A statistical framework for assessing pharmacological response and biomarkers with confidence

Dennis Wang[1,2], James Hensman[3], Ginte Kutkaite[4,5], Tzen S. Toh[6], GDSC Screening Team[#,7], Jonathan R Dry[8], Julio Saez-Rodriguez[9], Mathew J. Garnett[7], Michael P. Menden[4,5,10,*], Frank Dondelinger[11,*]

1. Sheffield Institute for Translational Neuroscience, University of Sheffield, Sheffield S10 2HQ, UK
2. Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
3. PROWLER.io, Cambridge, CB2 1LA, UK
4. Institute of Computational Biology, Helmholtz Zentrum München—German Research Center for Environmental Health, 85764, Neuherberg (Germany)
5. Department of Biology, Ludwig-Maximilians University Munich, 82152, Martinsried (Germany)
6. The Medical School, University of Sheffield, Sheffield, S10 2RX, UK
7. Wellcome Sanger Institute. Cambridge UK. CB10 1SA
8. Research and Early Development, Oncology R&D, AstraZeneca, Boston, MA, USA
9. Heidelberg University, Faculty of Medicine, Bioquant, 69120 Heidelberg
10. German Centre for Diabetes Research (DZD e.V.), 85764, Neuherberg (Germany)
11. Centre for Health Informatics, Computation and Statistics, Lancaster Medical School, Lancaster University, Lancaster LA1 4YW, UK

*corresponding authors: Michael P. Menden (michael.menden@helmholtz-muenchen.de) and Frank Dondelinger (f.dondelinger@lancaster.ac.uk)

GDSC Screening Team: Howard Lightfoot, Wanjuan Yang, Maryam Soleimani, Syd Barthorpe, Tatiana Mironenko, Alexandra Beck, Laura Richardson, Ermira Lleshi, James Hall, Charlotte Tolley, William Barendt

Abstract

Drug high-throughput screenings across large molecular-characterised cancer cell line panels enable the discovery of biomarkers, and thereby, cancer precision medicine. The ability to experimentally generate drug response data has accelerated. However, this data is typically quantified by a summary statistic from a best-fit dose response curve, whilst neglecting the uncertainty of the curve fit and the potential variability in the raw readouts. Here, we model the experimental variance using Gaussian Processes, and subsequently, leverage this uncertainty for identifying associated biomarkers with a new statistical framework based on Bayesian testing. Applied to the Genomics of Drug Sensitivity in Cancer, *in vitro* screening data on 265 compounds across 1,074 cell lines, our uncertainty models identified 24 clinically established drug response biomarkers, and in addition provided evidence for 6 novel biomarkers. We validated our uncertainty estimates with an additional drug screen of 26 drugs, 10 cell lines with 8 to 9 replicates. Our method is applicable to drug high-throughput screens without replicates, and enables robust biomarker discovery for new cancer therapies.

Introduction

The failure rate for new drugs entering clinical trials is in excess of 90%, with more than a quarter of drugs failing due to lack of efficacy [1,2]. The rapid development of technologies for deep molecular characterisation of clinical samples holds the promise to uncover molecular biomarkers that stratify patients towards more efficacious drugs, a cornerstone of precision medicine. In oncology, we can identify potential biomarkers of drug response in high-throughput screens (HTS) of patient-derived cell lines; these biomarkers need then to be validated in patients.

Assessment of cell line drug response typically involves treatment with multiple concentrations of the compound, followed by measurement of the amount of viable cells after a fixed period of time for each dose, and derivation of a dose-response curve. The drug response is commonly then summarised by measurements taken from this curve, most often the concentration required to reduce cell viability by half *i.e.* IC_{50}, or the area under the curve *i.e.* AUC. Currently the two largest *in vitro* drug screening studies, the Genomics of Drug Sensitivity in Cancer (GDSC) [3,4] and the Cancer Therapeutics Response Portal (CTRP) [5] have shown that some clinically actionable biomarkers of drug response can be concordantly discovered [4,6], and that different properties and mechanisms of drug response are best captured by different metrics dependent on the dose-response curve [7].

Most HTS efforts focus on increasing throughput [4,6] and thereby often neglect experimental replicates, which renders it impossible to correct for experimental noise, resulting in uncertainty for the estimated drug response metrics (e.g. IC_{50} value). Extrapolating IC_{50} values beyond the tested drug concentration range is particularly challenging and often unaccounted for in quality control metrics [8,9]. Most published studies using machine learning algorithms or mechanistic models for predicting drug response and biomarkers assume that the measured drug responses are precise [10–12]. If this assumption is not met and there is

high uncertainty in the measured drug response values, the utility of these methods for enhancing drug development may be severely limited [10–12]. Experimental noise can be reduced by adding experimental replicates, however, this either reduces the throughput of the screen or increases the cost. Most current models for curve fitting and describing dose-response data have primarily assumed that cell viability has a sigmoidal relationship to the logarithm of the dose concentrations of the drug [13,14]. While some models are more flexible by allowing many inflection points in the dose-response curve [15,16], their main output is a single drug response value that does not fully capture the uncertainty in the measurements [7].

Gaussian processes (GP) are a flexible, probabilistic modeling technique that has been successfully used to measure uncertainty in noisy gene expression datasets [17] and has been incorporated into machine learning prediction of cell fates [18]. This technique has been shown to cope well with regression tasks on dependent data and high dimensional covariates [19,20]. Instead of fitting a single function to the data, GPs allow for a flexible range of beliefs about the function underlying the data [21]. In the case of cell line drug responses, this can be conceptualised as fitting a range of curves that have equivalently strong fit to the data. We can sample from the inferred posterior distribution over functions, i.e. the variance between these curves, to generate uncertainty estimates of quantities of interest, in our case, properties of the dose-response such as IC_{50}.

In this study, we introduce a new GP regression approach for describing dose-response relationships in cancer cell lines that quantifies the uncertainty of the model fitted to measured responses. Instead of experimentally determining the confidence interval for drug response using many replicate experiments, our method computationally estimates these in each single experiment. Subsequently, we use our new dose-response models to identify genetic sensitivity and resistance biomarkers in standard statistical tests (e.g. ANOVA). We demonstrate how the flexibility of the GP dose-response modeling can be further exploited in a Bayesian framework to identify novel biomarkers. We also describe the variation in the level of drug response uncertainty across cancer types and drug classes. By accounting for the uncertainty in dose-response experiments, detection of clinically-actionable biomarkers can be enhanced.

Results

1. A probabilistic framework for measuring dose-response and predicting biomarkers

We analysed *in vitro* screening data on 265 compounds across 1,074 cell lines [4]. In those experiments, we quantified the amount of cytotoxicity after four days of compound treatments at each dose compared to controls (**Figure 1A**). The relationship between the dose and response (decrease in cell viability) was first described using a dose-response curve derived with a sigmoidal function (**Figures 1B and 1C**). This assumes that the number of viable cells decreases at an exponential rate, then slows down and eventually plateaus at

a lower limit. Since it was costly to test all possible doses, the sigmoid function was used to extrapolate the response at concentrations that had not been tested and to estimate overall measures of response, such as IC_{50} or AUC values, for downstream analysis. However, considering that each experiment tested only between five and nine dosage concentrations per experiment in GDSC, and a maximum of 16 in CTRP, the tightness of fit of the dose-response curve to the data points and therefore the level of uncertainty about the inferred response may vary. We utilised the probabilistic nature of GP models to quantify the uncertainty in the dose-response experiments as an alternate approach (**Figure 1D**). We sampled from the fitted GP and used the posterior distribution to quantify the uncertainty in curve fits for each experiment**.** We again generated summary statistics, IC_{50} and AUC values, by taking the average of the GP samples and also quantified the level of uncertainty for these statistics (**Figure 1E**). The GP model has the advantage that it models outliers at higher doses as one component of a two-component Beta mixture in the model (**see Methods**). Such outliers are typically the result of an experimental failure, and cannot be modeled using simple Gaussian noise without over-estimating the noise parameter.

After fitting the dose-response data using the sigmoid and GP models, we tested various biomarker hypotheses by examining the association between the overall response statistics from the models with genetic variants detected in the cell lines using a frequentist and a Bayesian approach (**Figures 1F-H**). For one biomarker hypothesis, as an example, we examined copy number alterations and point mutations in breast cancer cell lines in relation to the measured drug response of afatinib in those cells. The GP and sigmoid estimated IC_{50} from cell lines treated with afatinib were significantly different in cases with and without *ERBB2* amplification (ANOVA q-value = 4.12e-9; **Figure 1I**). The GP models provided an added benefit of providing uncertainty estimates that were incorporated into a Bayesian hierarchical model to further verify the association between *ERBB2* amplification and afatinib sensitivity (posterior probability = 0.001; **Figure1J**).

