ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Cite this article: Savage RS, Yuan Y. 2016 Predicting chemoinsensitivity in breast cancer with 'omics/digital pathology data fusion. *R. Soc. open sci.* **3**: 140501.
http://dx.doi.org/10.1098/rsos.140501

Received: 4 December 2014
Accepted: 8 January 2016

Subject Category:
Genetics

Subject Areas:
health and disease and epidemiology/bioinformatics

Keywords:
breast cancer, data integration, Bayesian

Author for correspondence:
Richard S. Savage
e-mail: r.s.savage@warwick.ac.uk

Predicting chemoinsensitivity in breast cancer with 'omics/digital pathology data fusion

Richard S. Savage[1,2] and Yinyin Yuan[3]

[1]Systems Biology Centre, and [2]Warwick Medical School, University of Warwick, Warwick, UK
[3]Division of Molecular Pathology, Centre for Evolution and Cancer, The Institute of Cancer Research, London, UK

Predicting response to treatment and disease-specific deaths are key tasks in cancer research yet there is a lack of methodologies to achieve these. Large-scale 'omics and digital pathology technologies have led to the need for effective statistical methods for data fusion to extract the most useful patterns from these diverse data types. We present *FusionGP*, a method for combining heterogeneous data types designed specifically for predicting outcome of treatment and disease. *FusionGP* is a Gaussian process model that includes a generalization of feature selection for biomarker discovery, allowing for simultaneous, sparse feature selection across multiple data types. Importantly, it can accommodate highly nonlinear structure in the data, and automatically infers the optimal contribution from each input data type. *FusionGP* compares favourably to several popular classification methods, including the Random Forest classifier, a stepwise logistic regression model and the Support Vector Machine on single data types. By combining gene expression, copy number alteration and digital pathology image data in 119 estrogen receptor (ER)-negative and 345 ER-positive breast tumours, we aim to predict two important clinical outcomes: death and chemoinsensitivity. While gene expression data give the best predictive performance in the majority of cases, the digital pathology data are much better for predicting death in ER cases. Thus, *FusionGP* is a new tool for selecting informative features from heterogeneous data types and predicting treatment response and prognosis.

1. Introduction

Post-genomic molecular biology is revolutionizing the study of cancer [1,2]. New measurement technologies are giving us fresh insights into the molecular and genetic mechanisms underlying the disease, and modelling these data gives us ways to predict

THE ROYAL SOCIETY PUBLISHING

the likely progression and outcome of disease in a given patient [3,4]. By identifying informative markers related to critical events, there is unprecedented potential for both the development of new prognostic/diagnostic tests, and also for furthering our understanding of cancer's key driving molecular and genetic mechanisms.

The digitisation of high-quality tumour pathology images now gives us a complementary perspective at the point of diagnosis [5]. Tumour cells can be identified in these images computationally and the resulting data on cell morphology and spatial distribution used as an additional source of information in analysing a given cancer case. We therefore have access to multiple data types for a given cancer patient. The most effective prediction of clinical outcomes then requires that we develop statistical algorithms that are able to effectively combine these diverse data types. This presents several challenges which we aim to address in this paper.

(i) *Challenge* 1: *outcome prediction*. Given a training set of measurements, we wish to predict the likely outcomes of treatment or prognosis for a new patient for whom we have analogous measurements. This can inform clinicians of the likely disease course and hence help design efficient treatment strategy.
(ii) *Challenge* 2: *biomarker discovery*. The available data are often high-dimensional. By identifying in a principled way which features are informative by allowing for the possibility of *sparse solutions*, we identify a small subset of the features are informative about clinical outcome.
(iii) *Challenge* 3: *effective integration of multiple data types*. To get the best possible results, we should include all relevant information and do so in a principled way. In particular, different types of data offer complementary perspectives on a given disease. We, therefore, wish to generalize the notion of feature selection to include multiple, possibly heterogeneous data types.

There have been a number of previous attempts to address various of these challenges, for example, Futschik *et al.* [6], Nevins *et al.* [7], Pittman *et al.* [8], Stephenson *et al.* [9], Gevaert *et al.* [10], Boulesteix *et al.* [11], Daemen *et al.* [12], Obulkasim *et al.* [13]. However, the field as yet lacks a comprehensive framework that is suitable for combining data from the types of highly heterogeneous biomedical data that are being generated in increasingly large volumes. To address this, we adopt a statistical modelling approach. Such a model must be able to handle high-dimensional, noisy data in a principled way and must be capable of learning a wide range of possible structure. Non-parametric Bayesian models provide just such a framework. By learning a level of structure appropriate to the data, they naturally encode a kind of Occam's razor, and their Bayesian nature lends itself naturally to the inclusion of multiple sources of data and/or prior knowledge. We have previously developed Bayesian non-parametric models for data integration in an unsupervised setting, e.g. Savage *et al.* [14], Yuan *et al.* [15], Kirk *et al.* [16], but there is much work still to be done in the supervised setting.

2. Material and methods

We present *FusionGP*, a Bayesian non-parametric method for integrating multiple data types, to perform classification and sparse biomarker discovery. The relationship between input features and outcome is modelled via a set of unknown latent functions (one per data type) that are constrained via a set of Gaussian process (GP) priors. The input features for each data type are selected via a slab-and-spike prior, where the probability of a feature being *switched on* is inferred as part of the inference process, allowing for sparse feature selection in a given data type, where the data support it.

2.1. Gaussian processes

We model the relationship between each data type and the target values via (unknown) latent functions. We assume the latent functions to be realizations of a zero-mean GP [17], noting that this approach has a successful track record in modelling molecular data [18–20].

For convenience, we adopt a compact notation for the latent function for the dth data type, f_d. Given this, we define the following GP prior over the space of unknown functions for the dth data type:

$$P(f_d) = (2\pi)^{-n/2}|K|^{-1/2}\exp\left(-\tfrac{1}{2}{f_d}^{\mathrm{T}}(K)^{-1}f_d\right). \tag{2.1}$$

For n data points, and where K is the $N \times N$ covariance matrix that defines the GP.

2.2. Feature selection

For GP models, the features contribute via the covariance function. To allow feature selection, we define additional indicator parameters, I_{jd} for the jth feature in the dth data type . When $I_{jd} = 1$, the jth feature is *switched on* and behaves as usual. When $I_{jd} = 0$, the feature is excluded from the analysis. In molecular data, we expect that in general the number of features that are informative may be relatively small. We, therefore, apply a sparsity prior to the I_{jd}, to encode this knowledge:

$$P(I_{jd}) \propto 2^{-a_d} w_d^{a_d} (1 - w_d)^{J_d - a_d}, \tag{2.2}$$

where $a_d = \Sigma_j I_{jd}$, the number of features switched on in data type d and J_d is the total number of features in data type d.

