Increased Proportion of Variance Explained and Prediction Accuracy of Survival of Breast Cancer Patients with use of Whole-Genome Multi-Omic Profiles

Al. Vazquez1*, Y. Veturi2, M. Behring2,4, S. Shrestha3, M. Kirst5,6, M.F.R. Resende Jr.5,6, G. de los Campos1,7

1Department of Epidemiology and Biostatistics, Michigan State University, 909 Fee Hall Room B601, East Lansing, MI, 48824, USA;  
2Biostatistics Department, University of Alabama at Birmingham, AL, USA;  
3Department of Epidemiology, University of Alabama at Birmingham;  
4Comprehensive Cancer Center, University of Alabama at Birmingham;  
5School of Forest Resources and Conservation, University of Florida, P.O. Box 110410, Gainesville, FL 32611, USA;  
6University of Florida Genetics Institute, University of Florida;  
7Statistics Department, Michigan State University, MI, USA.  
*Corresponding author: avazquez@msu.edu

Keywords: Prediction of complex traits, diseases risk, omics integration.

ABSTRACT

Whole-genome multi-omic profiles hold valuable information for the analysis and prediction of disease risk and progression. However, integrating high-dimensional multi-layer omic data into risk assessment models is statistically and computationally challenging. We describe a statistical framework (BGAM=Bayesian Generalized Additive Model) and present software for integrating multi-layer high dimensional inputs into risk assessment models. We used BGAM and data from The Cancer Genome Atlas for the analysis and prediction of survival after diagnosis of breast cancer. We developed a sequence of studies to (i) compare predictions based on single omics with that based on clinical covariates commonly used for the assessment of BC patients (COV), (ii) evaluate the benefits of combining COV and omics, (iii) compare models based on (a) COV and gene expression profiles from oncogenes with (b) COV and whole-genome gene expression (WGGE) profiles, and (iv) evaluate the impacts of combining multiple omics and their interactions. We report that: (i) WGGE and whole-genome methylation profiles (METH) offer more predictive power than any of the COV commonly used in clinical practice, (e.g., subtype and stage), (ii) adding WGGE or METH profiles to COV increases prediction accuracy, (iii) the predictive power of WGGE is considerably higher than that based on expression from large-effect oncogenes, and (iv) the gain in prediction accuracy when combining multiple omics is consistent. Our results show the feasibility of omic integration, highlight the importance of WGGE and METH profiles in breast cancer achieving gains of up to 7 points AUC over the COV in some cases.

INTRODUCTION

The continued development of high-throughput genomic technologies has fundamentally changed the genetic analyses of complex traits and diseases. These technologies provide large volumes of data from multiple ‘omic’ layers, including the genome (e.g., SNP, copy-number variants, mutations), epigenome (e.g., methylation), transcriptome (e.g., RNA-seq), proteome, etc. This information can be
used to develop models for understanding and prediction of disease risk and disease prognosis. Recently, several studies have uncovered unprecedented numbers of omic-factors associated to disease risk and progression. For instance, in the last decade Genome-Wide Association Studies (GWAS) have reported large numbers of SNPs (e.g., http://www.genome.gov/gwastudies/) and structural variants (e.g., CNVs(Beroukhim et al. 2010; Morrow 2010)) associated with disease risk. Likewise, several studies have reported methylation sites(Dedeurwaerder et al. 2011; Fackler et al. 2011; Fang et al. 2011) and genes with expression profiles associated to prognosis (Perou et al. 2000; Sørlie et al. 2001; Van’t Veer et al. 2002; Sotiriou and Pusztai 2009; Gy\Horffy et al. 2016). However, despite of the tremendous progress achieved, use of this information in clinical practice remains limited in part because the proportion of variance in disease risk or prognosis explained by the individual factors identified remains still limited.

Data integration can be an avenue for improving our understanding and our ability to predict disease risk and prognosis. Integration can take place by combining information from multiple sites across the genome as well as by integrating inputs from different omics. In prediction of complex traits and disease risk several studies (e.g., (de los Campos et al. 2010c; Makowsky et al. 2011; Vazquez et al. 2012) (Purcell et al. 2009; Yang et al. 2010 ) have demonstrated that the proportion of variance explained with use of whole-DNA profiles is considerably higher than the one achieved by models that use a limited number GWAS-significant variants. Likewise, several studies have demonstrated benefits of integrating data from multiple omics. For example, Chen and co-authors (Chen et al. 2012) demonstrated how integrated omic profiles can provide insights into the development of Type 2 diabetes. However, our ability to integrate whole-genome multi-layer omic data into risk assessment still lags behind.

Wheeler et al. (2014) and Vazquez et al. (2014) proposed using what Wheeler called “Omic Kriging” for prediction of complex traits and disease risk using multi-omic profiles. Kriging is a kernel smoothing technique commonly used in spatial statistics (e.g., Cressie 2015). From a statistical perspective Kriging is the ‘Best Linear Unbiased Predictor’ (BLUP) method commonly used in quantitative genetics (Henderson 1950; Robinson 1991) using pedigree (Henderson 1950, 1975) or DNA information (G-BLUP; VanRad 2008). Omic Kriging is a multi-kernel method (de los Campos et al. 2010a; b) where the resulting kernel is a weighted average of similarity matrices derived from different omics.

Although Omic Kriging represents a promising method for integrating multi-omic data, the method has potentially important limitations. Firstly, the approach assumes that the architecture of effects is homogeneous across omic layers. This assumption may not hold if some omics have a sparse architecture of effect (i.e., a few factors have sizable effects and the rest have no effect) and other omics have non-sparse effects architecture (i.e., all inputs have small effects). Secondly, Omic Kriging assumes implicitly that omics act in an additive manner (i.e., there are no interactions between omics). This may fail for instance if the effects of one layer (e.g., SNP) are modulated by a second layer (e.g., methylation).

In this study we describe a modeling framework that: (i) allows integrating high-dimension inputs from multiple omic layers, (ii) contemplate different effect architecture across layers and (iii) incorporates interactions between omics. The approach is a Bayesian Generalized Additive Model (BGAM) that integrates in a unified setting ideas from Generalized Additive Models (Hastie and Tibshirani, 1986), with Bayesian methods that allow for different architectures of effects (including estimation with or without shrinkage, and variable selection methods) and recently developed techniques
for modeling interactions between high dimensional inputs (Jarquín et al. 2014). Importantly, the
BGAM can be used with traditional quantitative traits, time-event (subject to censoring), ordinal and
binary (e.g., disease) outcomes.

We use BGAM and data from TCGA to develop models for analysis and prediction of breast cancer
(BC) outcomes. Breast cancer is considered one of the most lethal types of cancer (Boyle and Levin
2008). In the United States alone there are approximately 180,000 new cases of BC each year (Eifel et
al. 2000), and it has been estimated that about 12% of women will develop BC over their lifetime (Eifel
et al. 2000; Smigal et al. 2006). Advances in early detection and in adjuvant therapy have reduced
mortality due to BC. However, adjuvant therapy has important undesirable side effects on treated
patients. Some of the most serious ones include permanent infertility, heart damage, cognitive
impairment, and increased probability of developing other types of cancers (Eifel et al. 2000). Cancers
in approximately 40% of the BC patients are estimated to recur or metastasize (Weigelt et al. 2005).
However, because current models cannot accurately predict BC progression, approximately 80% of BC
patients are treated with adjuvant therapy. Thus, there are a substantial proportion of BC patients who
are being unnecessarily treated with adjuvant therapy. An accurate assessment of disease progression
could be used to implement a more precise approach for the treatment of BC patients and reduce the
impact of undesirable outcomes due to the therapy. Here we apply BGAM modeling framework to data
from the TCGA to develop models for prediction of the probability of survival after diagnostic of BC. In
Our application we compare multi-omic models with risk-assessments based on clinical covariates and
the expression profiles of large-effect genes included in the Oncotype DX (Genomic Health; Paik et al.
2004a, 2006) which is an FDA-approved platform used in clinical practice to predict BC progression.
Our analysis demonstrates that the integration of whole-omic profiles can increase the proportion of
inter-individual differences in survival and enhance prediction accuracy of BC outcomes above and
beyond of what can be achieved using clinical covariates (race, age, cancer subtype and stage) and
expression-based diagnostic tools (e.g., Oncotype DX).