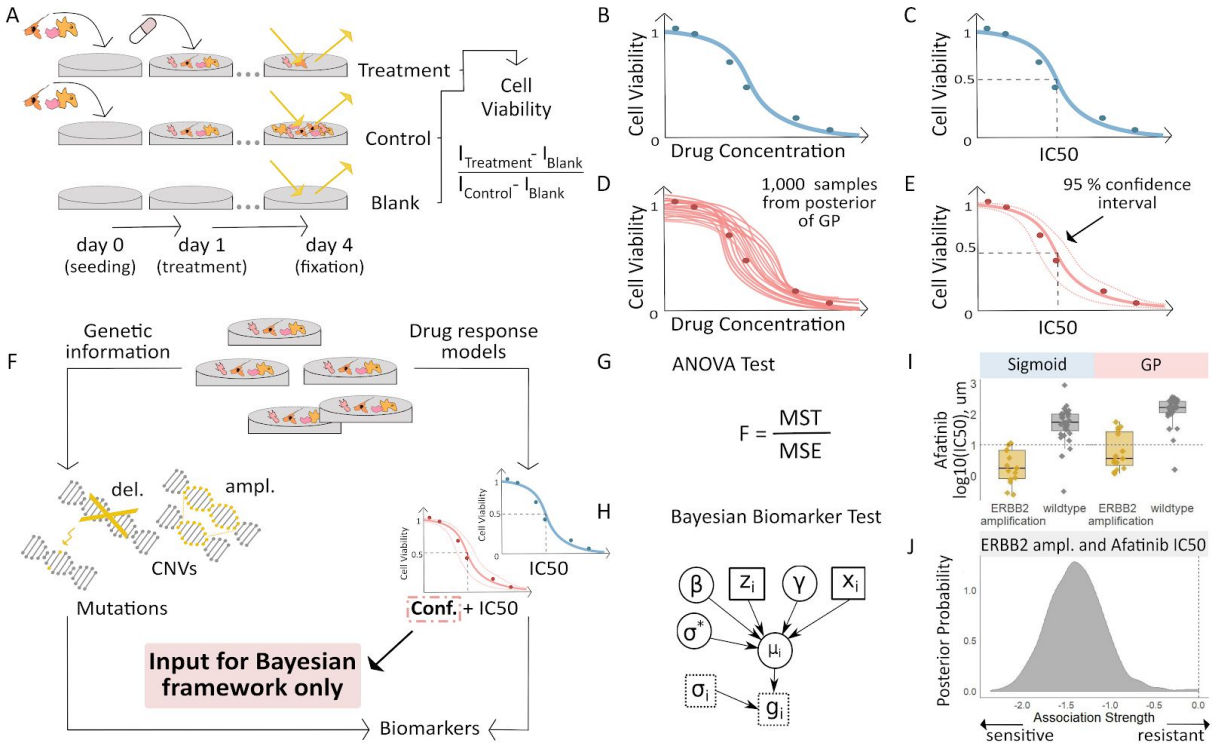

Figure 1: Workflow for fitting of Gaussian Process models to dose-response curves and estimating their uncertainty. (A) Large-scale drug screens test cell lines with different drugs and at different doses are used to obtain dose-response data. **(B)** Typically, for each drug tested in a cell line, the sigmoid model is fit to the drug-response data and **(C)** the overall measures of response (IC_{50}, AUC, etc.) are extracted. **(D)** For each drug tested in a cell line, we fit a GP model to the dose-response data. The GP allows us to sample from a distribution of possible dose-response curves, obtaining a measure of uncertainty. **(E)** From these curves, we can extract overall measures of response, such as IC_{50}, and importantly, their 95% confidence intervals. **(F)** Mutation markers for each cell line can be determined based on presence/absence of single nucleotide polymorphisms (SNPs) in key genes. Both the drug response estimates and the mutation markers are used to compute **(G)** the F-statistic for ANOVA, and **(H)** Bayesian test for biomarker association. The drug response summary measure g_i for cell-line i is modelled via a cell-line specific mean μ_i and standard error σ_i. The mean is defined as a linear effect β of the biomarker status z_i and a further effect γ from any remaining covariates x_i, such as tissue type. The parameter σ^* is the standard deviation of μ_i. **(I)** Boxplots illustrate the differences in the estimated mean IC_{50} of *ERBB2* amplified and non-amplified breast cancer cell lines treated with afatinib. An ANOVA test was used to test this difference in means but did not consider uncertainty in each IC_{50} estimate. **(J)** We estimated posterior distributions of gene association using the Bayesian model, *i.e.* the effect of a genetic mutation on the IC_{50} measurement of drug response. Distributions centered on zero indicate no effect while distributions on either side of zero indicate positive or negative effects of mutations on drug response.

2. Gaussian Processes provide estimates of dose-response uncertainty for single experiments

Both GP and sigmoid curve fitting produced comparable IC_{50} and AUC estimates. Precursor sigmoid curve fitting methods based on Markov Chain Monte Carlo simulations enabled error estimates in IC_{50} values [3], however, this was neglected in the state-of-the-art sigmoid curve fitting [16] due to missing propagation to biomarker identification. Here, we introduce the added benefit of sampling from the GP posterior, which provides the models in-build uncertainty obtained for these IC_{50} estimates. This is important for high-throughput drug screening experiments where there is often a high number of drugs and samples tested but very few replicate experiments. By applying the GP model to each experiment, we estimated the standard deviation for each IC_{50} or AUC value based only on data points from that single experiment. These single sample standard deviations were compared to the standard deviations measured from here provided replicate experiments, *i.e.* the same drug tested multiple times on the same cell line and at the same concentration. We applied our GP estimation method to data from replicate experiments of 26 drugs on 10 cell lines, which contained 260 test conditions and 8 to 9 replicates for each condition. We expected that an estimate of the uncertainty of the summary statistic, such as the standard deviation of the IC_{50} posterior samples, would be correlated with the dispersion between replicates. Here, we refer to the variability between (mean) estimates for replicates as the observation uncertainty, and the variability in the estimate for a single replicate as the estimation uncertainty.

We compared observation and estimation uncertainty across replicate experiments of all 260 conditions (**Figure 2A**). We found two trends in the relationship between observation and estimation uncertainty. First, for experiments where the estimated IC_{50} lies within the concentration range tested, the estimation uncertainty is positively correlated (Pearson correlation = 0.84, CI [0.76, 0.89]) with the observation uncertainty. Second, for experiments where the estimated IC_{50} lies beyond the maximum tested concentration, we observed a negative correlation (Pearson correlation = -0.39, CI [-0.51, -0.25]). We note that the latter experiments require extrapolation to estimate the IC_{50} beyond the concentration range, which increases the estimation uncertainty, but does not generally affect the observational uncertainty. However, we observed that the estimation uncertainty from our GPs for dabrafenib (BRAF inhibitor) tested in two independent studies on the same cell lines were comparable both within and beyond the concentration range (**Figure 2B**).

Since the replicate experiments were conducted in batches over a period of several months, we verified that the observed trends held regardless of batches (**Supplementary Figure S1**). Additionally, we examined the relationship between estimation uncertainty and observation uncertainty in a number of example experiments (**Figures 2C-E**). In the case of olaparib tested on PC-14, the uncertainty for the IC_{50} within each replicate experiment was high, but this level of uncertainty was consistent across all replicates even beyond the max concentration (**Figures 2C and 2F**). In other replicate experiments, both estimation and observation uncertainty were low (**Figures 2D and 2G**), or varied depending on whether the batch reported mostly IC_{50} values beyond the concentration range. Talazoparib tested in colorectal cancer line HCT-15 is a case where observation uncertainty was high, even though estimation uncertainty was low, and experiments in different batches showed different estimated IC_{50}s from very different dose-response curves (**Figures 2E and 2H**).

