 updated August 2019: [50]) to compute overlaps between the observed gene set and known oncogenic gene sets. Fig 5 displays 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.

Identifying dose-dependent genes in drug-resistance conditions

Acquired resistance to BRAF inhibitors is often observed in the clinic [52]. 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 resistant mechanisms to BRAF inhibitors [53]. 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 associated with dose-response, though none of

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

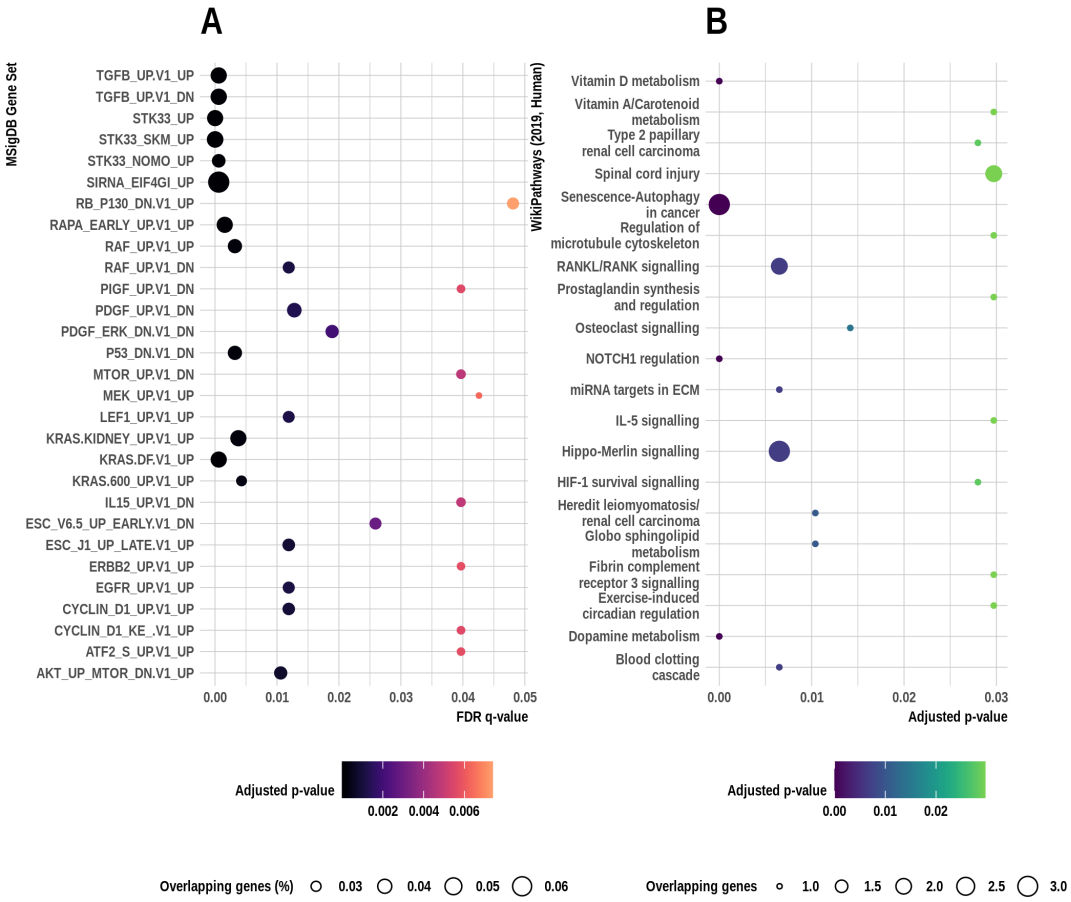

Fig 5. Overlaps between the observed gene set and oncogenic signatures in the Molecular Signatures Database (full data analysis); signalling pathways enriched for genes predictive of *BRAF* inhibitor response (resistant cell lines). (A) Full gene set names can be found in S8 Table. Overlaps have been 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 [51]. (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 full list of the pathways identified see S7 Table).

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 3) 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; S9 Table). Finally, we found 34 pathways enriched for genes predicting drug response of the

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

mutated cell lines to the examined BRAF inhibitors, of which the top 20 are depicted in Fig 5(B).

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 S5 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 shows robust cross-validated performance when it comes to predicting sensitivity to the administered drugs (see S3 Fig. which shows the correlation between predicted and true response). Predictive accuracy was evaluated under four different sub-scenarios: 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 really 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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
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 cells survival after drug administration, a negative (-) sign translates to a negative effect on cells survival and a neutral (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. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 of 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−).