In the present article we outline the main elements of BGAM modeling framework and present a
series of case studies in which we apply the methods to breast cancer cases from TCGA. The final
discussion section highlights the main findings of our study and offers a brief perspective on the
strengths and limitations of the BGAM framework. Our results show how the integration of omics in a
clinical model improves prediction accuracy for most Omics, but the improvements are higher by
combining clinical information with whole-genome methylation and gene expression profiles.

MODELING FRAMEWORK

Assume that the multi-layer omic data consists of a phenotype or disease outcome \( y_i \) \((i = 1, \ldots, n)\)
and sets of predictors coming from \( L \) input layers; these layers may include demographics, clinical
covariates and data from several omics. We denote the data from these layers as \( X = \{X_1, \ldots, X_L\} \). Here
\( X_l = \{x_{lj}\} \) denotes a set of predictors from the \( l^{th} \) data-layer and \( l=1, \ldots, L, i=1,\ldots,n \) and \( j=1,\ldots, p_l \) index
input layers \( (l) \), individuals \( (i) \) and predictors within an input layer \( (j) \), respectively.

Generalized Additive Model (GAM). Multi-layer inputs can be incorporated into a regression model
using the so-called Generalized Additive Model (GAM) framework (Hastie and Tibshirani 1986). In a
GAM a regression function is expressed as the sum of \( L \) smooth functions,

\[
\eta_i = f_1(x_{1i}, \alpha_1) + f_2(x_{2i}, \alpha_2) + \cdots + f_L(x_{Li}, \alpha_1) \tag{1}
\]
Each of these functions can be linear or non-linear for the inputs and can be specified parametrically or using semi-parametric methods (e.g., splines). Typically, these functions are indexed by a set of parameters \( (\alpha_t) \) that are estimated from data. When these parameters are high-dimensional (i.e., \( p_t \) is large), estimation is typically carried out using L2-penalized (i.e., Ridge-Regression) estimators (Hastie and Tibshirani 1986); this approach renders smooth functions with shrunken parameter estimates. The extent of shrinkage of estimates is controlled by regularization parameters. When there is only smooth function, an optimal value for the regularization parameter can be chosen using cross-validation methods (e.g., Golub et al. 1979). However, when there are multiple regularization parameters (e.g., one per term of the linear predictor), the cross-validation approach becomes infeasible and other approaches (e.g., mixed effects models or Bayesian methods) are needed.

For some high dimensional inputs (e.g., DNA markers, transcriptome) variable selection, as opposed to shrinkage, may be desirable. This can be achieved in penalized regressions by using penalties other than those based on the L2 norm e.g., with the L1 norm, as in the LASSO method, (Tibshirani 1996). Alternatively, variable selection and/or shrinkage can be obtained in a Bayesian setting by choosing particular types of prior distributions. The Bayesian approach has several attractive features. Firstly, within a Bayesian framework, multiple regularization parameters can be estimated from data without need of conducting extensive cross-validations. Secondly, Bayesian models can accommodate both shrinkage and variable selection in a unified framework. Finally, using methods described later, within the Bayesian framework one can accommodate interactions between inputs in high-dimensional sets. Therefore, in this study we adopt a Bayesian Generalized Additive Model (BGAM) framework for integrating multi-omic inputs.

**Bayesian Generalized Additive Model (BGAM).** For ease of presentation, we introduce the model for the case of a Gaussian outcome and assume that each of the functions entering in (1) are linear on their inputs. Cases involving non-Gaussian outcomes or functions that are non-linear on inputs are considered later on. For the purpose of illustration we consider only three input layers including a set of non-genetic covariates, \( X_1_i = \{x_{1ij}\}^{j=p_{1}}_{j=1} \) and two omics: \( X_2_i = \{x_{2ij}\}^{j=p_{2}}_{j=1} \) and, \( X_3_i = \{x_{3ij}\}^{j=p_{3}}_{j=1} \).

Extensions to more than three layers are straightforward. With this setting, the linear predictor becomes

\[
\eta_i = \mu + \sum_{j=1}^{p_{1}} x_{1ij} \alpha_{1j} + \sum_{j=1}^{p_{2}} x_{2ij} \alpha_{2j} + \sum_{j=1}^{p_{3}} x_{3ij} \alpha_{3j} \quad (2)
\]

above \( \alpha_{1} = \{\alpha_{1j}\}^{j=p_{1}}_{j=1} \), \( \alpha_{2} = \{\alpha_{2j}\}^{j=p_{2}}_{j=1} \) and \( \alpha_{3} = \{\alpha_{3j}\}^{j=p_{3}}_{j=1} \) are regression coefficients.

**Bayesian Likelihood.** Under Gaussian assumptions, the conditional distribution of the outcome given the parameters of the linear predictors is

\[
p(\gamma|X_1, X_2, X_3, \theta) = \prod_{i=1}^{n} \frac{\exp\left\{\frac{(y_i - \eta_i)^2}{2\sigma^2}\right\}}{2\pi\sigma^2} \quad (3)
\]

where \( \theta = \{\sigma^2, \mu, \alpha_1, \alpha_2, \alpha_3\} \) is a vector of model unknowns.

**Prior distribution.** In a Bayesian setting layer-specific architectures of effects can be accommodated using layer-specific priors. Therefore, we structure the joint prior distribution of effects as follows:

\[
p(\alpha_1, \alpha_2, \alpha_3, \sigma^2, \Omega_1, \Omega_2, \Omega_3) \propto p(\sigma^2) \prod_{i=1}^{3} \left\{ \prod_{j=1}^{p_i} p(\alpha_{ij} | \Omega_i) \right\} p(\Omega_i)
\]

where \( p(\sigma^2) \) is a prior for the error variance (e.g., a scaled-inverse chi-square), \( p(\alpha_{ij} | \Omega_i) \) are IID priors assigned to the effect of the \( j^{th} \)-input layer, \( \Omega_i \) is a set of layer-specific regularization hyper-parameters and \( p(\Omega_i) \) is a prior distribution assigned to these hyper-parameters.
Special cases. Estimation without shrinkage can be obtained by setting $p(\alpha_{ij} | \Omega_t)$ to be a flat prior (e.g., a normal prior centered at zero and with a very large variance). Shrunken estimates can be obtained by setting $p(\alpha_{ij} | \Omega_t)$ to be a normal prior centered at zero and with variance parameter $\Omega_t = \sigma^2_{\alpha ij}$ treated as unknown. This approach renders estimates comparable to those of Ridge Regression (Hayes et al. 2001) with an extent of shrinkage that is similar across effects. Differential shrinkage of estimates of effects can be obtained using priors from the thick tailed family such as the Double-Exponential or scaled-t distributions; these priors are used in the Bayesian Lasso (Park and Casella 2008) and in BayesA (Hayes et al. 2001). Finally, variable selection can be achieved by setting $p(\alpha_{ij} | \Omega_t)$ to be a finite mixture with a point of mass (or a very sharp spike) at zero and a relatively flat slab (George and McCulloch 1993; Ishwaran and Rao 2005).

Functions that are non-linear inputs can be accommodated by first mapping the original inputs (e.g., $X_i$) into a set of basis functions $\Phi_i = \{ \phi_{i1}(X_i), \phi_{i2}(X_i), \ldots \}$ and then using the transformed inputs, $\phi_{ij}(X_i)$, as covariates in the regression. This can be done either in parametric settings (e.g., with polynomials) or with semi-parametric specifications (e.g., using splines or kernels).

Gaussian processes. When the coefficients entering a linear term are assigned IID normal priors, the resulting function can be viewed as a draw from a Gaussian process. For instance, if $\alpha_{ij} \sim N(0, \sigma^2_{\alpha ij})$ then, the function $f_i = \Phi_i \alpha_i$ follows a normal distribution with null mean and covariance matrix given by $K_t \sigma^2_{\alpha ij}$ where $K_t = \Phi_i \Phi_i'$ is a covariance structure computed using cross-products of the basis functions. This treatment fully connects the BGAM with Reproducing Kernel Hilbert Spaces regressions methods (RKHS, e.g., Wahba 1990; Shawe-Taylor and Cristianini 2004) a framework that can be used to implement various types of parametric and semi-parametric regressions. Importantly, this framework can be implemented with almost any input sets, including text data, images, special data, graphs, etc. (Wahba 1990; de los Campos et al. 2009, 2010a).