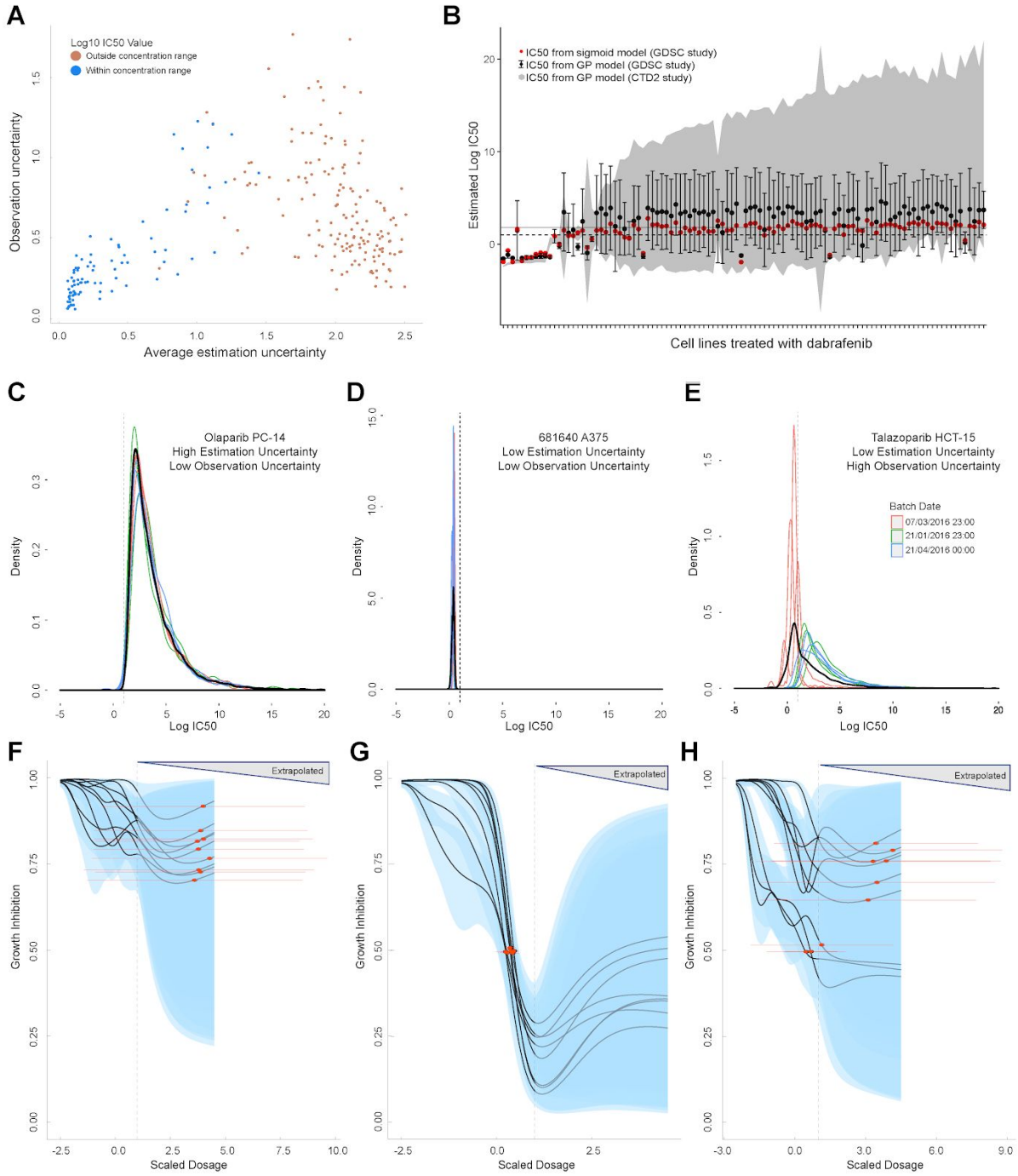

Figure 2: Comparison of GP estimates of uncertainty to replicate drug screening experiments. (A) Comparison between observational uncertainty (standard deviation over replicates of log10(IC_{50}) mean estimates) and estimation uncertainty (average over replicates of log10(IC_{50}) standard deviation) from each replication experiment. The colour of the points indicates whether the log10(IC_{50}) mean estimates were within or outside the maximum concentration range for each assay. **(B)** Mean IC_{50} and the estimation uncertainty from the GPs for a BRAF inhibitor (dabrafenib) tested in each cell line in two independent studies (GDSC and CTD2). Estimation uncertainty (error bars and grey shading) were larger beyond the max concentration in both GDSC (dashed line) and CTD2 (grey line). The point estimates of the IC_{50}s from the GPs (black dots) were also comparable to the published IC_{50}s (red dots). **(C-E)** Three sets of replicate experiments, representing different amounts of estimation and observation uncertainty. Each density represents the distribution of IC_{50}

values from the Gaussian process samples from each replicate experiment. The colours represent different experimental batches. Narrow distributions demonstrate low estimation uncertainty and overlapping distributions demonstrate low observation uncertainty. The thick black line represents the density obtained by pooling samples from all replicates and the dashed line shows the maximal dosage tested. GP curve fits corresponding to the three sets of replicate experiments showing IC_{50} estimates with **(F)** high uncertainty, **(G)** low uncertainty, and **(H)** mix of uncertainties depending on whether estimates are made within or beyond the max concentration. The blue areas represent the 95% confidence interval in the curve fits and extrapolated GP curves (light grey lines) are displayed up to five times the maximum concentration.

In order to examine the diversity of uncertainty estimates across experiments further, we described the relationship between AUC value of GP fits with their corresponding estimation uncertainty (**Figure 3**). We decided to use AUC here due to the greater uncertainty of estimating IC_{50}s beyond the maximum dose concentration. Since AUCs were computed within the tested concentration range, the estimation uncertainty for AUC was not substantially higher for cases where IC_{50}s were estimated within compared to beyond the maximum concentration (**Supplementary Figure S2A**). The difference between the AUC estimates from the GP compared to the published GDSC sigmoid curve fits was greatest for experiments showing a partial response (AUC between 0.4 and 0.9), whilst at the same time these experiments also had the highest estimation uncertainty (**Figure 3A**). Our visual examination of the raw dose-response data from those experiments revealed evidence of poor quality readouts, for instance, where cell viability increases with increasing drug dose (**Supplementary Figure S2B**). We were able to quantify the quality of these readouts by estimating the Spearman correlation coefficient based on the raw cell viability counts and the dose concentrations (**Figure 3B**). A negative Spearman correlation indicates that cell viability decreases as dosage increases (as expected) while a positive Spearman correlation indicates the opposite. The experiments with high estimation uncertainty from our GPs were also the experiments with high Spearman correlation pointing to poor quality.

Next, we investigated whether there were any attributes of experiments that would correspond to high estimation uncertainty and poor quality results. Labelling of experiments based on cell culture conditions, dose and cancer type revealed no obvious associations with estimation uncertainty (**Supplementary Figure S3A-E**). However, there was a large spread in the uncertainty estimates for AUC when we grouped the experiments into target pathways based on the primary targets of the tested drugs (**Figure 3C; Supplementary Figure S3F**). While most drugs had similar average AUC point estimates between 0.6 and 0.8, suggesting they all had a spread of experiments showing resistance and sensitivity, the average estimation uncertainties varied across target pathways. Interestingly, similar target pathways (e.g. chromatin histone methylation and chromatin histone acetylation) had very different levels of estimation uncertainty. Within each of these target pathways, we also see different distributions of estimation uncertainties (**Figure 3D**). Most target pathways have a bi-modal distribution representing compounds that have low uncertainty in the cases of clear sensitivity or resistance, and high uncertainty in the cases of partial responders. Chromatin histone methylation drugs in particular had a much longer right tail towards higher estimation uncertainties that are associated with poor experimental readouts.