This is, therefore, a product of Bernoulli distributions (one per indicator parameter), with an additional factor of 2^{-a_d} which penalizes individual features being switched on. This equation is, therefore, the joint distribution for all I_{jd} for a given data type.

The hyperparameters w_d are inferred and allow the model to determine the appropriate level of sparsity, given the data. We apply a binomial prior on w_d, using this to encode our prior belief that at least a small number of the features will be informative. We also enforce a hard prior of $w_d \leq 0.5$, because we do not wish the prior itself to favour switching features on:

$$w \sim \beta(10, 10). \tag{2.3}$$

2.3. Covariance function

One of the great advantages of GP models is that we have access to a wide range of covariance functions, all using the same overall model framework. For the synthetic data analyses in this paper, we use four different covariance functions:

$$K_{\text{linear}}(x_i, x_j) = A_0 + \boldsymbol{x_i}.\boldsymbol{x_j}, \tag{2.4}$$

$$K_{\text{SE}}(x_i, x_j) = \exp\left(-\frac{r^2}{2l^2}\right), \tag{2.5}$$

$$K_{\text{Matern}}(x_i, x_j) = \left(1 + \frac{\sqrt{3}r}{l}\right) \exp\left(-\frac{\sqrt{3}r}{l}\right) \tag{2.6}$$

and

$$K_{\text{sum}}(x_i, x_j) = K_{\text{linear}}(x_i, x_j) + K_{\text{SE}}(x_i, x_j), \tag{2.7}$$

where 'SE' stands for 'square exponential' and $r = |\boldsymbol{x_i} - \boldsymbol{x_j}|$.

For simplicity, all analyses presented in this paper use the same covariance function for data types in said analysis. This is chosen for convenience—*FusionGP* does not require this. *FusionGP* uses the GPML MATLAB library [21], which means that it can be run with a wide range of different covariance functions. We are grateful to the GPML authors for their provision of this library and would recommend it highly to anyone working with GPs in MATLAB.

2.4. Model specification

To model binary outcomes, we choose the probit function [22]:

$$P(D|\boldsymbol{f_d}) \sim \text{probit}. \tag{2.8}$$

We model each data type using an (unknown) latent function, that relates the input features $\boldsymbol{x}$ to the outcome of interest, y, via the probit likelihood. We wish these latent functions to be highly flexible so that they can capture the complex biological structure underlying the data. We therefore choose to draw the latent functions from zero-mean GPs:

$$\boldsymbol{f_d} \sim \text{GP}(0, K_d). \tag{2.9}$$

To include information from multiple data sources, we use the approach of Girolami & Zhong [23]. In this approach, we assume that the datasets are conditionally independent of one another given the target values. This results in a model that is easy to work with and is nevertheless extremely powerful.

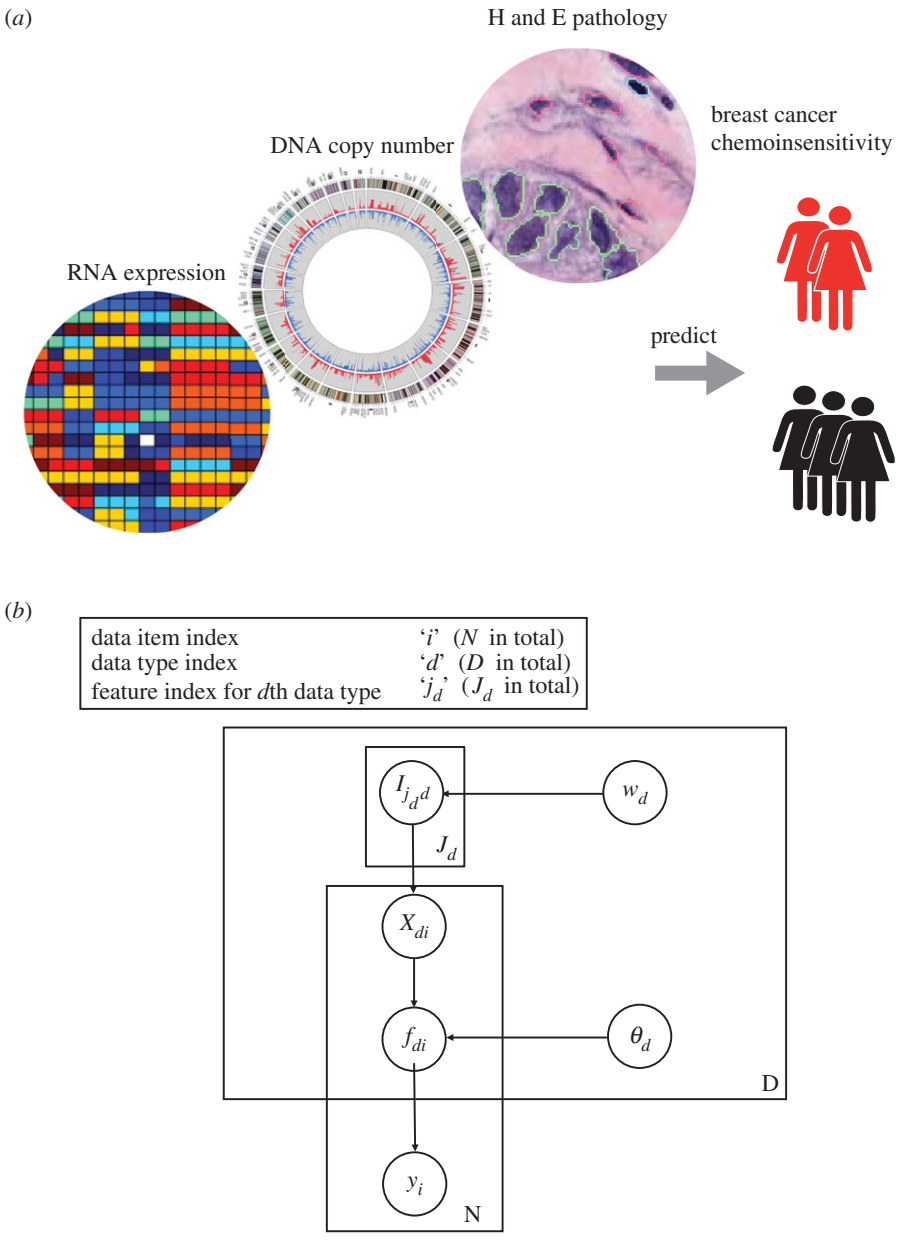

Figure 1. (*a*) The aim of *FusionGP* is to predict clinical outcome such as chemoinsensitivity by integrative analysis of multiple data types. (*b*) Graphical model for *FusionGP*. y_i are the target labels. f_{di} is the latent function value for the ith data item in the dth data type. θ_d are the hyperparameters that define the covariance function for the dth data type. X_{di} is the data value for the ith data item in the dth data type. $I_{j_d d}$ is the feature selection indicator parameter for the j_dth feature in the dth data type. w_d are the hyperparameters that encode sparsity in the feature selection.