Interactions between input layers. The model of expressions (1) and (2) assumes that layers act additively. However, many applications may require modeling interactions between layers. Accommodating interactions can be particularly challenging when the number of inputs in the interacting layers is large. For instance, with 10,000 expression profiles and 10,000 SNPs, modeling all possible first-order interactions requires using 100 million contrasts. Dealing with interactions explicitly is not feasible. Therefore, we propose to deal with interactions implicitly using Gaussian processes with covariance structures based on the patterns induced by the so-called “reaction norm model”. This approach has been used for modeling interactions between genetic factors and environmental covariates in plants and animals (Gregorius and Namkoong 1986; Calus et al. 2002; Su et al. 2006; Jarquin et al. 2014). Recently, Jarquin et al. (2014) develop methods for reaction norms involving high dimensional genetic (e.g., SNP) and high-dimensional environmental inputs. The authors show that the covariance patterns induced by a reaction norm model can be expressed as the Schur (or Hadamard) product of kernels that evaluate input similarity at each of the interacting layers. An example of the use of this method is provided in the fourth case study of the following section.

Non-Gaussian outcomes (e.g., binary or ordered categorical) can be accommodated using the probit or logit link; in a Bayesian MCMC setting the probit link can be implemented easily using data augmentation (Albert and Chib 1993).

Software. All the models described above can be implemented using the BGLR R-package (Pérez and de los Campos 2014). This software implements BGAM for continuous, binary and ordinal outcomes and offers users the possibility of specifying at each of the layers parametric and semi-
parametric methods for shrinkage and variable selection. Further details about the software can be found in Pérez and de los Campos (2014) and in the following website: https://github.com/gdlc/BGLR-R/.

CASE STUDIES

In this section we investigate the association between patient survival and several predictors that can be assessed at diagnosis, including: information commonly used by clinicians to assess BC patients (hereinafter we refer to these as “clinical covariates”), gene expression profiles (RNA-seq), methylation, copy-number variant and micro-RNA. All these omics were assessed at the primary tumor. We consider several research questions, and for each of these questions we designed a case study that involves the comparison of several models, each of which is a special case of the BGAM framework described in the previous section. All the case studies are based on data from BC patients from TCGA. The motivation for each of the case studies is briefly presented next.

Case Study I. Clinical information such as tumor subtype or cancer stage is used to assess risk of possible cancer outcomes; precise prediction of outcomes improves the decision of which treatments options to use for each patient. Although the clinical covariates are predictive of the likelihood of disease progression, after accounting for differences attributable to these clinical predictors important inter-individual differences in the BC outcome remain. Gene expression has been demonstrated to be associated to BC progression (Sørlie et al. 2001; Sørlie et al. 2003). Therefore, in our first case study (CS-I) we assess the relative contribution to variance and to prediction accuracy of whole-genome gene expression profiles (WGGE). We compare models based on WGGE with others based on clinical covariates commonly used in clinical practice (breast cancer subtype, stage, age at the cancer diagnosis, histological subtype and race). In the study we assess the contribution to variance and to prediction accuracy of WGGE alone and in combination with clinical covariates. Sørlie et al. (2001) demonstrated that clusters derived from the gene expression profiles are associated to breast cancer subtypes. Our COV (M7) model and all other models that incorporate all clinical covariates already accounts for breast cancer subtypes as dummy variables and therefore incorporates clustering. Several studies have demonstrated the association of gene expression patterns and BC outcome. However, these studies are based on data, that has been conditioned by some dimension reduction method (e.g., clustering, principal components). We argue that the consideration of WGGE is essential in capturing the diverse information on this trait with complex biology.

Case Study II. Our first case study accounts for the main effects of commonly used clinical covariates and those of WGGE. However, the patterns of gene expression and the prognosis of the cancer present substantial variation in both the different cancer sub-types and the different stages of development of the disease. Therefore, in our second case study (CS-II) we focus on a particular cancer subtype: luminal types at early stage--this is the most prevalent subtype. For early stage luminal patients there is a well-established commercial gene-expression platform [Oncotype DX; Genomic Health Inc, Redwood City, CA (Genomic Health; Paik et al. 2004a, 2006)] that has been approved by the Food and Drug Administration for use as a diagnostic tool. Oncotype DX analysis is based on the profile of a genetic signature consisting of only a few genes. We argue that the use of whole-genome gene expression profiles can lead to a larger proportion of variance explained and higher prediction accuracy than the one that can be achieved using the expression profiles of a few genes. Therefore, in our CS-II we compare models based on: (i) clinical covariates, (ii) clinical covariates plus the expression profile of genes included in the Oncotype DX, and (iii) clinical covariates and WGGE. The models were fitted and
compared based on data from patients with luminal types at early stage only, lymph node negative and all lymph nodes.

Case Study III. Information from omics other than the transcriptome, such as DNA information (e.g., copy-number variants) or data from the epigenome can also contribute to inter-individual differences in survival. Therefore, in our third case study (CS-III) we consider the use of omics other than WGGE, including: micro-RNA (miRNA), methylation and copy-number variant (CNV). For each omic we assess the proportion of variance explained and prediction accuracy of the omic alone and in conjunction with clinical covariates. In all cases we consider one omic at a time and conduct separate analyses for each of the omics.

Case Study IV. In our previous CS we have assessed omics separately or in combination with COV. In our fourth case study (CS-IV) we evaluate the benefits of integrating two omics, WGGE and METH, and COV simultaneously; we explore this both with an additive model and with a specification that contemplates interactions between omics.

Data

The Cancer Genome Atlas (TCGA) offers data of BC patients with demographic, clinical, omic and follow-up information from where survival information can be derived. Since data is still being collected, follow-up time is short for most patients. Therefore, our response variable was defined as subjects that either died (1) or were alive (0) and had at least three years of follow up. All male records and females with incomplete follow-up or inconsistent clinical records (e.g., death shortly after diagnosis of BC in an early stage without any record of progression) were removed. Also, women with distant metastasis at the time of diagnosis, or patients with history of a previous cancer were removed. After edition of these samples data was reduced from over 1,000 to 797 samples, from which only 285 met the minimum follow up criteria. Thus, baseline data set comprised of 285 patients; these included subjects with concordant data that were either dead (n=60) or alive (n=225) and had a minimum follow up time of 3 years. Not all these patients had complete data for all the omics. Therefore, in some of the case studies we further narrowed the set of patients to those that have complete data for the inputs relevant to the specific analysis. The original dataset offered by TCGA was reduced to patients with at least three years follow-up, because follow-up is still too short [in the original TCGA data, the follow up time averages (± SD) 2.05(±1.14) years of last contact time for those still alive].

In case studies I, III and IV models were obtained by regressing alive status (0/1) on the inputs that follow. These inputs were selected based on their association with survival in preliminary analysis. Case study II is a more homogeneous population, and fewer covariables were used (see case study II section).

- **Demographics**, including age at diagnosis (mean±SD were 55.6±12.6), race/ethnicity (white Caucasian /African American).

- **Clinical information from the tumor**, including histologic type (whether the invasive tumor arose from lobular tissue [n=35] or from ductal breast tissue [n=251]), subtype classification based on the membrane receptors present in the tumor cell (Luminal A: 179, Luminal B: 24, Her2-Neu: 69, and triple negative: 13), and stage, as defined by the American Joint Committee on Cancer (Edge et al. 2010) (from I to IV; the number of patients per stage were 58, 159 and 68 in stages I, II and higher, respectively).