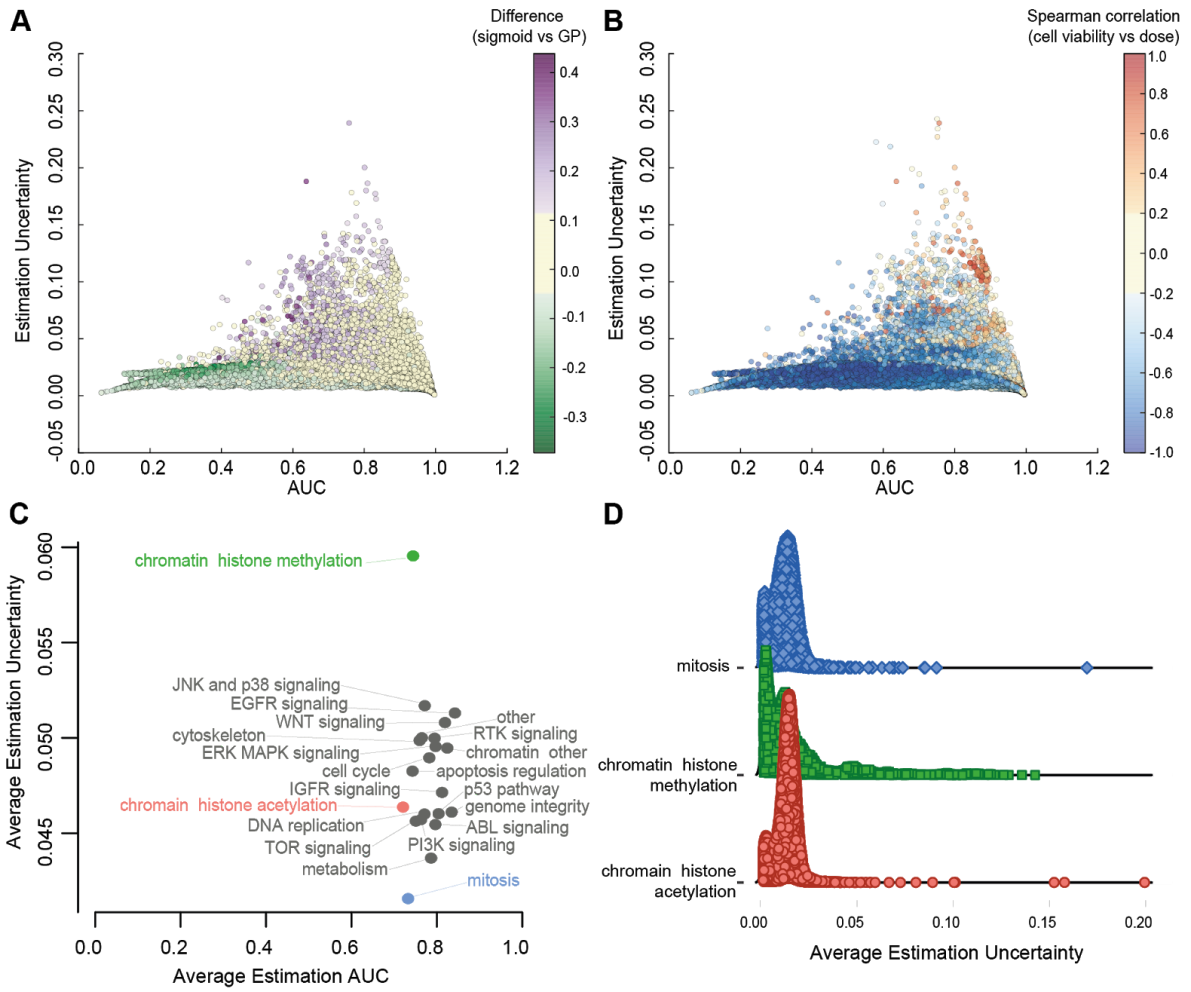

Figure 3: Relationship between AUC and uncertainties estimated from GPs across all experiments. (A) Coloured by difference between the AUC estimated by sigmoid vs GP fits. **(B)** Coloured by Spearman correlation between cell viability and dose concentration in the raw data. Poorer experiments (orange-red) result in greater uncertainty and positively correlated with cell viability increasing with higher dose. **(C)** Average uncertainty and AUC for experiments with uncertain fits (estimation uncertainty > 0.03) with drugs grouped by their target pathway. **(D)** Distribution of estimation uncertainty for drugs targeting chromatin histone methylation, chromatin histone acetylation, and mitosis.

3. Curve fits using Gaussian Processes can help identify clinically relevant biomarkers

The IC_{50} values are highly conconcordant for sigmoid and GP curve fittings, showing an average weighted Pearson correlation of 0.88 (CI [0.86; 0.90]) across individual drugs, and cancer types (**Figure 4A**). Strong agreement is found when true responding cell lines were observed in the screen (**Figure 4B**). For example, if >10% of cell lines responded within the concentration range, *i.e.* IC_{50} < maximum tested concentration, then a weighted Pearson correlation > 0.75 was consistently achieved for all drugs. We found positive correlations for all drugs, even when comparing exclusively non-responding cell lines, where all the IC_{50} values are extrapolated beyond the maximum dosage range. Drug response values are concordantly fitted with both methods for sensitive cell lines (**Figure 4C**, mean log10(IC_{50}) in

μM of 0.02 CI [-0.05; 0.09]), whilst extrapolated non-responders tend to lead to more conservative and higher IC_{50} values fitted with GP (**Figure 4C**, mean log10(IC_{50}) in μM of 1.10, CI [1.03; 1.18]). While the average fits from the sigmoid and GP models identify known clinical biomarkers, there are clearly differences for individual cell lines, especially when the IC_{50} value has been extrapolated beyond the dosage range, that may help identify new biomarkers.

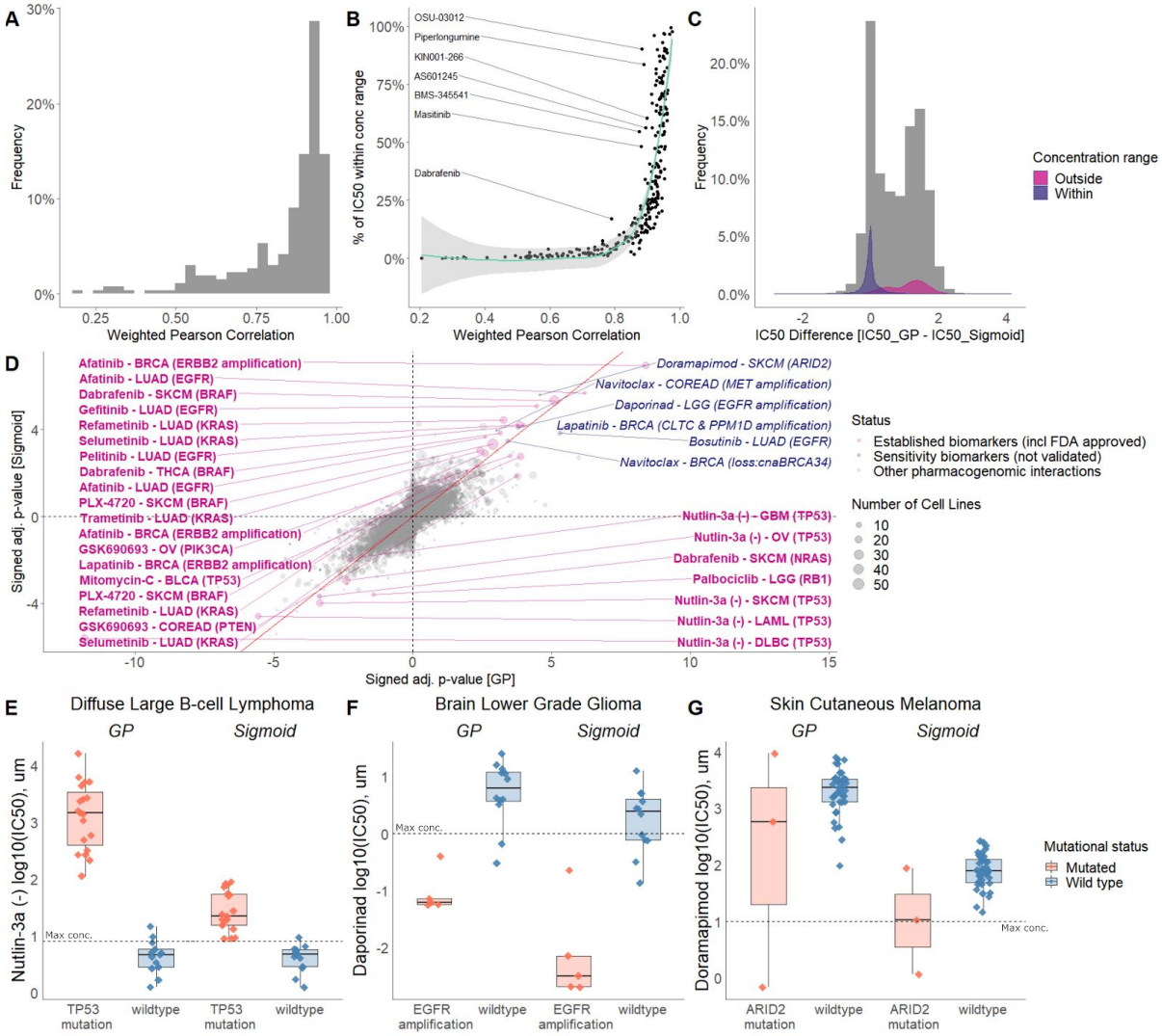

Figure 4: Comparison of sigmoid and GP curve fitting. (A) Weighted Pearson correlation of each drug within cancer types. **(B)** Comparing the concordance of sigmoid and GP curve fitting when stratifying for percentage of cell lines with IC_{50} value lower than maximum concentration. **(C)** IC_{50} value difference between GP and sigmoid curves. **(D)** Drug response biomarker comparison based on both curve fittings. Additional biomarker examples for **(E)** diffuse large B-cell lymphoma (DLBCL) treated with nutlin-3a (MDM2 inhibitor) and stratified by *TP53* mutants; **(F)** Low Grade Glioma (LGG) treated with daporinad (NAMPT inhibitor) and stratified by *EGFR* amplification; **(G)** Skin cutaneous melanoma (SKCM) treated with doramapimod (p38 & JNK2 inhibitor) and stratified with *ARID2* mutations.