The overall latent function is therefore a linear combination of the individual latent functions:

$$\boldsymbol{f} = \sum_d A_d \boldsymbol{f_d}, \tag{2.10}$$

where A_d is the scaling parameter for the dth latent function. We note that this model has the great merit of scaling well with the number of datasets. For example, the number of covariance parameters is proportional to the number of datasets.

The features in each data type are assigned indicator parameters, I_{jd}, subject to the sparsity priors discussed in the above section on feature selection. The covariance matrix for a given data type is computed using only the features for which $I_{jd} = 1$. This, therefore, encodes sparse feature selection into the *FusionGP* model. A graphical model representation of the *FusionGP* model is shown in figure 1.

2.5. Inference

Owing to the form of the likelihood, this model does not have a closed form. We, therefore, use Markov chain Monte Carlo (MCMC) to perform inference. The challenge in doing this is one of computation time. MCMC analysis typically requires at least $O(10^4 - 10^5)$ evaluations of the likelihood, and inference with GPs requires Cholesky decompositions of the covariance matrix for each evaluation, which scale as the cube of the number of data items. For types of data considered in this paper, the number of data items will be tractable.

In the context of GPs, Cholesky decompositions tend to become challenging to use for more than $O(10^4)$ data items. Given that this method uses MCMC (rather than parameter optimization), we suspect that for more than a few thousand data items, the proposal model will be sufficiently slow as to be intractable. However, there are now many effective ways to scale GPs to larger numbers of data items, so this would probably provide a plausible way forward in such cases.

We note that the Cholesky decomposition has no dependence on the number of features. For the covariance functions considered in this paper, computing the covariance matrix is linear in the number of features, and will typically not be the dominant step in the algorithm. *FusionGP*, therefore, scales extremely well with the both the number of features in each data type, and also the total number of features across all data types.

We sample the logarithm of the covariance hyperparameters (w_d and A_d) using a Metropolis–Hastings algorithm with Gaussian proposal distribution, the variance of which we tune using samples from an initial burn-in period of the MCMC chain. We choose to work with the log-parameters for convenience as the parameters cannot take negative values.

When using GPs, we marginalize over the (unknown) latent function. Because in *FusionGP* the likelihood is not closed form, we do this as part of the MCMC algorithm. Specifically, we define a set of parameters that are the latent function values at the training points. We sample these using a Metropolis–Hastings algorithm with Gaussian proposal distribution, with variance again tuned using samples from an initial burn-in period of the MCMC chain. The latent function values are then effectively marginalized as part of the MCMC sampling process. We note that there are more sophisticated sampling schemes that one could use here, such as elliptical slice sampling [24]. We found a simpler scheme adequate for the results presented in this paper, but future improvements can certainly be made to the sampling scheme.

2.6. Markov chain Monte Carlo performance

For each analysis in this paper, we run five MCMC chains and combine the results, after removing the first 50% as burn-in and sparse-sampling the chains by a factor of 10. Each chain is run for 48 h in the University of Warwick's High Performance Computing cluster. The results presented in this paper are, therefore, produced using a fixed total run-time.

We note that MCMC convergence can be challenging for models dealign with high-dimensional feature selection, and *FusionGP* is no exception. We assess the convergence of our analyses by looking by eye at a range of one-dimensional marginal posteriors (histograms), comparing the equivalent plot for each of the chains (which should give the same distribution if the chains are converged). This could be further improved in future by taking a more in-depth statistical approach, but our practical experience is that this approach works well in practice.

3. Examples

We validate *FusionGP* on both synthetic and real breast cancer data. The synthetic data are generated to provide a rigorous test of *FusionGP's* performance where the ground truth is known, and to facilitate comparisons with existing methods. The breast cancer dataset is taken from the recently published METABRIC study of over 2000 patients [25].

3.1. Synthetic data

We generate two synthetic data types, each with the same underlying signal, and different noise realizations drawn from the same Gaussian distribution. These data are designed to represent the case where we have a significant number of features, of which a relatively small proportion are informative.

Each data type contains 500 items (250 outcome = true and 250 outcome = false cases). There are 600 features, 100 of which contain signal and the remaining 500 of which contain only noise. Each signal

feature is generated by taking the known target values and adding Gaussian noise, whose variance is fixed for a given data type. The noise features are generated solely by drawing from the same Gaussian noise distribution, i.e. without any signal.

3.2. METABRIC breast cancer data

METABRIC is a large-scale study of the genomic and transcriptomic landscape of breast cancer [25]. The METABRIC data contains copy number, gene expression, histopathological haematoxylin and eosin (H&E) images, and clinical information for over 2000 patients. H&E stained images of tumour slides are potentially highly complementary to the molecular data types. The METABRIC images have been quantitatively analysed by Yuan *et al.* [5]. Nuclear morphological features of different types of cells including cancer, lymphocytes and stromal cells which include fibroblasts and endothelial cells were quantitatively measured. These features include topological and image moment features to measure elongation, size, texture, etc., yielding 100 features in total. Median, variance and skewness were used to characterize the distribution of nuclear features in each tumour. As a result, we obtained imaging data for each patient tumour where each of the morphological features was summarized by median, variance or skewness, totalling 900 features.

We considered two clinical outcomes: death and chemoinsensitivity. All are defined as binary outcomes for simplicity of modelling. This will result in some information loss on the case of death, which would be more properly modelled as a (censored) survival time. However, this choice allows us to focus in paper more on other aspects of the model. We, therefore, considered only samples from patients who have either survived for 10 years, or who are confirmed to have died prior to that point. This leaves us with 464 samples for which we have copy number, gene expression and image feature data. Of these 464 samples, 119 were estrogen receptor (ER) negative (ER−) and 345 were ER positive (ER+). In addition to analysing the whole set of samples, we also considered the ER− and ER+ subsets, as these subsets are biologically distinct disease from one another. We, therefore, define the clinical outcomes as follows.

— *Death*. The patient died of breast cancer within 10 years of diagnosis with breast cancer.
— *Chemoinsensitivity*. The patient died (as above) and had been treated with chemotherapy.