- **Omics data included** gene-expression profiles from RNA-seq, whole genome methylation, micro RNA, and copy-number variants (CNV). Gene expression profiles were assessed using RNA-seq technology sequenced on an Illumina HiSeq 2,000 platform. Normalized expression counts per-gene
were used. Workflows for the creation of level 3 RNA data are previously detailed (Li et al. 2010; Wang et al. 2010). Copy number variant data were derived from Affymetrix Genome-Wide SNP array 6.0. Mean log2 ratios were used as a measure of per-segment CNV. Full processing details are documented in a Broad Institute GenePattern pipeline (“GenePattern”). Source data for methylation were generated with the Illumina Infinium HumanMethylation450 BeadChip and were processed by the Johns Hopkins GSC to derive beta values for CpG sites and their association with gene regions using methylumi (Pidsley et al. 2013). Micro RNA values are quantified as reads per million (RPM) from the Illumina HiSeq miRNA 2,500 platform. Short sequence reads were aligned to RCh37-lite reference genome using Burrows-Wheeler Alignment tool (BWA; Li and Durbin 2009), and normalized as RPMs (Network 2012). In TCGA samples were randomly assigned to plates; therefore, there should not be association between batch and survival outcomes. However, to confirm this we conducted analyses of dispersion due to batch, (see supplementary File S1 Table S1.1).

Data analysis

Each of the case studies includes a baseline model plus extensions obtained by including different combinations of omics. In all cases, the response (survival Yes/No) was regressed on predictors using a threshold model (Gianola and Foulley 1983; Agresti 2012) as implemented in the BGLR R-package (Pérez and de los Campos 2014). In each study, models were first fitted to all the individuals that had complete data for the set of predictors used in the case study. From this analysis, we report parameter estimates (e.g., variance components) and the posterior means of the log likelihoods.

Model specification. The effects of clinical covariates were regarded as fixed, while the effects of different omics were regarded as random. For simplicity all random effects were assumed to be IID Gaussian, with omic-specific variance parameters. We also conducted analyses using priors that induce variable selection. In other studies, these models have not shown strong differences in risk to disease (Vazquez et al. 2015). Results of these analyses are given in the Supplementary File S1, Table S1.2. Variance parameters were assigned scaled-inverse chi-square priors with 5 degrees of freedom (this gives a weakly informative prior) and scale parameters computed according to the rules described in Perez and de los Campos (2014); this is the default treatment of variances implemented in BGLR. For each model, we run 500,000 iterations of a Gibbs sampler; the first 20,000 samples were discarded as burn in and the remaining samples were thinned at a thinning interval of 5 (see Supplementary File S1, Figures S1.1 and S1.2 and Table S1.3). For all case studies we report the log likelihood, effective number of parameters in the model and the deviance information criteria.

Prediction accuracy was assessed using cross-validations (CV). We implemented a total of 200 independently generated 10-fold CV. Prediction accuracy was assessed using the CV-Area Under the Receiver-Operating Characteristic Curve (AUC, e.g., Fawcett 2006). Therefore, for each study and model we had a total of 200 estimates of AUC. Models were compared based on the average AUC and also by counting the proportion of CV (out of 200) for which a given model had higher AUC than another. For CV analyses models were fitted using 80,000 iterations collected after discarding the first 15,000 samples; furthermore samples were thinned at an interval of 5. For all case studies we report the average and the SD (across 200 CV) of the AUC and the proportion of times that a model had an AUC greater than other models, also computed using results from 200 CV. Code to implement models here described are provided in Supplementary File S2 and in the website: https://github.com/anainesvs/VAZQUEZ_etal_GENETICS_2016.
**Data availability**

The data used in this study is publically available, collected and distributed by The Cancer Genome Atlas (TCGA), NIH/NCI project. Data can be obtained at: https://tcga-data.nci.nih.gov/tcga/. Additionally, to ensure reproducibility of our analysis we provide the lines of code used to execute this study in supplementary file 2, and at our abovementioned github repository.

**Case Study I: Integrating clinical covariates and whole-genome gene expression**

The first case study (CS-I) was designed to assess the marginal association between survival and individual risk factors composed of clinical covariates (e.g., age, race, etc.), and to quantify the gains in prediction accuracy that can be achieved by adding gene-expression data on a model that accounts for the clinical information. Six sets of risk factors were considered; these include two demographics (age and race), three clinical features of the cancer (whether it is a lobular carcinoma, cancer subtype, and pathological stage) and gene expression profiles (RNA-seq) from the primary tumor.

**Sequence of models.** A total of eight models were fitted, including six single-risk-factor models (labeled as M1-M6), a model based on all predictors except gene expression (M7 also labeled as COV), and a model that included all the available predictors (M8, labeled as COV+WGGE).

**Results.** Table 1 provides goodness of fit statistics, measures of model complexity and estimates of prediction accuracy for each of the 8 models fitted in CS-I. Among the single-factor models, the one that fitted the data best and had the highest CV-AUC was the model using whole genome gene expression (WGGE, M6); clearly, WGGE was the most informative input.

The comparison of the results obtained with models COV and COV+WGGE indicate that information from WGGE profiles can improve the assessment of survival, even after accounting for the predictors commonly considered in clinical practice. The increase in CV-AUC obtained when WGGE was added on a model that includes all COV was in average 1.7 points (COV+WGGE) higher than that of the model based on COV, and the comparison across 200 CV shows that the model COV+WGGE outperformed the model based on clinical covariates (COV) 99% of the time. In other words, the increase in prediction accuracy was consistent.

**Case Study II: Genetic Signatures vs Whole Genome Gene Expression Profiles within Cancer Subtypes**

Case Study I showed that the assessment of BC survival could be improved by using WGGE profiles from the tumor tissue. CS-I is an analysis not specific to a cancer subtype, although they are considered in the model. The clinical value of WGGE for BC has been demonstrated before (Sørlie et al. 2001), and gene expression profiles from oncogenes are often used to assess BC patients; an example of this is the Oncotype DX platform (Genomic Health; Paik et al. 2004a, 2006), which is based on the expression profiles of 21 genes. Oncotype DX has been validated for to assess BC outcome among patients affected by tumors of the luminal (ER+) cancer subtype which are in early stage of the disease and do not have distant or nodal metastasis. Therefore in this CS we focus only in Luminal cancers, and compare the relative contribution to variance and to prediction accuracy of the expression profile of the Oncotype DX with that of whole-genome gene expression profiles.
Data are comprised of a subset of the patients (N=186) used in CS-I that qualify for the Oncotype-DX; these are patients who had estrogen receptor positive tumors at stages I or II. Results are presented for all early stage Luminal patients (ER+ or PR+), and only for early stage Luminal patients with negative lymph nodes (the target population of the Oncotype-DX). The platform includes 21 genes, 16 “risk” genes and five reference (“housekeeping”) genes for the purpose of normalizing the data (Paik et al. 2004a). From RNA-seq WGGE profiles we retrieve the expression profiles from all the risk genes, except RPLP0.

Sequence of models. The baseline model (COV) included age at diagnosis, race and ethnicity. Tumor subtype and stage were not included as covariates of the baseline models because all patients had luminal tumors in early stage. The baseline model was first extended by adding the random effects of the expression of the genes included in the Oncotype-DX panel (COV+ONCO). Subsequently, we extended the COV+ONCO model by adding the random effects of 17,899 genes not included in the Oncotype panel (we label this model as COV+WGGE, standing for covariates plus Whole-Genome Gene Expression). The effects of race and age were treated as fixed, and those of the gene expression profiles of the genes included in either COV+ONCO or COV+WGGE were assigned iid normal priors with null mean and unknown variance (variances were assigned scaled-inverse chi-square priors).

TABLE 2, ABOUT HERE

Results from CS-II are given in Table 2. In this study we report AUC based on luminal types, all luminals and only the ones with lymph node negatives. Estimates of variance components revealed that contribution to variance of the expression of the gene in the Oncotype DX was low (.027). On the other hand, the use of WGGE lead to a sizable fraction of variance explained. Indeed, the estimated variance component associated to WGGE was 0.439 amounts to 30% of the variance in risk that is not explained by COV (computed as 0.439/1.439). However, due to small sample size, the posterior credibility region for the estimate variance component associated to WGGE was wide. The DIC (“smaller is better”) also suggests that the best model was the one including COV and WGGE. And the cross-validation analyses based on all luminal cases revealed that adding information from the genes included in the Oncotype-DX (COV+ONCO) improved AUC relative to the baseline model in 2.7 units, and that adding WGGE increased AUC (also relative to COV) by 6.5 units. The analyses based on patients with lymph node negative also revealed a sizable increase in prediction AUC when using WGGE (compared to the baseline model, the model using COV and WGGE had 6.6 point in AUC and COV+WGGE outperformed COV in 99% of the 200 cross-validations). However, for here the prediction AUC of the model using COV+ONCO was similar to the one obtained with COV only. Therefore, we conclude that using WGGE lead to a higher proportion of variance of risk explained and higher prediction accuracy than the one that can be achieved using the expression profiles of a few genes.