To highlight the overall agreement of both curve fitting methods, we systematically tested 26 clinically established biomarkers of drug response (**Figure 4D, Supplementary Figure S4A-C, Supplementary Table S1**) using previously established association tests [4], 24 of which were significantly reproduced regardless of sigmoid or GP curve fitting (10% FDR).

For example, both curve fittings captured the association of BRAF inhibitors (PLX4720, progenitor of vemurafenib; and dabrafenib) with *BRAF* mutations in melanoma (**Supplementary Figure S5A-C**) [22]. Dabrafenib is a potent BRAF inhibitor and in addition we detected *BRAF* mutations as a sensitivity marker in thyroid carcinoma (**Figure 4D**, **Supplementary Figure S5D**). Another example are the EGFR inhibitors, afatinib and gefitinib, that are concordantly correlated with drug sensitivity in *EGFR* mutant cell lines in lung adenocarcinoma (**Supplementary Figure S5E-G**) [23,24]. *ERBB2*(*HER2*) amplification in breast cancer was also recapitulated as a biomarker of sensitivity to the dual EGFR/ERBB2 inhibitor lapatinib (**Supplementary Figure S5H**) [25]. Among the 26 clinical biomarkers, we consistently found drug resistance of *TP53* mutants to MDM2 inhibition with nutlin-3a in five different cancer types (**Figure 4E, Supplementary Figure S5I-L**). Overall, the majority of expected clinical and preclinical biomarkers are reproduced, regardless of the drug response curve fitting method.

We concordantly and significantly identified 6 novel (not yet clinically established) drug sensitivity biomarkers (0.1% FDR) regardless of the applied drug response curve fitting method. Investigating two different curve fitting algorithms, and retrieving the same biomarkers can be considered as a test of robustness, which in our case concordantly highlighted non-gold standard associations for prioritising experimental validation. For example, daporinad (also known as FK866 and APO866) is a small molecule inhibitor of nicotinamide phosphoribosyltransferase leading to inhibition of NAD+ biosynthesis. It has been clinically tested in melanoma (ClinicalTrials.gov Identifier: NCT00432107), Refractory B-CLL (NCT00435084) and Cutaneous T-cell Lymphoma (NCT00431912), whilst showing anti-proliferative effect in glioblastoma cell lines [26]. Therapeutic potential when combining with other drugs used to treat gliomas [27,28] has been suggested, while we additionally and concordantly identify *EGFR* amplification as a biomarker (**Figure 4F**).

Another novel and concordant identified biomarker is doramapimod response (also known as BIRB-796) in *ARID2* mutant melanoma cell lines (**Figure 4G**). Doramapimod is a small molecule p38 MAPK inhibitor and has been reported in different cancer types (in combination with other drugs) including cervical cancer, paracrine tumours and myeloma [29,30]. ARID2 is part of chromatin remodelling complex and is involved in DNA repair in hepatocellular carcinoma cells [31] and enriched in melanomas [32,33]. In conclusion, different curve fitting approaches lead to concordantly and novel identified biomarkers, thereby increasing the robustness in those findings, and consequently enabling to prioritise hypotheses.

4. Improved biomarker detection by taking into account uncertainty in a Bayesian framework

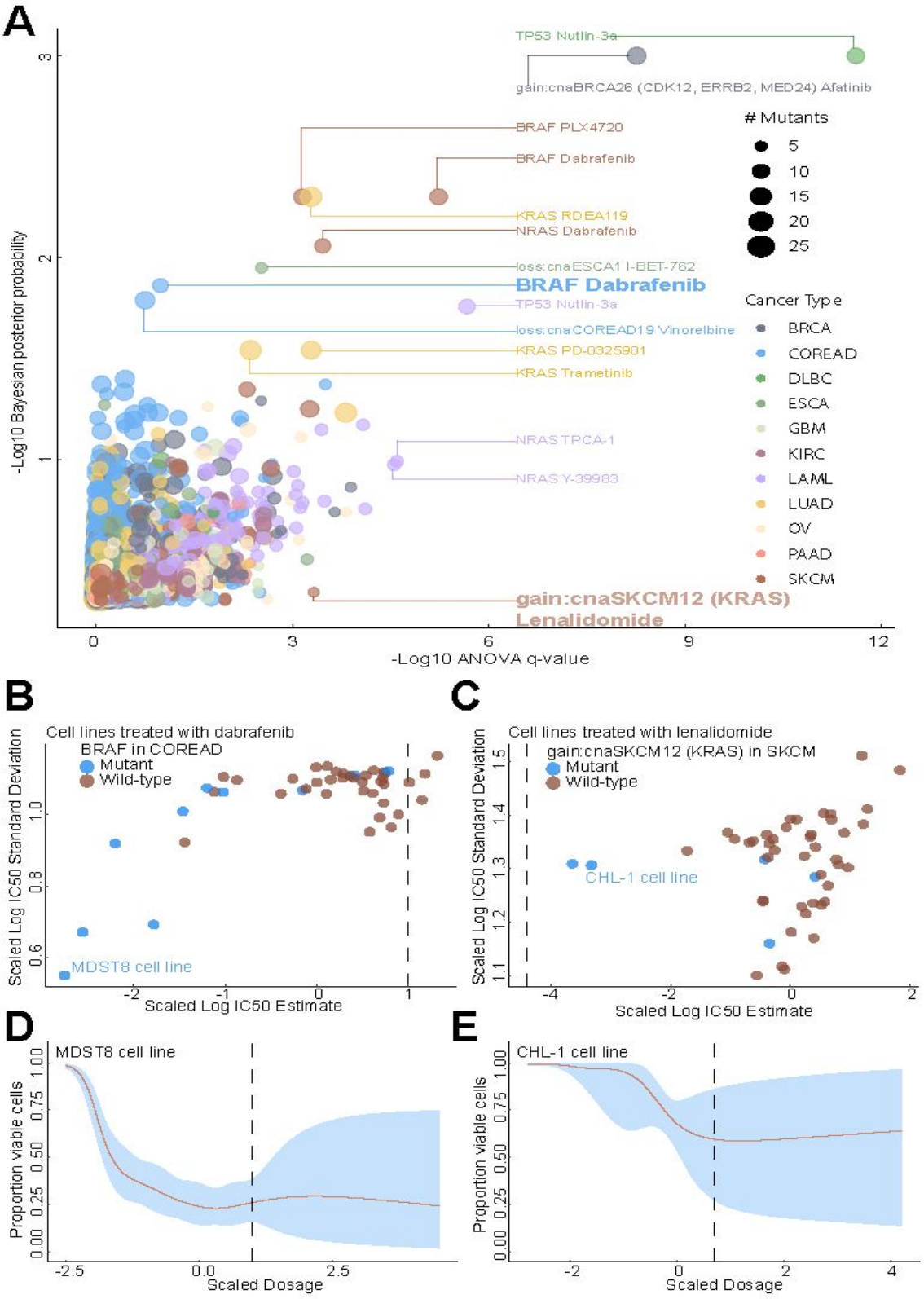

Figure 5: Comparison of Bayesian testing and ANOVA using the GP IC_{50} estimates. (A) Scatterplot of biomarker associations with IC_{50} drug response. The y-axis shows the negative log10 transformed posterior probability of a sign change in the effect under the Bayesian testing model, while the x-axis shows the negative log10 of the q-value from ANOVA testing. The size of the circles is proportional to the number of mutants or copy number variations in the given type of cancer cell line. **(B)** GP estimates for the mean and

standard deviation of the log(IC_{50}) from colorectal cell lines tested with BRAF inhibitor dabrafenib, which showed significant association with *BRAF* mutation in the Bayesian test. **(C)** Estimated IC_{50} and its uncertainty for skin cutaneous melanoma cell lines tested with the immunomodulatory drug lenalidomide, which showed significant association with *KRAS* copy number alteration in the ANOVA test. Black vertical lines show the location of the maximum experimental drug dosage. Dose-response curve of the **(D)** MDST8 colorectal cancer cell lines with *BRAF* mutation treated with dabrafenib. The black dotted line represents the maximum concentration of the drug used to treat the cell lines. The blue area represents the 95% confidence intervals in the dose-response fits. **(E)** Similar to **(D)** but for CHL-1 skin cutaneous melanoma cell lines with *KRAS* copy number alteration treated with lenalidomide.