4. Results

4.1. Results on synthetic data

Tables 1 and 2 show the results of *FusionGP* analyses of the synthetic data. Table 2 gives 'oracle' results, where only 'signal' features are used in the analysis. Table 1 gives the results for the case where the correct features are not known, so the algorithm must use feature selection. Comparing the two sets of results, we can see that while the presence of noise features has an adverse impact on the outcome prediction area under curve (AUC), we are still able to make useful predictions even when there are five times as much noise as informative features. The biomarker AUC values show that correctly identifying the informative 'biomarker' features is more challenging than simply making predictions. This may be an indication as to one reason why biomarker discovery in genomic cancer data has proven such a challenge to date.

Table 3 shows comparison analyses using the Random Forest (RF) model [26], a stepwise logistic regression generalized linear model (GLM) method and a simple Support Vector Machine (SVM) with no feature selection. These represent well-known classification algorithms that would be sensible default choices for analysis of single data types such as those considered in this paper. The results show that all three comparison methods are significantly impacted by the presence of large numbers of noise features, under-performing in comparison to *FusionGP*. We also note that even in the oracle case, *FusionGP* shows somewhat superior performance.

4.2. Results on METABRIC data

We ran twofold cross-validation analyses to assess the performance of *FusionGP* on the METABRIC breast cancer data. For these analyses, we chose the SE covariance function, as it was best-performing on the synthetic data runs. Tables 4–6 show the prediction AUC from the cross-validation. We compared the AUC scores of *FusionGP* and each of the three comparison methods. The differences are computed for each single data type comparison in tables 4–6. Figure 2 shows scatter plots of the single data type results for *FusionGP*. Comparing the distributions of AUC scores for each method using the Wilcoxon rank sum

Table 1. Synthetic data results. (Shown are area-under-curve (AUC) values from receiver operating characteristic (ROC) curves for both outcome and biomarker predictions.)

run	covar.	outcome AUC	biomarker AUC
synth. (2 types)	SE	0.74 ± 0.04	0.59 ± 0.08
synth. (2 types)	Matern	0.70 ± 0.02	0.59 ± 0.04
synth. (2 types)	sum	0.70 ± 0.02	0.58 ± 0.04
synth. (2 types)	linear	0.71 ± 0.01	0.59 ± 0.10
synth. (1 type)	SE	0.68 ± 0.04	0.68 ± 0.04
synth. (1 type)	Matern	0.68 ± 0.04	0.62 ± 0.14
synth. (1 type)	sum	0.69 ± 0.04	0.62 ± 0.11
synth. (1 type)	linear	0.68 ± 0.04	0.66 ± 0.03

Table 2. Synthetic data 'oracle' results, where the correct features are known. (Shown are area-under-curve (AUC) values from ROC curves.)

run	covar.	outcome AUC
oracle (2 types)	linear	0.86 ± 0.02
oracle (2 types)	SE	0.87 ± 0.03
oracle (2 types)	Matern	0.86 ± 0.02
oracle (2 types)	sum	0.86 ± 0.03
oracle (1 type)	linear	0.78 ± 0.05
oracle (1 type)	SE	0.79 ± 0.04
oracle (1 type)	Matern	0.79 ± 0.04
oracle (1 type)	sum	0.79 ± 0.05

Table 3. Synthetic data results for comparison methods. (For the 'oracle' runs, only the informative features are included in the analysis.)

data	method	prediction AUC	biomarker AUC
synth. (1 type)	RF	0.63 ± 0.03	0.57 ± 0.01
synth. (1 type)	stepwise GLM	0.61 ± 0.03	0.54 ± 0.01
synth. (1 type)	SVM (no selection)	0.66 ± 0.02	—
oracle (1 type)	RF	0.74 ± 0.01	—
oracle (1 type)	stepwise GLM	0.74 ± 0.01	—
oracle (1 type)	SVM (no selection)	0.75 ± 0.03	—

test, we obtain $p = 0.66$ (RF), 8×10^{-4} (stepwise GLM) and 4×10^{-3} (SVM). We therefore conclude that *FusionGP* significantly outperforms both stepwise GLM and SVM methods, and marginally outperforms RF for the METABRIC data types.

For analysis using death as outcome, we found a striking difference between the ER− and ER+ item subsets. For the ER+ items, the molecular data are highly informative. However, the combination of copy number and gene expression data produce results consistent with those obtained for gene expression alone. We suggest that this is due to a duplication of information in that the effect of the gene expression accompany copy number alteration. In ER− tumours both molecular data types are poorly informative. The image features, however, are highly informative, suggesting that cell morphology information are important in predicting survival in breast cancer.

Table 4. ROC curve AUC results from analysing different combinations of METABRIC data types. (The best-performing method/s in each row are highlighted in bold. CN, copy number; GE, gene expression.)

	data	*FusionGP*	RF	GLM	SVM
chemoinsens.	CN	**0.72 ± 0.01**	0.68 ± 0.01	0.63 ± 0.05	0.61 ± 0.06
	GE	**0.78 ± 0.02**	0.74 ± 0.02	0.71 ± 0.01	0.70 ± 0.02
	image	**0.55 ± 0.04**	**0.58 ± 0.06**	**0.58 ± 0.01**	**0.57 ± 0.01**
	GE/CN	**0.78 ± 0.02**	—	—	—
	all	**0.78 ± 0.03**	—	—	—
death	CN	**0.69 ± 0.01**	0.68 ± 0.01	0.53 ± 0.02	0.62 ± 0.08
	GE	**0.75 ± 0.01**	0.74 ± 0.01	0.64 ± 0.05	0.68 ± 0.03
	image	**0.60 ± 0.01**	**0.59 ± 0.01**	0.49 ± 0.02	**0.60 ± 0.01**
	GE/CN	**0.75 ± 0.01**	—	—	—
	all	**0.75 ± 0.01**	—	—	—

Table 5. ROC curve AUC results, using only the 119 ER− patients. (The best-performing method/s in each row are highlighted in bold.)

	data	*FusionGP*	RF	GLM	SVM
chemoinsens.	CN	**0.59 ± 0.10**	**0.60 ± 0.14**	0.52 ± 0.07	0.54 ± 0.16
	GE	**0.59 ± 0.07**	**0.60 ± 0.02**	0.51 ± 0.05	0.55 ± 0.02
	image	**0.61 ± 0.07**	**0.60 ± 0.03**	0.55 ± 0.01	**0.59 ± 0.10**
	GE/CN	**0.59 ± 0.11**	—	—	—
	all	**0.59 ± 0.08**	—	—	—
death	CN	**0.56 ± 0.07**	**0.54 ± 0.06**	0.51 ± 0.02	0.46 ± 0.01
	GE	**0.55 ± 0.03**	**0.56 ± 0.06**	**0.51 ± 0.06**	**0.52 ± 0.04**
	image	**0.72 ± 0.03**	**0.73 ± 0.04**	0.69 ± 0.02	0.69 ± 0.01
	GE/CN	**0.55 ± 0.02**	—	—	—
	all	**0.67 ± 0.01**	—	—	—