Case Study III: Comparison between omics

In the two previous studies we assessed the performance of models based on clinical covariates and gene-expression information from the tumor cells. In this study we compare the relative performance of models based on the other “omics” available: (i) copy-number variants (CNV), (ii) methylation (METH), and (iii) micro-RNA.

Data. This study includes data from patients that had information for at least one of the omics considered. Figure 1 shows a Venn diagram(Oliveros 2007) representing the number of patients with omic data by layer. The number of individuals with complete omic data is relatively small (n=127). Therefore, when fitting models for a given omic, we used all the individuals that had information for
that omic. These lead to three different sets of patients (we labeled them as sets 1-3, corresponding to individuals with CNV, METH and miRNA, respectively) to which models were fitted.

**FIGURE 1, ABOUT HERE**

**Sequence of models.** For each set of patients, we compared the performance of a model based on covariates only (COV) with that of a model based on covariates only and one with covariates plus data from the corresponding omic (COV+CNV, COV+METH, COV+miRNA). As before, models were fitted using the BGLR R-package with COV as fixed effects and omics as random effects, where the effects of the omics were treated as IID drawn from a normal distribution with null mean and unknown variance.

**Results.** Table 3 shows estimates of goodness of fit, model complexity, variance components, and prediction accuracy (AUC in cross-validation) by model. The comparison of models based on COV only with those based on one omic (either METH, CNV or miRNA) suggest that a model using METH fits the data better and predicts equally well (actually slightly more accurate) survival than a model based on COV. This suggests that METH information is capturing differences due to tumor subtype and stage. This was not observed for models based on CNV or miRNA, in these two cases the model based on COV outperformed the prediction accuracy of the models based on either CNV or miRNA only.

The estimates of variance components derived from models using COV plus one omic (COV+CNV, COV+METH, COV+miRNA) shows METH and CNV explained a large fraction of inter-individual differences in risk that cannot be explained by COV; however, the 95% posterior credibility regions are all wide. According to DIC the models using COV plus one omic were all better than the model using COV only; however, the differences in DIC were, relative to the model based on COV, large for the case of COV+METH and COV+CNV and very small for the model COV+miRNA (only about 2 points). Finally, the evaluation of prediction accuracy suggests that adding either METH or CNV on a model based on COV increased prediction accuracy significantly (99% of the time in 200 cross-validations) but by 1.5 to 1.7 units of AUC. Considering all the results from these case study it appears that among the three omics evaluated METH was the one that explained larges proportion of variance in risk and contributed most to prediction power both when considered alone or in combination with COV.

**TABLE 3, ABOUT HERE**

**Case study IV: integrating multiple omics**

Among the four omics considered in the previous studies, METH and WGGE appeared to be the ones that explain larges proportion of variance and achieved the highest levels of prediction accuracy both when considered alone or in combination with COV. Therefore, in this case study, we consider integrating these two omics, together with COV, into a risk assessment model. Furthermore, we evaluate the impacts of including interactions between the two omics using a reaction norm model.

**Data** includes the individuals (N=218) that had complete information for COV, METH and WGGE.

**Sequence of models.** The baseline model (COV) included is the same as the ones described in CS-I. This model was first expanded by adding METH and WGGE additively (COV+METH+WGGE) and subsequently further expanded by adding interactions between omics (COV+METH× WGGE) using a
reaction norm model. As before, COV were included as fixed effects and omics as random effects. In all cases, the random effects were assumed to be Gaussian, with omic-specific variance. The additive model COV+METH+WGGE had two variance parameters linked to the main effects of each of the omics included and COV+METH×WGGE had three variance parameters, two for main effects and one for interactions.

**Results.** Table 4 shows the results obtained in CS-IV. In the additive model (COV+METH+WGGE), the two omics explained about 27% of the variance in risk that was not accounted for COV (this is estimated as the sum of the two variance components divide by the sum of the two variance components plus the error variance, which in the probit model is one). When interactions were added the estimated variance components of the main effects of each omic went down (this relative to the additive model) and the total proportion of variance in risk explained by omics (including main effects and interactions) stay roughly the same. The posterior mean of the log likelihood of the model COV+METH+WGGE was 15.8 higher than that of the COV model; this indicates that adding the two omics increased goodness of fit markedly. When interactions were added the change in the log-likelihood relative to COV+METH+WGGE was more modest. DIC (“smaller is better”) indicates a clear superiority of the additive model with two omics, relative to COV, and almost no difference between COV+METH+WGGE and COV+METH×WGGE. Finally, the evaluation of prediction accuracy from cross-validation showed that: (i) the baseline model had a reasonably good AUC (.724), (ii) the additive model improved the performance by 3 points in AUC, importantly this increase happened in 99% of the 200 cross-validations, and (iii) adding interactions did not clearly improved prediction accuracy (in 60% of the cross-validations the additive model was better than the model having interactions and in the other 40% the opposite happened).

**TABLE 4, ABOUT HERE**

**DISCUSSION**

The availability of multi-omic data sets has increased recently and this trend is expected to continue. Modern omic data sets can be big (large-N), high-dimension (each subject can have information on hundreds of thousands of variables), and have a multi-layer structure (e.g., data may involve clinical information, demographics, lifestyle and multiple omics). While recent advances in computational power and methodology have enhanced our ability to analyze these data sets, the availability of methods and data analysis tools for integrating high-dimensional-multi-layer inputs for the prediction of disease risk is lacking.

Statistical models for the analysis of multi-layer omic data should: (i) be able to integrate data from multiple omics, (ii) cope with high dimensional inputs, (iii) allow for different architecture of effects across layers and (iv) accommodates interactions between risk factors, including interactions between two or more high-dimensional sets. In this study we described a Bayesian Generalized Additive Model (BGAM) framework that fulfill those requirements. BGAMs integrate ideas from different sources, including: (a) Generalized Additive Models (GAM; Hastie 2008), (b) Bayesian regularized regressions (George and McCulloch 1993; Ishwaran and Rao 2005), and (c) modern approaches for modeling interactions between high-dimensional inputs primarily developed for the study of genetic-by-environment interactions (Jarquin et al. 2014). OmicKriging, a multi-omic risk assessment method (Vazquez et al. 2014; Wheeler et al. 2014), can be seen as a special case of the BGAM, which assumes additive action across omics and an homogenous architecture of effects (with Gaussian assumptions).
across layers. Within the BGAM framework, some of these assumptions can be relaxed by specifying different prior distributions of effects across layers, by using layer-specific regularization parameters (e.g., layer-specific variances), and by incorporating interactions within or between layers using either parametric or semi-parametric procedures. The BGLR R-package (Pérez and de los Campos, 2014) allows incorporating all these features for quantitative (censored or not), ordinal and binary traits. In our application we used the BGLR with data from TCGA to build risk assessment models for prediction of survival of BC patients using clinical covariates and multiple omics.

Omic information (e.g., gene expression patterns) can reveal important processes taking place at the cellular level. Previous studies (Wheeler et al. 2014) have shown successful integration of multi-layer omics for prediction of cell-phenotypes. In these studies, the phenotype was measured at the same cells where omics were assessed. Prediction of whole-organism phenotypes is considerably more challenging due to inter-cell variations in omics and traits and because the link between the cellular processes at the tissues where omics were assessed and target phenotype/disease may be weak due to multiple intervening factors. Perhaps for this reason, the integration of multiple omics for prediction of whole-organism phenotypes has been much more limited. For instance, using OmicKriging Wheeler et al. (2014) did not observe benefits in integrating DNA and gene-expression information for prediction of a pharmacogenetic trait (change in low-density lipoprotein cholesterol after simvastatin treatment), relative to models based on DNA information only.