We observed a number of cases where the Bayesian and ANOVA tests disagree (**Figure 5A; Supplementary Table S2**). For instance, *BRAF* mutations in colorectal cancer were detected as a sensitivity biomarker for dabrafenib by the Bayesian test, but less significant by the ANOVA test. This association had been repeatedly reported in *in vitro* models [4,5] and also found in melanoma cases [22], whilst not in colorectal cancer patients due to feedback activation of ERK-signalling mediated via *EGFR* [34,35]. We note in **Figure 5B** that the Bayesian test takes advantage of the additional information that sensitive mutant cell lines have low estimation uncertainty, while the small number of resistant mutant cell lines have high estimation uncertainty, causing them to have less influence on the biomarker detection. On the other hand, the ANOVA model detected the *KRAS* copy number alteration as a resistance biomarker for lenalidomide (immunomodulatory drug) partial sensitivity in skin cutaneous melanoma (SKCM), whilst not detected by our Bayesian approach. While on the linear IC50 scale there is some difference between the small number of mutant cell lines and wildtypes, the Bayesian model considered that the estimated responses of the mutant cell lines had high uncertainty (**Figure 5C**). While discrepancies between Bayesian and ANOVA tests have to be taken with caution, they may highlight novel biological insights which would be missed when applying only a single model.

Discussion

The GP approach developed in this manuscript has several advantages compared to the traditional approach of fitting sigmoidal drug response curves. Firstly, these flexible, non-parametric models can be used to fit a wider variety of dose-response curves than the parametric sigmoidal models, e.g. curves of unexpected shapes may reflect biological signals of off-target effects. Secondly, the GP models provide straightforward uncertainty quantification of any summary statistic that can be calculated on a dose-response curve, a fact that we take advantage of in developing our hierarchical Bayesian model for biomarker testing. Thirdly, the GP model can deal with outlying measurements better than a sigmoidal model, due to formulating it as a mixture model with one component representing the latent GP process of the drug response, and the second component accounting for outliers.

The increased flexibility of the GP model comes at a price. Most notably, because we do not impose a specific functional form, there are few constraints on the behaviour of the curve outside the range of observed dosages. This leads to the counter-intuitive behaviour that the posterior mean estimate of drug response can go up when extrapolating beyond the maximum dosage. Note, however, that this goes along with a commensurate increase in the posterior variance (**Figure 5D**,**E**). In other words, the model is highlighting that extrapolation

beyond the observed dosage range is highly uncertain, and the posterior mean estimate should not be relied on. It would be possible to constrain this behaviour by introducing artificial data points at a high concentration, or less crudely by imposing monotonicity constraints via virtual derivative observations [36]. However, these methods would limit the flexibility of our method and lead us to underestimate the uncertainty of the posterior mean, as well as increasing the computational burden, and so we have not pursued them in this work.

We have systematically compared the application of GP to sigmoid models across a pan-cancer drug screen. We demonstrated that our GP estimates of the IC_{50} values and their subsequently predicted biomarkers using ANOVA are reliable when compared to estimates from the sigmoid models. In addition, the GP models provide useful information about the uncertainty associated with the drug response quantification.

However, there is a crucial difference between estimation uncertainty on a single experiment and observational uncertainty across multiple replicates of the same experiment, which incorporates measurement error, technical and biological variation. We are interested in the former to assess the quality of the fit, and therefore the reliability of the estimated IC_{50}. We hypothesized that estimation uncertainty characterises observational uncertainty within the dose concentration range tested, but extrapolating beyond the concentration range would be challenging. We have verified this by applying our estimation method to a replication data set of 26 drugs tested on 10 different cell lines, with 8 to 9 replicates for each drug-cell line experiment. We conclude that while estimation uncertainty is a useful indicator for within-concentration IC_{50} values, it cannot be used as a proxy for observation uncertainty when the IC_{50} is extrapolated beyond the tested concentration range. Indeed, overall drug responses and biomarkers from independent drug screens were consistent when comparing similar dose ranges [8]. Any difference between replicate experiments may be due to batch effects or other unobserved factors that are not necessarily reflected in the estimation error. While previous studies have attempted to capture uncertainty by measuring the spread of the residuals from the fitted curves, such as root mean square error, they were not able to capture these false positive biomarkers by setting strict cutoffs [37].

Our Bayesian biomarker model extends the classical ANOVA testing, since it is able to leverage the estimation uncertainty of the IC_{50} values. We showed that taking estimation uncertainty into account in the Bayesian model can lead to both inclusion and exclusion of putative biomarkers. For example, the Bayesian model highlighted the association between *BRAF* mutation in colorectal cancer and BRAF inhibitor response. Targeting BRAF signaling has recently been confirmed as a viable option for metastatic colorectal cancer cases with *BRAF* mutations [38]. In contrast, the Bayesian model excluded a suggestion from ANOVA of association between *KRAS* mutation with lenalidomide response in melanoma. Lenalidomide has thus far had no clinical success in *KRAS* mutant cases nor melanoma [39,40].

Although we systematically tested for drug-biomarker associations, we did observe common behaviour for certain cell types or classes of drugs. The high uncertainty in the response estimates of chromatin histone methylation targeting compounds for instance may be due to the large number of factors contributing to epigenetic regulation of cells [41]. It would be straightforward to extend the GP model to allow for sharing information across drugs or cell lines of similar class, by using either shared hyperparameters or a hyperprior on the hyperparameters. We have not implemented this approach in our work here as our aim was

to show the advantage of fitting individual drug-response using GPs, and extending the method to fitting multiple curves jointly would increase the memory and computational requirements significantly. It is our hope to continue expanding the suite to multiple dimensions of dose-response and biomarker prediction needed for drug combinations, which is predominantly based on synergy modelling with either Loewe Additivity or Bliss Independence [42]. In cases where multiple statistical models converge to concordant biomarkers, this increases the reproducibility of the evidence, potential for clinical translatability and ultimately enables precision medicine.

Methods

Drug screening

We analysed 1,074 cancer cell lines tested with 265 compounds from a high-throughput screen resulting in 225,384 experiments that were previously published [4]. Compounds were tested with 5 to 9 titration points, whilst either diluted with 4- or 2-fold, respectively. Cells were seeded on day zero, left in the microtiter plate for 24 hours to retain linear growth, and consecutively treated for 3 days. After those 3 days of treatment, cellTiterGlo staining is used to quantify ATP levels within each well. In parallel, untreated cells and blank wells were also measured to estimate and normalise cell viability.

Compounds within the replicate study were screened across a 7 point dose response curve with a half-log dilution and 1000 fold range. The duration of drug treatment was 72 hours and cell viability was measured using CellTiter-Glo (Promega). Each cell line and compound pair was screened in technical triplicate, three assay plates generated simultaneously, and across three biological replicates with 46 and 44 days between the first to second and second to third replicates respectively. Cell viability measurements for these experiments can be found in **Supplementary Table S3**.

Preprocessing

Prior to analysis, we scaled the raw observed fluorescent intensities for each drug/cell line combination using the observations from the blank and negative control wells as follows. Let $R = \{r_1, r_2..., r_n\}$ be the observed intensities for *n* dosages. Let B be the mean of the intensities for the blank wells on the same plate as the experiment, and C be the mean of the intensities of the negative control wells (no drug added). Then the relative cell viability V can be calculated as:

$$V = \frac{R - B}{C - B}$$

For the purpose of fitting the Gaussian process models, we additionally rescale the dosages to avoid having to adapt the length-scale hyperparameter to the maximum dosage. We rescale the $\log_2$-transformed dosages $d = \{d_1, d_2..., d_n\}$ as follows:

$$d' = \frac{d + 1}{max(d) + 1}$$

Note that IC_{50} values have been back-transformed to the log10 drug dosage scale for

comparability with those reported in Iorio et al. (2016).