Table 6. ROC curve AUC results, using only the 345 ER+ patients. (The best-performing method/s in each row are highlighted in bold.)

	data	*FusionGP*	RF	GLM	SVM
chemoinsens.	CN	**0.68 ± 0.01**	0.62 ± 0.01	0.58 ± 0.03	0.62 ± 0.02
	GE	**0.70 ± 0.02**	0.66 ± 0.03	**0.64 ± 0.11**	0.60 ± 0.01
	image	0.52 ± **0.04**	**0.58 ± 0.08**	0.46 ± 0.08	0.50 ± 0.09
	GE/CN	**0.70 ± 0.02**	—	—	—
	all	**0.70 ± 0.03**	—	—	—
death	CN	**0.71 ± 0.02**	**0.71 ± 0.03**	0.58 ± 0.11	0.60 ± 0.03
	GE	**0.75 ± 0.01**	0.74 ± 0.01	**0.75 ± 0.01**	0.68 ± 0.01
	image	**0.62 ± 0.04**	**0.64 ± 0.04**	0.52 ± 0.04	**0.60 ± 0.05**
	GE/CN	**0.75 ± 0.01**	—	—	—
	all	**0.75 ± 0.01**	—	—	—

For predicting chemoinsensitivity, we found that for the ER+ samples the molecular data are most informative. For the ER− samples, our results suggest all three data types contain some relevant information. Of note for chemoinsensitivity is the performance of the all-samples case. The predictive performance for this case is significantly better than that of either ER− or ER+ cases. There must,

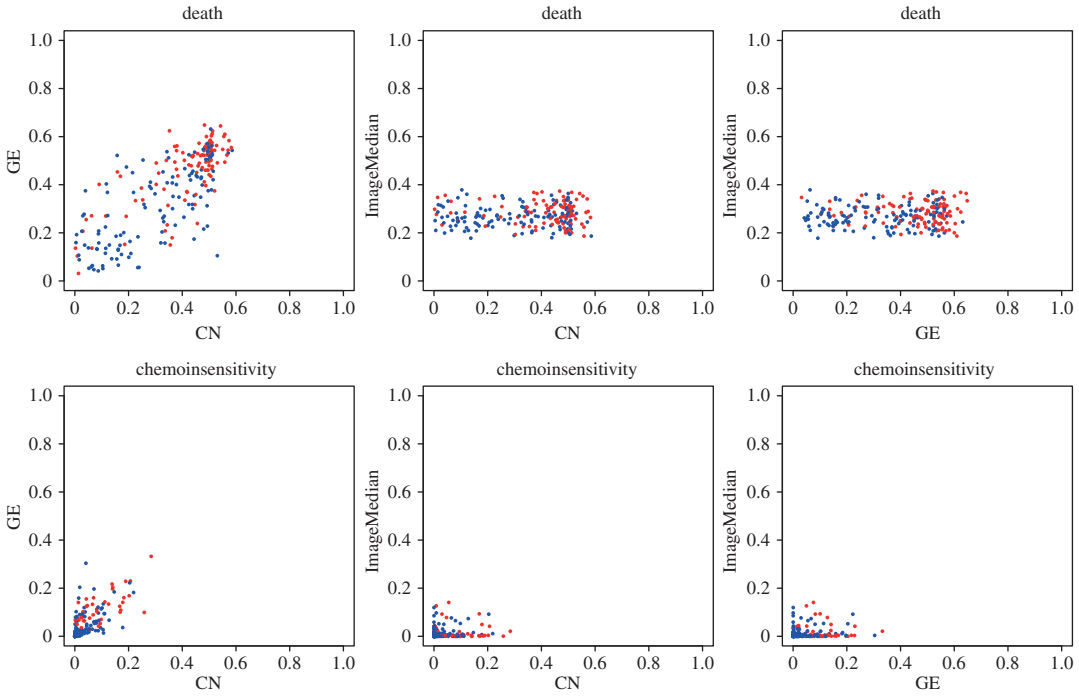

Figure 2. Scatter plots for outcome predictions of all patients, comparing results for single data types. Outcome, FALSE is shown in blue, outcome, TRUE is shown in red. CN, copy number; GE, gene expression.

therefore, be significant information sharing between the ER− and ER+ cases, in contrast with the analyses for death, where the two ER subsets are highly distinct.

We compared *FusionGP* for single data types to the RF, stepwise GLM (logistic regression) and SVM with no feature selection. Of the outcome/single-data-type combinations considered in tables 4–6, *FusionGP* is either the best performing method or within one standard deviation of the best result in 26 of the 27 cases. RF also performed well, with joint-best performance in 16 cases and overall best performance in one case. Stepwise GLM and the SVM were generally outperformed by both *FusionGP* and RF, with a few exceptions.

For the image features, we found that across all analyses about a third of them were selected. This suggests that there is duplicate information, which is expected due to the nature of their construction. For example, acircularity (median acirc) and HuÃŢs first moment or I1 (median I1) both measure shape irregularity. There is minimal downside here in terms of predictive performance, although it suggests that it may be possible to construct a smaller set of equally informative image features, which would have some benefit in terms of the run time of any analysis.

We analysed the most strongly selected features for each data type. For the molecular types, we looked for enrichment in gene ontology (GO) ontologies and KEGG pathways. We focused on biological process in GO to identify processes underlying features predicted to be informative for important clinical phenomenon. Figure 3 shows specific biological processes enriched in copy number features that were found to be predictive of chemoinsensitivity by our algorithm in ER− and ER+ tumours. Multiple cell cycle processes forming large cell cycle modules are evident in both ER− and ER+ tumours, signifying the role of cell-cycle pathways across ER status yet highlighting different cell-cycle phases for different ER status, e.g. transition phases including G2/M and G1/S for ER− and G1 and G2 for ER+. Notably, activation of innate immune response is significantly enriched in the selected features for predicting chemoinsensitivity only in ER− tumours. This is consistent with recent evidences showing the importance of immune response in achieving pathological complete response to chemotherapy in ER− tumours [27,28]. Our result again confirmed their observations yet extended to the genomics level, indicating specific genomic aberrations that could underly innate immune response and influence the response of ER− patients to chemotherapy.