Recently, Yuan et al. (2014) considered integrating omics with clinical covariates for prediction of survival in four different types of cancer (ovarian, kidney renal, glioblastoma multiforme and lung squamous cell carcinoma). In most cases the authors did not find a significance gain in prediction accuracy by combining omics with clinical covariates relative to the covariate-only model. In a few combinations of cancers and omics the authors reported a statistically significant gain in prediction accuracy, but the magnitude of the gain was very low. In our study we found significant gains in prediction accuracy when integrating either GE or METH, with gains in AUC ranging from 2 to 7 points. An important difference between the study by Yuan et al. and this study is that the modeling approach used here (BGAM) assigned different regularization parameters for different sets of inputs. This allows the model to weight differentially information from clinical covariates and from different omics. To illustrate the importance of assigning different priors/regularization parameters for different omics we conducted a sensitivity analysis where we fitted the model incorporating COV, GE and METH of case study IV without assigning different priors/regularization parameters for each of the three inputs sets. The results are presented in Supplementary File S1, Table S1.4. Assigning the same prior/regularization parameters to all the effects resulted in a substantial loss in AUC: from 0.754 (model COV+GE+METH, case study IV) to levels of AUC of the order of 0.56 when the same inputs were assigned the same prior/regularization parameters and 0.64 when the same inputs were assigned the same prior in a variable selection model (Bayes B).

Breast carcinoma becomes lethal after migrating from the breast with the development of distant metastasis on organs (e.g., brain or liver). An important strength of our application is that all the omics used for prediction of survival of BC patients were assessed at the primary tumor: the tissue where the disease is unfolding. An additional strength of this application is that the overwhelming majority of cancer samples are primary tumor only. Our response variable considered alive status (0/1), however one could also regress survival time as a censored outcome on covariates and omics using parametric (e.g., Weibull, log-normal) or semi-parametric regression (e.g., Cox proportional hazard regression, Cox
Several risk factors are associated with the likelihood of developing distant metastasis and ultimately survival. The risk factors commonly considered when assessing cancer patients, including tumor type, subtype, and stage, were found to be significantly associated with survival in TCGA. Other factors commonly considered when assessing BC patients, including lymph node invasion, marginal status (whether cancerous cells are present in the remaining margins at the site of surgery), increased size of the primary tumor, and level of loss of histopathology differentiation in the tumor cells themselves were not significantly associated with survival when a full set of COV was included. Consequently, our baseline COV model included demographics (race and age at diagnosis) and the three clinical covariates that had significant association with survival (lobular/ductal, tumor subtype and stage).

Cancer subtype and stage are the most important predictors considered by a clinician when assessing BC patients. Our study shows that these two predictors are indeed the clinical COV that offer highest prediction accuracy. Our CS-I also shows that WGGE had more predictive power than any of the predictors commonly used in clinical practice, including cancer subtype and state, which are well established in the literature (Koscielny et al. 1984; Carter et al. 1989; Rosen et al. 1989; Elston et al. 1991; Sørlie et al. 2001; Weigelt et al. 2005) as clinical predictors BC progression and survival.

Gene expression is informative of cancer subtype and stage; indeed, GE patterns are predictive of intrinsic subtypes, which are then confirmed by receptor subtype (Sørlie et al. 2001). However, our results suggest even after accounting for all the variables commonly used to assess cancer patients, including stage and cancer subtypes, the addition of WGGE can further improve prediction accuracy. The gains in prediction accuracy obtained when adding WGGE were moderate in magnitude (2.0-2.5 points in AUC) when we considered all the cancer subtypes together, to very relevant (7 points in AUC) when models were fitted to a particular subtype, as it was the case of our CS-II. Because the COV model includes cancer stage and subtype, which are correlated with gene-expression derived clusters (Sørlie et al. 2001), the gains in prediction accuracy obtained with the addition of WGGE cannot be attributed to clustering. To demonstrate this we derived the leading 5 principal components (PC) of GE and tested the significance of adding these PC as predictors in the COV model using a likelihood ratio test. The results (see Supplementary File 1, Table S1.5) indicated that after accounting for COV the leading 5 PC did not have a significant effect on survival. Therefore, we conclude that the gains in prediction accuracy observed are largely due to patterns other than the clustering obtainable with the first PC from GE.

The predictive power of gene expression profiles was established in the literature more than a decade ago. However, risk assessment is typically based on the expression profiles of a few large-effect genes (Paik et al. 2004b; Glas et al. 2006). Results from SNP data in other context suggest that the information of large number of markers may increases the variance explained of the phenotype than pre-selecting small numbers of SNPs (Allen et al. 2010; Vazquez et al. 2010). While the expression profiles of pre-selected genes are certainly predictive of BC outcomes, valuable information may be lost when ignoring the non-selected genes. The results form our CS-II confirm this hypothesis. Indeed, our results indicate that the use of WGGE profiles lead to a larger proportion of variance in risk explained and provide higher prediction accuracy of BC patient survival than the one that can be obtained using the expression profiles of a few oncogenes. With modern sequencing technologies, assessing WGGE has become feasible and it should be economically viable. We argue that the use of WGGE for assessment of BC patients should receive more attention.
In addition to WGGE, CNV and METH offer some promising results. Methylation has been pointed out as an interesting set to predict plant traits (Hu et al. 2015). In our study, methylation considered alone offered higher prediction accuracy than a model based on clinical predictors, including both cancer subtype and stage. Further studies are needed to assess whether the association between methylation and survival is due to common factors (e.g., carcinogenic factors that affect methylation pattern and BC progression at the same time) or due to mediation (e.g., that the effects of carcinogenic factors may be mediated by methylation). On the other hand, our results did not show large variance associated to miRNA in survival of BC patients. Further studies with larger sample sizes will be needed to determine whether our results for the models involving miRNA are due to lack of power or due to weak association between miRNA profiles and survival.

Methylation additively integrated to whole genome gene expression explains about 30% of the variance in risk that was not explained by COV and the AUC of the model was 3 points greater than the one achieved when using COV. This gain in AUC is slightly greater (about 1 point greater) than what we achieved in CS-I and CS-III when we added one of the omics at a time. This suggests that even though METH and WGGE provide, to some extent, redundant information, such redundancy is not complete and there may be some benefits of including both omics in a model. When we added interactions we did not observe improvements in performance relative to the additive model. Neither the proportion of variance explained by omics nor prediction increased relative to the additive model. Further studies with higher sample size and perhaps with analyses within cancer subtype are needed to fully explore the potential benefits of including multiple omics with omic-by-omic interaction.

In this study we demonstrate how clinical information can be integrated with whole-genome omic data derived from several omic layers, including the genome (e.g., CNV), epigenome (METH) and transcriptome (microRNA and WGGE). With some of these omics we found statistically significant and, in some cases, substantial gains in prediction accuracy relative to models based on clinical COV. However, our ability to detect improvement may have been limited by three main factors: (i) small sample size, most of the models that we consider involves large number of effects. Although Bayesian methods allow handling high-dimensional predictors even in settings where the number of effects exceeds sample size, the accuracy of estimates of individual effects is low when the number of effects is large relative to sample size. Therefore, considerably larger number of samples will be needed to realize the potential contribution to prediction accuracy; (ii) in three of our four CS we treated BC as a single disease, and included the cancer subtype in the model. When BC is treated as a homogeneous disease a large fraction of inter-individual difference in survival can be attributed to cancer sub-type. We decided to carry out CS-I, III and IV based on all BC cases because carrying out analyses within cancer subtype would have reduced the sample size. In the only case where we consider a within-subtype analysis, our CS-II, we detected gains in prediction accuracy considerably larger than when considering all subtypes jointly. This suggests that omics may contribute significantly to prediction of inter-individual differences in progression and survival within sub-types, hence, paving the way towards a more precise approach to the treatment of BC patients. Finally, (iii) TCGA is a relatively new repository for breast cancer and hence, follow-up time is short for many patients, limited follow up time reduces the information content of each of the cases. In the near future the availability of large data sets comprising clinical information and multi-layer omic data is granted to increase. Such data sets will allow researchers to explore the limits of what multi-layer omic data can contribute for prediction of BC progression and patient survival.
LITERATURE CITED

Agresti A., 2012 *Categorical Data Analysis*. Wiley, Hoboken, NJ.

Albert J. H., Chib S., 1993 Bayesian analysis of binary and polychotomous response data. Journal of the American statistical Association 88: 669–679.