Sigmoid drug response model

The GDSC estimates in Iorio et al. (2016) were obtained using a sigmoid fit to the drug response curve, using the same pre-processing of the fluorescent intensities as described above. The particular sigmoid model used is the one described in Vis et al. (2016). In brief, if we have shape parameter s_i and position parameter ρ_{ij} for cell line i and drug j, then cell viability can be represented as a function of dosage d:

$$f(d, s_i, \rho_{ij}) = \frac{1}{1 + \exp\left(\frac{d - \rho_{ij}}{s_i}\right)}$$

Note that this allows for cell line/drug specific position parameters, but shape parameters that only vary by cell line and are shared across drugs. The position parameter ρ_{ij} corresponds to the estimated IC_{50} for cell line i and drug j. For full details, see Vis et al. (2016).

Gaussian process drug response model

For simplicity, we drop the subscripts ij and present the model for a single drug and cell line combination. We model the drug response y via a two-component Beta mixture such that:

$$P(\mathbf{y}|\mathbf{f}, s_1, \mu_2, s_2, \pi) = \pi Beta^{\mu}(\mathbf{y}|\Phi^{-1}(\mathbf{f}), s_1) + (1 - \pi) Beta^{\mu}(\mathbf{y}|\mu_2, s_2)$$

where $Beta^{\mu}$ is the reparameterization of the Beta distribution in terms of the mean μ and a scale parameter s, and Φ^{-1} is the probit function (the inverse of the standard normal cumulative distribution function). Component 1 represents the drug response, which is driven by a latent Gaussian process $\mathbf{f}$, while component 2 represents outliers that deviate from the overall dose response trend. We set the scale parameters $s_1 = 50$ and $s_2 = 11$ and specify $\mu_2 = 0.9$ to reflect our belief that outliers will mostly be erroneous measurements of resistance. We set $\pi = 0.999$ as we believe that outliers are rare.

We place a standard Gaussian process prior on $\mathbf{f}$, such that:

$$P(\mathbf{f}|d, \Psi) = \mathcal{MVN}(\mathbf{f}|\mathbf{m}, C_{\Psi}(d, d'))$$

where m is the mean drug response, and $C_{\Psi}(d, d')$ is a covariance function with hyperparameters Ψ; in practice we choose a combined linear-Matern3/2 as a flexible option, though other choices are possible. Information sharing across drugs and cell lines can be achieved via common hyperparameters, or shared hyperpriors in a hierarchical model; we choose the former for the application in this paper and empirically set the variance and length scale parameters for the Matern to 0.2 and 0.3 respectively, and the variance parameter for the linear kernel to 0.1.

Inference is performed using variational learning [43], via the GPFlow software [44]. We choose variational learning over alternatives such as Markov chain Monte Carlo due to its speed, which allows us to process large drug response panels in a realistic time frame.

Calculation of summary statistics

Summary statistics of drug response can be calculated straightforwardly by sampling from the posterior of the Gaussian process. Generally, let $g(\mathbf{d}, \mathbf{y})$ be a function that calculates a summary statistic τ from a dose-response curve with dosages $\mathbf{d}$ and responses $\mathbf{y}$, then we can obtain a posterior estimate of the mean of the summary statistic by first sampling N dose-response curves from the posterior of the GP model, and then calculating the average:

$$\bar{\tau} = \frac{1}{N} \sum_{l}^{N} g(\mathbf{d}_l, \mathbf{y}_l)$$

A similar procedure can be used to calculate the posterior estimate of the standard deviation.

Although we can extract other response statistics from our curve fits, the most common are the IC_{50} and the area under the drug response curve (AUC). On the $\log_2$ dosage scale the dosages are equally spaced, and hence AUC can be straightforwardly estimated by the mean function:

$$g_{AUC}(\mathbf{d}, \mathbf{y}) = \frac{1}{n} \sum_{m}^{n} y_m$$

where m indexes over the n dosages. For the IC_{50}, estimation for a single curve is complicated by the fact that the curve may not cross the 50% viability threshold within the observed dosage range (non-crossing sample). We therefore extrapolate the GP samples to 10 times the maximum ($\log_2$) experimental dosage and specify $g_{IC50}(\mathbf{d}, \mathbf{y})$ as:

$$g_{IC50}(\mathbf{d}, \mathbf{y}) = d_m \text{ such that } y_m = 0.5 \text{ if } \exists y_m \leq 0.5$$

Note that this ignores samples where for all dosages, $y_m \geq 0.5$; one could devise a multivariate sufficient statistic that takes this information into account, but we have found that in general there is a reasonable amount of correlation between $g_{IC50}(\mathbf{d}, \mathbf{y})$ and the number of non-crossing samples for a given cell-line/drug combination.

Comparison of GP and sigmoid IC_{50} values

Concordance between IC_{50} values based on sigmoid and GP curve fitting is quantified with Pearson correlation for each drug. To account for tissue specificity and the varying number of cell lines assessed per tissue type, we employed the average weighted Pearson correlation (pw) of the sigmoid-curve versus GP-curve fitted IC_{50} values for the individual cancer types (i).

The weight for a given cancer type i was denoted as $\sqrt{n_i - 1}$, where n_i is the total number of cell lines treated with the drug within this tissue type. The following metric was applied,

$$pw = tanh\left(\frac{\sum_{i=1}^{N}\sqrt{n_i - 1}\, arctanh(p_i)}{\sum_{i=1}^{N}\sqrt{n_i - 1}}\right)$$

where p_i is unweighted Pearson correlation within a cancer type (i) and a total number of tested cancer types is $N = 30$. For a given drug and tissue type combination, at least 10 cell lines need to be treated ($n_i \geq 10$).

Differences in IC_{50} values for each drug response value *j* were consistently defined as

$$df_j = IC50_{j,GP} - IC50_{j,sigmoid}$$

with a total number of tested cell line and drug combinations equalling to $N_j = 171,937$.

Bayesian biomarker testing

Standard statistical approaches for testing the influence of biomarkers on drug response mostly rely on analysis of variance (ANOVA) testing. An ANOVA can be understood as a linear model of the dependent variable g_i (in this case, a summary measure of drug response such as IC_{50}):

$$g_i = \alpha + \beta z_i + \gamma x_i + \epsilon_i$$

where x_i is an indicator variable denoting the group membership of data point i. In our application, the data points are cell lines, z_i indicates group membership, for example the mutation status of a given SNP, and x_i indicates any other covariates that we wish to correct for, such as tissue type. The parameter α captures the global mean of the drug response, while β captures the effect of mutation status on the drug response, γ is the effect of covariates, and ϵ_i is independent Gaussian noise.

This model, while useful, fails to account for the fact that our Gaussian process model provides estimates σ_i of the uncertainty (or standard error) associated with the mean IC_{50} estimates g_i. In order to make use of these uncertainty estimates, we take an idea from Bayesian meta-analysis, and integrate them via a hierarchical model:

$$g_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$
$$\mu_i \sim \mathcal{N}(\alpha + \beta z_i + \gamma x_i, \sigma^{*2})$$

where μ_i is the mean drug response estimate for cell line i, and σ^{*2} is the variance across cell lines (the variance of ϵ_i in the ANOVA example). Note that this model can be reduced to:

We further specify a sparse Gaussian prior $\beta \sim \mathcal{N}(0, 0.1)$ on the effect size parameter to discourage false positives and reflect our prior belief that most mutations are not associated with drug response. We also place a sparse exponential prior $\sigma^{*2} \sim Exp(10)$ to regularize the variance parameter. Finally, $\alpha \sim \mathcal{N}(0, \tau^2)$ is a Gaussian prior on the global mean with

standard error $\tau \sim Gamma(1,1)$. Early exploratory results showed that using the estimates of σ_i directly placed too much weight on experiments with very low estimation uncertainty, leading to unrealistic posterior estimates of the effect size β. To attenuate this, we used a transformed estimate σ_i^c, where the effect of parameter c was explored over the range [0,1], and empirically set to 0.25 for the results reported in this paper.

Inference in this model is performed using Hamiltonian Monte Carlo via the Stan software package Carpenter et al. (2017). We report the posterior mode of β as well as the posterior probability of observing $\beta > 0$ (if the posterior mode is positive) or $\beta < 0$ (if the posterior mode is negative).

Supplements

Supplementary figures

1. Supplementary Figure S1: Investigation of batch effects in the replicate data.
2. Supplementary Figure S2: High estimation uncertainty independent of concentration range.
3. Supplementary Figure S3: AUC estimation uncertainty
4. Supplementary Figure S4: Comparison of sigmoid and GP curve fitting.
5. Supplementary Figure S5: Drug response biomarker comparison based on both curve fittings.

Supplementary tables

1. Supplementary Table S1 - Summary of pharmacogenomic associations based on ANOVA
2. Supplementary Table S2 - Pharmacogenomic associations based on Bayesian testing of GP curve fits
3. Supplementary Table S3 - Raw and curve fitted replicate dataset

Supplementary source code

All code for curve fitting and Bayesian biomarker detection can be found at https://github.com/FrankD/GPDrugModels.