We examined copy number and gene expression features for predicting deaths in ER+ because of their good performance as indicated in table 6. Compared with figure 3, large cell-cycle modules can also be found in the biological processes enriched in both feature types (figure 4). Of particular note, however,

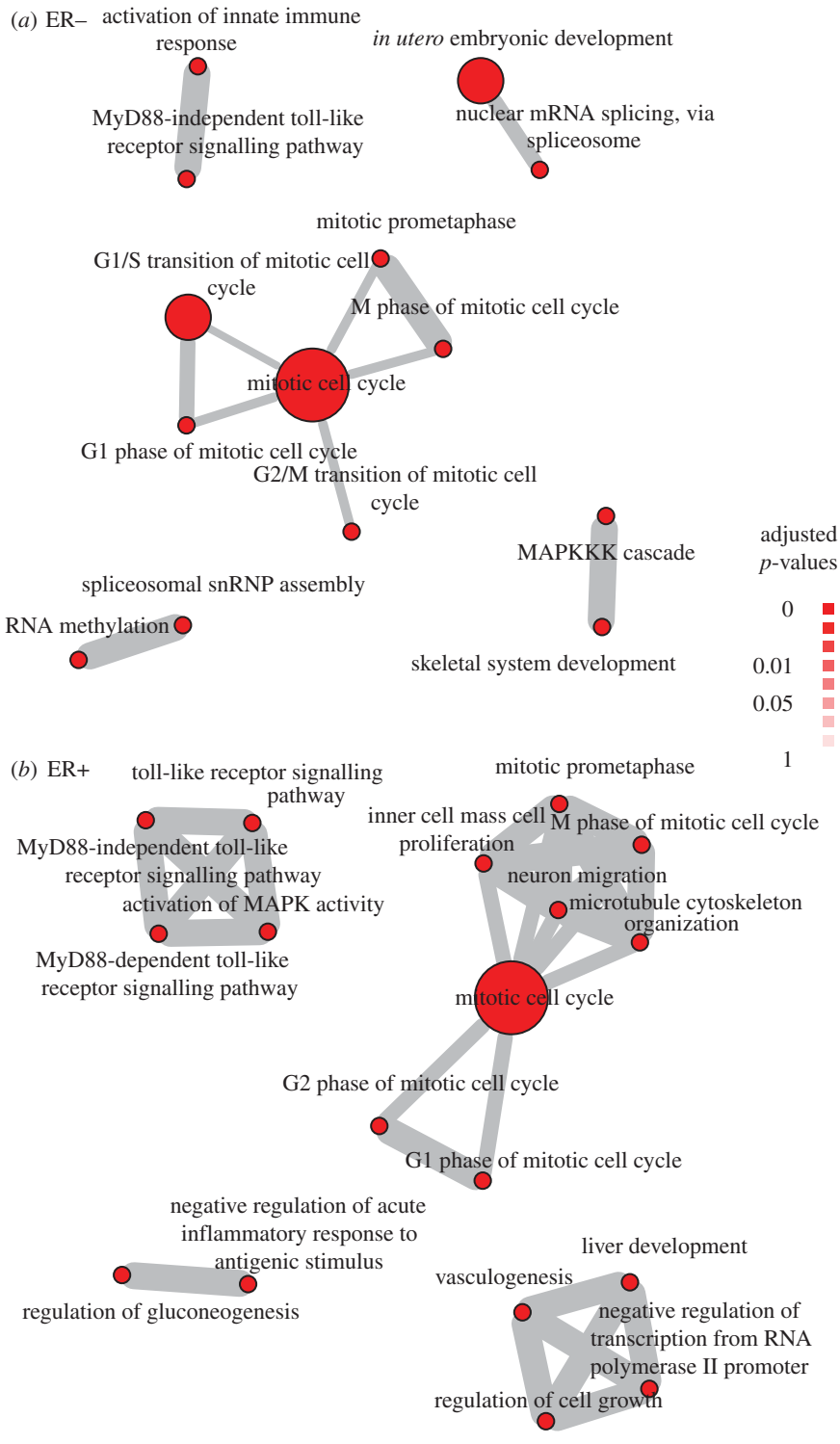

Figure 3. Enrichment maps for *chemoinsensitivity* as outcome for (*a*) ER— and (*b*) ER+ tumours. Shown are biological process that are statistically over-represented in the copy number features for the single data type *FusionGP* analysis, as determined by a hypergeometric test ($p < 1^{-4}$). The size of the red circles indicates the number of genes with a given pathway; the grey lines show where the same gene is shared across a pair of pathways, with line thickness indicating the number of genes. Enrichment maps were generated using the R package *HTSanalyzeR*. Singleton/unconnected nodes are not shown.

is the DNA repair processes enriched only in the copy number features. It is well known that BRCA1, a key mediator of the DNA repair pathways, remain one of the most important genes for breast cancer. Specifically for ER+ patients, loss of BRCA1-mediated transcriptional activation of ER expression can lead to increased resistance to ER antagonists [29]. Thus, the presence of genomic features involved in