Allen H. L., Estrada K., Lettre G., Berndt S. I., Weedon M. N., Rivadeneira F., Willer C. J., Jackson A. U., Vedantam S., Raychaudhuri S., others, 2010 Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832–838.

Beroukhim R., Mermel C. H., Porter D., Wei G., Raychaudhuri S., Donovan J., Barretina J., Boehm J. S., Dobson J., Urashima M., 2010 The landscape of somatic copy-number alteration across human cancers. Nature 463: 899–905.

Boyle P., Levin B., 2008 International Agency for Research on Cancer. World Health Organization: World Cancer Report.

Calus M. P. L., Groen A. F., De Jong G., 2002 Genotype X environment interaction for protein yield in Dutch dairy cattle as quantified by different models. Journal of dairy science 85: 3115–3123.

Carter C. L., Allen C., Henson D. E., 1989 Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. Cancer 63: 181–187.

Chen R., Mias G. I., Li-Pook-Than J., Jiang L., Lam H. Y., Chen R., Miriami E., Karczewski K. J., Hariharan M., Dewey F. E., others, 2012 Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 148: 1293–1307.

Cox D. R., 1972 Regression models and life tables. Journal of the Royal Statistical Society 34: 187–220.

Cressie N., 2015 *Statistics for Spatial Data*. Wiley-Interscience, Hoboken, NJ.
de los Campos G., Gianola D., Rosa G., 2009 Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. Journal of animal science 87: 1883–1887.

de los Campos G., Gianola D., Rosa G., Weigel K., Vazquez A., Allison D., 2010a Semi-parametric marker-enabled prediction of genetic values using reproducing kernel Hilbert spaces regressions. 9th World Congress on Genetics Applied to Livestock Production, Leipzig, Germany. CD-ROM Communication 520.

de los Campos G., Gianola D., Rosa G. J., Weigel K. A., Crossa J., 2010b Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genetics Research 92: 295–308.

de los Campos G., Gianola D., Allison D. B., 2010c Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet 11: 880–886.

Dedeurwaerder S., Desmedt C., Calonne E., Singhal S. K., Haibe-Kains B., Defrance M., Michiels S., Volkmar M., Deplus R., Luciani J., others, 2011 DNA methylation profiling reveals a predominant immune component in breast cancers. EMBO molecular medicine 3: 726–741.

Edge S., Byrd D. R., Compton C. C., Fritz A. G., Greene F. L., Trotti A., 2010 AJCC Cancer Staging Manual. Springer-Verlag, New York, NY.

Eifel P., Azelson J., Costa J., Crowley J., Curran J., Deshley A., Fulton S., Hendricks C., Kemeny M., Kornblith A., 2000 National Institutes of Health Consensus Development Conference Statement: adjuvant therapy for breast cancer. Journal of the National Cancer Institute.

Elston C. W., Ellis I. O., others, 1991 Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology 19: 403–410.
Fackler M. J., Umbricht C. B., Williams D., Argani P., Cruz L.-A., Merino V. F., Teo W. W., Zhang Z., Huang P., Visvananthan K., others, 2011 Genome-wide methylation analysis identifies genes specific to breast cancer hormone receptor status and risk of recurrence. Cancer research 71: 6195–6207.

Fang F., Turcan S., Rimner A., Kaufman A., Giri D., Morris L. G., Shen R., Seshan V., Mo Q., Heguy A., others, 2011 Breast cancer methylomes establish an epigenomic foundation for metastasis. Science translational medicine 3: 75ra25–75ra25.

Fawcett T., 2006 An introduction to ROC analysis. Pattern recognition letters 27: 861–874.

George E. I., McCulloch R. E., 1993 Variable selection via Gibbs sampling. Journal of the American Statistical Association 88: 881–889.

Gianola D., Foulley J., 1983 Sire evaluation for ordered categorical data with a threshold model. Genet Sel Evol 15: 201–224.

Glas A. M., Floore A., Delahaye L. J., Witteveen A. T., Pover R. C., Bakx N., Lahti-Domenici J. S., Bruinsma T. J., Warmoes M. O., Bernards R., others, 2006 Converting a breast cancer microarray signature into a high-throughput diagnostic test. BMC genomics 7: 278.

Golub G. H., Heath M., Wahba G., 1979 Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21: 215–223.

Gregorius H.-R., Namkoong G., 1986 Joint analysis of genotypic and environmental effects. Theoretical and applied genetics 72: 413–422.

Gyorffy B., Bottai G., Fleischer T., Munkácsy G., Budczies J., Paladini L., Børresen-Dale A.-L., Kristensen V. N., Santarpia L., 2016 Aberrant DNA methylation impacts gene expression and prognosis in breast cancer subtypes. International Journal of Cancer 138: 87–97.

Hastie T., Tibshirani R., 1986 Generalized Additive Models. Statistical Science 1: 297–318.
Hayes B. J., Goddard M. E., others, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Henderson C. R., 1950 Estimation of genetic parameters. Ann. Math. Stat. 21: 309.

Henderson C. R., 1975 Best linear unbiased estimation and prediction under a selection model. Biometrics 31: 423–447.

Hu Y., Morota G., Rosa G. J., Gianola D., 2015 Prediction of Plant Height in Arabidopsis thaliana Using DNA Methylation Data. Genetics 201: 779–793.

Ishwaran H., Rao J. S., 2005 Spike and slab variable selection: frequentist and Bayesian strategies. Annals of Statistics: 730–773.

Jarquín D., Crossa J., Lacaze X., Du Cheyron P., Daucourt J., Lorgeou J., Piraux F., Guerreiro L., Pérez P., Calus M., others, 2014 A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoretical and Applied Genetics 127: 595–607.

Koscielny S., Tubiana M., Le M. G., Valleron A. J., Mouriesse H., Contesso G., Sarrazin D., 1984 Breast cancer: relationship between the size of the primary tumour and the probability of metastatic dissemination. British journal of cancer 49: 709.

Li H., Durbin R., 2009 Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754–1760.

Li B., Ruotti V., Stewart R. M., Thomson J. A., Dewey C. N., 2010 RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics 26: 493–500.

Makowsky R., Pajewski N. M., Klimentidis Y. C., Vazquez A. I., Duarte C. W., Allison D. B., de los Campos G., 2011 Beyond Missing Heritability: Prediction of Complex Traits. PLoS Genet 7: e1002051.
Morrow E. M., 2010 Genomic copy number variation in disorders of cognitive development. Journal of the American Academy of Child & Adolescent Psychiatry 49: 1091–1104.

Network C. G. A., 2012 Comprehensive molecular portraits of human breast tumours. Nature 490: 61–70.

Oliveros J. C., 2007 VENNY. An interactive tool for comparing lists with Venn Diagrams. Accessed 5/2016

Paik S., Shak S., Tang G., Kim C., Baker J., Cronin M., Baehner F. L., Walker M. G., Watson D., Park T., others, 2004a A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. New England Journal of Medicine 351: 2817–2826.

Paik S., Shak S., Tang G., Kim C., Baker J., Cronin M., Baehner F. L., Walker M. G., Watson D., Park T., Hiller W., Fisher E. R., Wickerham D. L., Bryant J., Wolmark N., 2004b A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. New England Journal of Medicine 351: 2817–2826.

Paik S., Tang G., Shak S., Kim C., Baker J., Kim W., Cronin M., Baehner F. L., Watson D., Bryant J., others, 2006 Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor–positive breast cancer. Journal of clinical oncology 24: 3726–3734.

Park T., Casella G., 2008 The bayesian lasso. Journal of the American Statistical Association 103: 681–686.

Pérez P., Campos G. de los, 2014 Genome-wide regression & prediction with the BGLR statistical package. Genetics: genetics–114.

Perou C. M., Sørlie T., Eisen M. B., Rijn M. van de, Jeffrey S. S., Rees C. A., Pollack J. R., Ross D. T., Johnsen H., Akslen L. A., Fluge O., Pergamenschikov A., Williams C., Zhu S. X., Lønning P. E., Børresen-Dale A. L., Brown P. O., Botstein D., 2000 Molecular portraits of human breast tumours. Nature 406: 747–752.