Acknowledgements

The M.J.G. laboratory is supported by the Wellcome Trust (206194). D.W. is supported by the NIHR Sheffield Biomedical Research Centre, Rosetrees Trust (ref: A2501), and the Academy of Medical Sciences Springboard (ref: SBF004/1052). We thank Benjamin Sidders

and Oliver Stegle for feedback on the methodology.

Competing interests

Jonathan Dry is an employee of AstraZeneca.

References

1. Cook, D. *et al.* Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discov.* **13**, 419–431 (2014).
2. Arrowsmith, J. & Miller, P. Phase II and Phase III attrition rates 2011–2012. *Nat. Rev. Drug Discov.* **12**, 569–569 (2013).
3. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
4. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740–754 (2016).
5. Rees, M. G. *et al.* Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat. Chem. Biol.* **12**, 109–116 (2016).
6. Seashore-Ludlow, B. *et al.* Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discov.* **5**, 1210–1223 (2015).
7. Fallahi-Sichani, M., Honarnejad, S., Heiser, L. M., Gray, J. W. & Sorger, P. K. Metrics other than potency reveal systematic variation in responses to cancer drugs. *Nat. Chem. Biol.* **9**, 708–714 (2013).
8. Haverty, P. M. *et al.* Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* **533**, 333–337 (2016).
9. Haibe-Kains, B. *et al.* Inconsistency in large pharmacogenomic studies. *Nature* **504**, 389–393 (2013).

10. Costello, J. C. *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **32**, 1202–1212 (2014).
11. Menden, M. P. *et al.* Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nature Communications* vol. 10 (2019).
12. Silverbush, D. *et al.* Cell-Specific Computational Modeling of the PIM Pathway in Acute Myeloid Leukemia. *Cancer Res.* **77**, 827–838 (2017).
13. Dawson, D. A. *et al.* Evaluation of an asymmetry parameter for curve-fitting in single-chemical and mixture toxicity assessment. *Toxicology* **292**, 156–161 (2012).
14. Wang, Y., Jadhav, A., Southal, N., Huang, R. & Nguyen, D.-T. A grid algorithm for high throughput fitting of dose-response curve data. *Curr. Chem. Genomics* **4**, 57–66 (2010).
15. Di Veroli, G. Y. *et al.* Combenefit: an interactive platform for the analysis and visualization of drug combinations. *Bioinformatics* **32**, 2866–2868 (2016).
16. Vis, D. J. *et al.* Multilevel models improve precision and speed of IC50 estimates. *Pharmacogenomics* **17**, 691–700 (2016).
17. Lopez-Lopera, A. F. & Alvarez, M. A. Switched Latent Force Models for Reverse-Engineering Transcriptional Regulation in Gene Expression Data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **16**, 322–335 (2019).
18. Boukouvalas, A., Hensman, J. & Rattray, M. BGP: identifying gene-specific branching dynamics from single-cell data with a branching Gaussian process. *Genome Biol.* **19**, 65 (2018).
19. Rasmussen, C. E. & Williams, C. K. I. Gaussian Processes for Machine Learning. (2005) doi:10.7551/mitpress/3206.001.0001.
20. Shi, J. Q. & Choi, T. Gaussian Process Regression Analysis for Functional Data. (2011) doi:10.1201/b11038.
21. Tian, L. *et al.* Gaussian process emulators for quantifying uncertainty in CO2 spreading

predictions in heterogeneous media. *Computers & Geosciences* vol. 105 113–119 (2017).

22. Chapman, P. B. *et al.* Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. *New England Journal of Medicine* vol. 364 2507–2516 (2011).
23. Yang, J. C.-H. *et al.* LUX-Lung 3: A randomized, open-label, phase III study of afatinib versus pemetrexed and cisplatin as first-line treatment for patients with advanced adenocarcinoma of the lung harboring EGFR-activating mutations. *Journal of Clinical Oncology* vol. 30 LBA7500–LBA7500 (2012).
24. Tamura, K. & Fukuoka, M. Gefitinib in non-small cell lung cancer. *Expert Opin. Pharmacother.* **6**, 985–993 (2005).
25. Konecny, G. E. *et al.* Activity of the Dual Kinase Inhibitor Lapatinib (GW572016) against HER-2-Overexpressing and Trastuzumab-Treated Breast Cancer Cells. *Cancer Res.* **66**, 1630–1639 (2006).
26. Zhang, L.-Y. *et al.* Anti-proliferation effect of APO866 on C6 glioblastoma cells by inhibiting nicotinamide phosphoribosyltransferase. *Eur. J. Pharmacol.* **674**, 163–170 (2012).
27. Lucena-Cacace, A., Umeda, M., Navas, L. E. & Carnero, A. NAMPT as a Dedifferentiation-Inducer Gene: NAD+ as Core Axis for Glioma Cancer Stem-Like Cells Maintenance. *Front. Oncol.* **9**, (2019).
28. Lucena-Cacace, A., Otero-Albiol, D., Jiménez-García, M. P., Peinado-Serrano, J. & Carnero, A. NAMPT overexpression induces cancer stemness and defines a novel tumor signature for glioma prognosis. *Oncotarget* **8**, 99514 (2017).
29. Jin, X. *et al.* The p38 MAPK inhibitor BIRB796 enhances the antitumor effects of VX680 in cervical cancer. *Cancer Biol. Ther.* **17**, 566–576 (2016).
30. Yasui, H. *et al.* BIRB 796 enhances cytotoxicity triggered by bortezomib, heat shock protein (Hsp) 90 inhibitor, and dexamethasone via inhibition of p38 mitogen-activated

protein kinase/Hsp27 pathway in multiple myeloma cell lines and inhibits paracrine tumour growth. *Br. J. Haematol.* **136**, 414–423 (2007).

31. Oba, A. *et al.* ARID2 modulates DNA damage response in human hepatocellular carcinoma cells. *J. Hepatol.* **66**, 942–951 (2017).
32. Hodis, E. *et al.* A Landscape of Driver Mutations in Melanoma. *Cell* **150**, 251 (2012).
33. Ding, L. *et al.* Clonal Architectures and Driver Mutations in Metastatic Melanomas. *PLoS One* **9**, e111153 (2014).
34. Prahallad, A. *et al.* Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature* **483**, 100–103 (2012).
35. Corcoran, R. B. *et al.* Combined BRAF, EGFR, and MEK Inhibition in Patients with BRAFV600E-Mutant Colorectal Cancer. *Cancer Discovery* vol. 8 428–443 (2018).
36. Riihimäki, J. & Vehtari, A. Gaussian processes with monotonicity information. in *Proceedings of the thirteenth international conference on artificial intelligence and statistics* 645–652 (jmlr.org, 2010).
37. Cokelaer, T. *et al.* GDSCTools for mining pharmacogenomic interactions in cancer. *Bioinformatics* **34**, 1226–1228 (2018).
38. Kopetz, S. *et al.* Encorafenib, Binimetinib, and Cetuximab in V600E-Mutated Colorectal Cancer. *N. Engl. J. Med.* **381**, 1632–1643 (2019).
39. Gandhi, A. K. *et al.* Immunomodulatory Effects in a Phase II Study of Lenalidomide Combined with Cetuximab in Refractory KRAS-Mutant Metastatic Colorectal Cancer Patients. *PLoS One* **8**, (2013).
40. Glaspy, J. *et al.* Results of a multicenter, randomized, double-blind, dose-evaluating phase 2/3 study of lenalidomide in the treatment of metastatic malignant melanoma. *Cancer* **115**, 5228–5236 (2009).
41. Michalak, E. M., Burr, M. L., Bannister, A. J. & Dawson, M. A. The roles of DNA, RNA and histone methylation in ageing and cancer. *Nat. Rev. Mol. Cell Biol.* **20**, 573–589

(2019).

42. Vlot, A. H. C., Aniceto, N., Menden, M. P., Ulrich-Merzenich, G. & Bender, A. Applying synergy metrics to combination screening data: agreements, disagreements and pitfalls. *Drug Discov. Today* **24**, 2286–2298 (2019).
43. Hensman, J., Fusi, N. & Lawrence, N. D. Gaussian processes for Big data. in *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence* 282–290 (AUAI Press, 2013).
44. Matthews, A. G. de G. *et al.* GPflow: A Gaussian Process Library using TensorFlow. *Journal of Machine Learning Research* **18**, 1–6 (2017).