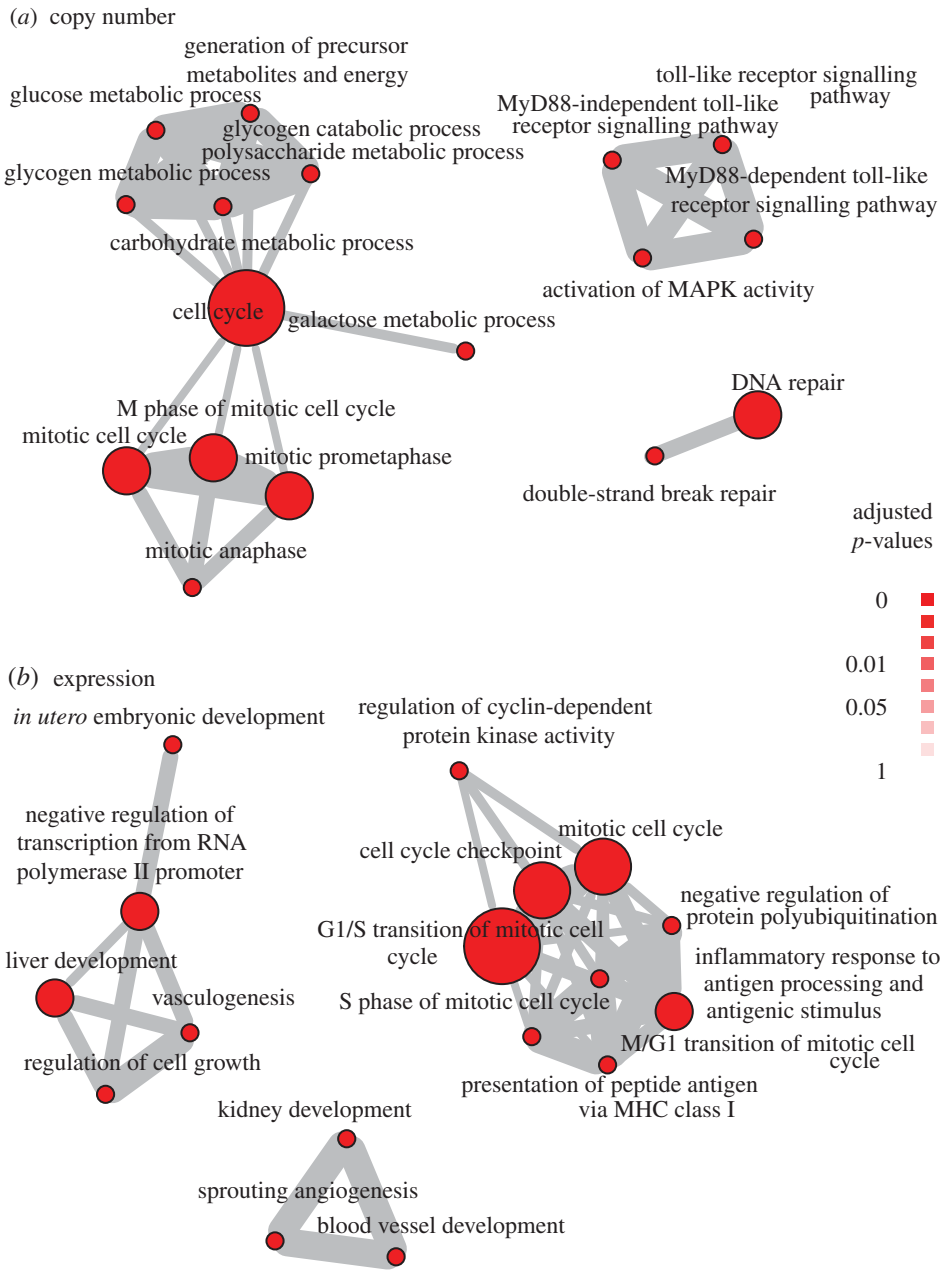

Figure 4. Enrichment maps for *deaths* as outcome for (*a*) copy number and (*b*) expression features selected for ER+ tumours. Shown are biological process that are statistically over-represented in the corresponding feature types for the single data type *FusionGP* analysis, as determined by a hypergeometric test ($p < 1^{-4}$).

DNA repair when predicting deaths in ER+ patients support the accuracy of our algorithm in searching for biologically meaningful features.

Gene expression features useful for predicting ER+ deaths include those involved in angiogenesis (figure 4*b*). This indicates that microenvironmental changes such as angiogenesis, which can be captured by gene expression data generated from our whole-tumour material, can be found in patients with poor prognosis. Angiogenesis is a known hallmark of poor prognosis in breast cancer as supported by multiple lines of evidences [30]. Meanwhile, ER is known to be capable of mediating angiogenesis, but its mechanism remains elusive Losordo & Isner [31], and our result here not only supports this but also reveals specific genes from our unsupervised analysis that may help define the regulatory mechanism. Taken together, results from our proposed model are supported by current knowledge of important genomic aberrations and microenvironmental features implied in insensitivity to chemotherapy and poor prognosis, in addition to identifying new features that can potentially help define mechanisms of these critical events.

5. Conclusion

We have presented *FusionGP*, a non-parametric Bayesian method for combining multiple data types to predict binary clinical outcomes. *FusionGP* generalizes the notion of feature selection for biomarker discovery, allowing for simultaneous, sparse feature selection across multiple heterogeneous data types.

Results on synthetic data show that *FusionGP* is effective at making predictions using noisy, sparse data and that it can identify the informative features. Combining two synthetic data types leads to superior predictive results, and also that *FusionGP* outperforms the RFs classifier, stepwise logistic regression and also a simple SVM. We suggest that this is because these standard methods do not explicitly account for the sparse nature of the synthetic data.

We present a range of analyses of the METABRIC breast cancer dataset, including gene expression, copy number alterations and H&E image-derived morphological feature data, and considering as outcomes *death* and *chemoinsensitivity*. We note that because the results are for a single collection of datasets, and using cross-validation rather than an independent test set, some caution in the interpretation of these results is warranted. With this in mind, from these analyses we draw a number of conclusions.

— *FusionGP* consistently outperforms both stepwise GLM and SVM methods, and marginally outperforms RF for the METABRIC data types.
— For prediction of disease-specific death, the molecular data are highly informative for the ER+ tumours, whereas for the ER− tumours the image features significantly outperform the molecular data.
— The enriched biological processes for the ER+ cases highlight specific genomic alterations in the DNA repair pathway, which is indicative of DNA repair-mediated ER expression activation and imply an influence on response to ER treatment.
— For prediction of chemoinsensitivity, ER− and ER+ tumours share common cell-cycle-related processes in the feature selected, yet differs in processes signifying the role of immune infiltration in the microenvironment in response to chemotherapy in ER− tumours.
— We therefore conclude that it is important to account for both the ER subtypes and also common underlying structure that is shared by all tumour samples.

We note that in practical terms, complex methods such as *FusionGP* have much longer run times than many of the equivalent standard methods. For example, if one is interested mainly in making predictions using the METABRIC data, one can achieve only marginally worse results using the much-quicker RF algorithm. *FusionGP* does also give a useful ability to better explain the biology underlying the data, as we have done in this paper. Nevertheless, we note that improvements to the run time of *FusionGP* would greatly improve its utility.

Given these results, we believe a number of interesting future directions present themselves. The image features, and to some extent the molecular features, contain significant redundancy. Developing models that learn a new set of (uncorrelated) latent features, rather than simply selecting the existing features, may allow us to refine our biological insights. It would also be interesting to encode more complex structures between the different data types—for example, we know that copy number alterations will effect gene expression patterns. Previous experience suggests that modelling such structure can add significant complexity to the modelling process, but it is certainly worthy of further investigation. Finally, fast approximations for the *FusionGP* algorithm will become increasingly valuable as multi-type cancer datasets becoming increasingly large.

We live in an exciting time for cancer research, where a range of new measurement technologies are giving us unprecedented access to the inner workings of tumours and cancer cells. Developing new methods to better use this data is vital if we are to make the most of this wave of new data.