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 modelled using a varying coefficient modelling 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 disease indications 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 cells 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 are associated with the drug response,

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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. However, 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 [25], or a combination of network analysis and penalised regression, e.g. the sparse network-regularized partial least squares method [17], 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. Employing the proposed dose-varying model gives a detailed picture of different drugs effect and can be extremely valuable in predicting drug response for agents with small therapeutic range and high toxicity levels. In addition, the method can be easily extended for different cell lines-drug combinations as well as different types of molecular data (e.g. RNA-seq gene expression, methylation or mutational profiles). Finally, due to the structure of our model, enrichment with additional low-dimensional covariates, such as drug chemical information, is straightforward.

Funding

This work was financially supported by the North West Social Science Doctoral Training Partnership (ref: ES/P000665/1) as part of Evanthia's Koukouli doctoral studies (ref: 2035874). Dennis Wang is supported by the NIHR Sheffield Biomedical Research Centre, Rosetrees Trust (ref: A2501), and the Academy of Medical Sciences Springboard (ref: SBF004/1052).

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

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.

3. Corrie PG. Cytotoxic chemotherapy: clinical aspects. Medicine. 2008;36(1):24–28.

4. Relling MV, Dervieux T. Pharmacogenetics and cancer therapy. Nature reviews cancer. 2001;1(2):99.

5. 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).

6. 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.

7. 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.

8. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603–607.

9. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

10. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. science. 2006;313(5795):1929–1935.

11. Hyman DM, Taylor BS, Baselga J. Implementing genome-driven oncology. Cell. 2017;168(4):584–599.

12. 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:1812.05691. 2018 Dec 13.

13. Fan J, Ren Y. Statistical analysis of DNA microarray data in cancer research. Clinical Cancer Research. 2006;12(15):4469–4473.

14. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 2012;483(7391):570.

15. Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Advances in bioinformatics. 2015 Jun 11;2015.

16. Zhang N, Wang H, Fang Y, Wang J, Zheng X, Liu XS. Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. PLoS computational biology. 2015;11(9):e1004498.

17. Chen J, Zhang S. Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. Bioinformatics. 2016;32(11):1724–1732.

18. Toh TS, Dondelinger F, Wang D. Looking beyond the hype: Applied AI and machine learning in translational medicine. EBioMedicine. 2019 Aug 26.

19. 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.

20. Wang D, Hensman J, Kutkaite G, Toh TS, GDSC Screening Team, Dry JR, et al. A statistical framework for assessing pharmacological response and biomarkers with confidence; BioRxiv. 2020 Jan 1.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

21. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. PLoS one. 2013;8(4):e61318.

22. Ruffalo M, Thomas R, Chen J, Lee AV, Oesterreich S, Bar-Joseph Z. Network-guided prediction of aromatase inhibitor response in breast cancer. PLoS computational biology. 2019;15(2):e1006730.

23. Qian C, Sidiropoulos ND, Amiridi M, Emad A. From Gene Expression to Drug Response: A Collaborative Filtering Approach. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2019. p. 7465–7469.

24. Geeleher P, Cox NJ, Huang RS. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. Genome biology. 2014;15(3):R47.

25. 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.

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

27. 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.

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

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

30. Qu A, Li R. Quadratic inference functions for varying-coefficient models with longitudinal data. Biometrics. 2006;62(2):379–391.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

32. 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.

33. 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.

34. 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.

35. 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.

36. 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.

37. 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.

38. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nature methods. 2016;13(12):966.

39. 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.

40. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

41. Sharma SP. RAS mutations and the development of secondary tumours in patients given BRAF inhibitors. The Lancet Oncology. 2012;13(3):e91.

42. 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.

43. 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.

44. 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. doi:10.1093/nar/gkx1064.

45. Rangaswami H, Bulbule A, Kundu GC. Osteopontin: role in cell signaling and cancer progression. Trends in cell biology. 2006;16(2):79–87.

46. 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.

47. 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.

48. 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.

49. 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.

50. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102(43):15545–15550.

51. 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 (Methodological). 1995;57(1):289–300.

52. Solit DB, Rosen N. Resistance to BRAF inhibition in melanomas. New England Journal of Medicine. 2011;364(8):772–774.

53. 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).

Supporting information captions

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.

S2 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.

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.

S4 Table. Full gene rankings based on the estimated area under the

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

S5 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.

S6 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 to the full scale analysis results.

S7 Table. Signalling pathways linked to genes predictive of BRAF inhibitor

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.18.158907; this version posted June 18, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 to the resistant cell line analysis results. **S8 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. **S9 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.