Data accessibility. All data have been deposited at the European Genome-Phenome Archive hosted by the European Bioinformatics Institute, under accession number EGAS00000000083. They can also be accessed upon request via the METABRIC data access committee (metabric@cruk.cam.ac.uk). The *FusionGP* MATLAB code and synthetic data can be downloaded from https://sites.google.com/site/fusiongppaper/.

Authors' contributions. R.S.S. developed the statistical model and implemented it in code; Y.Y. developed and applied the input feature sets for the METABRIC data; both authors conceived of the study, analysed the results, drafted the manuscript and gave final approval for publication.

Competing interests. We declare we have no competing interests.

Funding. R.S.S. acknowledges support from an MRC Biostatistics Fellowship. Y.Y. acknowledges support from the Institute of Cancer Research.

References

1. Gray J, Druker B. 2012 Genomics: the breast cancer landscape. *Nature* **486**, 328–329. (doi:10.1038/486328a)
2. Lander E. 2011 Initial impact of the sequencing of the human genome. *Nature* **470**, 187–197. (doi:10.1038/nature09792)
3. Paik S *et al.* 2004 A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826. (doi:10.1056/NEJMoa041588)
4. Van't Veer L *et al.* 2002 Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536. (doi:10.1038/415530a)
5. Yuan Y *et al.* 2012 Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci. Transl. Med.* **4**, 161er6. (doi:10.1126/scitranslmed.3005298)
6. Futschik M, Sullivan M, Reeve A, Kasabov N. 2003 Prediction of clinical behaviour and treatment for cancers. *Appl. Bioinform.* **2**, 53–58.
7. Nevins JR, Huang ES, Dressman H, Pittman J, Huang AT, West M. 2003 Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Hum. Mol. Genet.* **12**, R153–R157. (doi:10.1093/hmg/ddg287)
8. Pittman J *et al.* 2004 Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proc. Natl Acad. Sci. USA* **101**, 8431–8436. (doi:10.1073/pnas.0401736101)
9. Stephenson AJ, Smith A, Kattan MW, Satagopan J, Reuter VE, Scardino PT, Gerald WL. 2005 Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. *Cancer* **104**, 290–298. (doi:10.1002/cncr.21157)
10. Gevaert O, Smet FD, Timmerman D, Moreau Y, Moor BD. 2006 Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics* **22**, e184–e190. (doi:10.1093/bioinformatics/btl230)
11. Boulesteix A, Porzelius C, Daumer M. 2008 Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics* **24**, 1698–1706. (doi:10.1093/bioinformatics/btn262)
12. Daemen A, Gevaert O, Ojeda F, Debucquoy A, Suykens JAK, Sempoux C, Machiels J-P, Haustermans K, De Moor B. 2009 A kernel-based integration of genome-wide data for clinical decision support. *Genome Med.* **1**, 39. (doi:10.1186/gm39)
13. Obulkasim A, Meijer G, van de Wiel M. 2011 Stepwise classification of cancer samples using clinical and molecular data. *BMC Bioinform.* **12**, 422. (doi:10.1186/1471-2105-12-422)
14. Savage RS, Ghahramani Z, Griffin JE, de la Cruz BJ, Wild DL. 2010 Discovering transcriptional modules by Bayesian data integration. *Bioinformatics* **26**, i158–i167. (doi:10.1093/bioinformatics/btq210)
15. Yuan Y, Savage RS, Markowetz F. 2011 Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput. Biol.* **7**, e1002227. (doi:10.1371/journal.pcbi.1002227)
16. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. 2012 Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* **28**, 3290–3297. (doi:10.1093/bioinformatics/bts595)
17. Rasmussen CE, Williams CKI 2006 *Gaussian processes for machine learning*. Cambridge, MA: The MIT Press.
18. Chu W, Ghahramani Z, Falciani F, Wild DL. 2005 Biomarker discovery in microarray gene expression data with Gaussian processes. *Bioinformatics* **21**, 3385–3393. (doi:10.1093/bioinformatics/bti526)
19. Kirk P, Stumpf M. 2009 Gaussian process regression bootstrapping: exploring the effects of uncertainty in time course data. *Bioinformatics* **25**, 1300–1306. (doi:10.1093/bioinformatics/btp139)
20. Stegle O, Denby KJ, Cooke EJ, Wild DL, Ghahramani Z, Borgwardt KM. 2010 A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *J. Comput. Biol.* **17**, 355–367. (doi:10.1089/cmb.2009.0175)
21. Rasmussen C, Nickisch H. 2010 Gaussian processes for machine learning (gpml) toolbox. *J. Mach. Learn. Res.* **11**, 3011–3015.
22. Gelman A 2004 *Bayesian data analysis*. Boca Raton, FL: CRC Press.
23. Girolami M, Zhong M. 2007 Data integration for classification problems employing Gaussian process priors. In *Advances in Neural Information Processing Systems. Proc. of the 2006 Conf.*, vol. 19. Cambridge, MA: MIT Press.
24. Murray I, Adams RP. 2010 Slice sampling covariance hyperparameters of latent Gaussian models. *Adv. Neural Inform. Process. Syst.* **23**, 1732–1740.
25. Curtis C *et al.* 2012 The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352.
26. Breiman L. 2001 Random Forests. *Mach. Learn.* **45**, 5–32. (doi:10.1023/A:1010933404324)
27. Calabro A *et al.* 2009 Effects of infiltrating lymphocytes and estrogen receptor on gene expression and prognosis in breast cancer. *Breast Cancer Res. Treat.* **116**, 69–77. (doi:10.1007/s10549-008-0105-3)
28. Loi S *et al.* 2013 Prognostic and predictive value of tumor-infiltrating lymphocytes in a phase iii randomized adjuvant breast cancer trial in node-+ breast cancer comparing the addition of docetaxel to doxorubicin with doxorubicin-based chemotherapy: big 02-98. *J. Clin. Oncol.* **31**, 860–867. (doi:10.1200/JCO.2011.41.0902)
29. Ratanaphan A. 2012 A DNA repair brca1 estrogen receptor and targeted therapy in breast cancer. *Int. J. Mol. Sci.* **13**, 14 898–14 916. (doi:10.3390/ijms131114898)
30. Gujam FJA, Going JJ, Edwards J, Mohammed ZMA, McMillan DC. 2014 The role of lymphatic and blood vessel invasion in predicting survival and methods of detection in patients with primary operable breast cancer. *Crit. Rev. Oncol./Hematol.* **89**, 231–241. (doi:10.1016/j.critrevonc.2013.08.014)
31. Losordo D, Isner J. 2001 Estrogen and angiogenesis a review. *Arterioscler. Thromb. Vasc. Biol.* **21**, 6–12. (doi:10.1161/01.ATV.21.1.6)