Pidsley R., Wong C. C., Volta M., Lunnon K., Mill J., Schalkwyk L. C., 2013 A data-driven approach to preprocessing Illumina 450K methylation array data. BMC genomics 14: 293.
Purcell S. M., Wray N. R., Stone J. L., Visscher P. M., O’Donovan M. C., Sullivan P. F., Sklar P., Ruderfer D. M.,
McQuillin A., Morris D. W., others, 2009 Common polygenic variation contributes to risk of
schizophrenia and bipolar disorder. Nature 460: 748–752.

Robinson G. K., 1991 That BLUP is a Good Thing: The Estimation of Random Effects. Statist. Sci. 6: 15–32.

Rosen P. P., Groshen S., Saigo P. E., Kinne D. W., Hellman S., 1989 Pathological prognostic factors in stage I
(T1N0M0) and stage II (T1N1M0) breast carcinoma: a study of 644 patients with median follow-up of 18
years. Journal of Clinical Oncology 7: 1239–1251.

Shawe-Taylor J., Cristianini N., 2004 Kernel methods for pattern analysis. Cambridge university press.

Smigal C., Jemal A., Ward E., Cokkinides V., Smith R., Howe H. L., Thun M., 2006 Trends in breast cancer by race
and ethnicity: update 2006. CA: a cancer journal for clinicians 56: 168–183.

Sørlie T., Perou C. M., Tibshirani R., Aas T., Geisler S., Johnsen H., Hastie T., Eisen M. B., Rijn M. van de, Jeffrey S.
S., others, 2001 Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical
implications. Proceedings of the National Academy of Sciences 98: 10869–10874.

Sorlie T., Tibshirani R., Parker J., Hastie T., Marron J. S., Nobel A., Deng S., Johnsen H., Pesich R., Geisler S.,
Demeter J., Perou C. M., Lønning P. E., Brown P. O., Børresen-Dale A.-L., Botstein D., 2003 Repeated
observation of breast tumor subtypes in independent gene expression data sets. Proc. Natl. Acad. Sci.
U.S.A. 100: 8418–8423.

Sotiriou C., Pusztai L., 2009 Gene-expression signatures in breast cancer. New England Journal of Medicine 360:
790–800.

Su G., Madsen P., Lund M. S., Sorensen D., Korsgaard I. R., Jensen J., 2006 Bayesian analysis of the linear reaction
norm model with unknown covariates. Journal of animal science 84: 1651–1657.
Tibshirani R., 1996 Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological): 267–288.

VanRaden P. M., 2008 Efficient methods to compute genomic predictions. Journal of dairy science 91: 4414–4423.

Van’t Veer L. J., Dai H., Van De Vijver M. J., He Y. D., Hart A. A., Mao M., Peterse H. L., Kooy K. van der, Marton M. J., Witteveen A. T., others, 2002 Gene expression profiling predicts clinical outcome of breast cancer. nature 415: 530–536.

Vazquez A., Rosa G., Weigel K., Campos G. de los, Gianola D., Allison D., 2010 Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. Journal of dairy science 93: 5942–5949.

Vazquez A. I., de los Campos G., Klimentidis Y. C., Rosa G. J., Gianola D., Yi N., Allison D. B., 2012 A Comprehensive Genetic Approach for Improving Prediction of Skin Cancer Risk in Humans. Genetics 192: 1493–1502.

Vazquez A. I., Wiener H., Shrestha S., Tiwari H., Campos G. de los, 2014. Integration of Multi-Layer Omic Data for Prediction of Disease Risk in Humans. In: 10th World Congress on Genetics Applied to Livestock Production, Vancouver Canada, (213), American Society of Animal Sciences.

Vazquez A. I., Klimentidis Y. C., Dhurandhar E. J., Veturi Y. C., Paérez-Rodríguez P., 2015 Assessment of whole-genome regression for type II diabetes. PloS one 10: e0123818.

Wahba G., 1990 Spline models for observational data. Siam (59).

Wang K., Singh D., Zeng Z., Coleman S. J., Huang Y., Savich G. L., He X., Mieczkowski P., Grimm S. A., Perou C. M., others, 2010 MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic acids research 38: e178–e178.
Weigelt B., Peterse J. L., Van’t Veer L. J., 2005 Breast cancer metastasis: markers and models. Nature reviews cancer 5: 591–602.

Wheeler H. E., Aquino-Michaels K., Gamazon E. R., Trubetskov V. V., Dolan M. E., Huang R. S., Cox N. J., Im H. K., 2014 Poly-Omic Prediction of Complex Traits: OmicKriging. Genetic epidemiology 38: 402–415.

Yang J., Benyamin B., McEvoy B. P., Gordon S., Henders A. K., Nyholt D. R., Madden P. A., Heath A. C., Martin N. G., Montgomery G. W., others, 2010 Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42: 565–569.

Yuan Y., Van Allen E. M., Omberg L., Wagle N., Amin-Mansour A., Sokolov A., Byers L. A., Xu Y., Hess K. R., Diao L., others, 2014 Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nature biotechnology 32: 644–652.

1- ACKNOWLEDGEMENTS

We thank Kyle Grimes for editing the manuscript. We also thank The Cancer Genome Atlas network for data access (http://cancergenome.nih.gov/). A.I.V. acknowledges financial support from National Institutes of Health grant 7-R01-DK-062148-10-S1; A.I.V. and G.D.L.C. acknowledge support from National Institutes of Health grants R01-GM-099992 and R01-GM-101219 and National Science Foundation grant 1444543, subaward UFDSP00010707. A.I.V, M.B. and S.S. acknowledge financial support from American Cancer Society Institutional Research Grant 60-001-53-IRG, University of Alabama at Birmingham-Comprehensive Cancer Center.
Figure 1: Venn diagram with the number of patients that had information by omic layer (CNV=Copy Number Variant, miRNA=miRNA, RNA=RNA abundance measured with RNA-seq).
## Tables and Figures

### Table 1. Parameter estimates, model goodness of fit, model complexity and prediction accuracy (Case Study 1).

| Model | Predictors | 200-Cross-Validations |
|-------|------------|-----------------------|
|       | Whole Data Analysis |                      |
|       | Predictors | Log Likelihood | Effective number of Parameters (pD) | Deviance Information | Average Cross-validation AUC (3) | Proportion of times (out of 200 cross-validations) that the model in column had AUC greater than the model in row |
|       | Age at diagnosis | Race (1) | Lobular (Y/N) | Tumor Subtype | Pathological Stage | Gene Expression M2 | M3 | M4 | M5 | M6 | M7 (COV) | M8 (COV+WGGE) |
| M1    | X | -146.1 | 2.1 | 294.3 | .557a (0.007) | .14 | <.01 | >.99 | >.99 | >.99 | >.99 | >.99 |
| M2    | X | -147.5 | 2.0 | 296.9 | .525a,b (0.023) | .59 | >.99 | >.99 | >.99 | >.99 | >.99 | >.99 |
| M3    | X | -144.3 | 2.0 | 290.6 | .526b (0.020) | >.99 | >.99 | >.99 | >.99 | >.99 | >.99 | >.99 |
| M4    | X | -138.6 | 4.1 | 281.3 | .618c (0.013) | .14 | >.99 | >.99 | >.99 | >.99 | >.99 | >.99 |
| M5    | X | -142.4 | 2.0 | 286.9 | .596c (0.012) | >.99 | >.99 | >.99 | >.99 | >.99 | >.99 | >.99 |
| M6    | X | -132.4 | 15.5 | 280.3 | .659d (0.011) | >.99 | >.99 | >.99 | >.99 | >.99 | >.99 | >.99 |
| M7: COV | X | -146.3 | 3.2 | 295.8 | .704e (0.007) | >.99 | >.99 | >.99 | >.99 | >.99 | >.99 | >.99 |
| M8: COV+WGGE | X | -131.3 | 17.6 | 280.3 | .721f (0.010) | >.99 | >.99 | >.99 | >.99 | >.99 | >.99 | >.99 |

(1): African American Yes/No; (2): estimated posterior mean of the log-likelihood; (3): average over 200 ten-fold cross-validations. Same letter in the superscript of the Average AUC indicates that the models are no different (empirical p-value<0.05).
Table 2. Parameter estimates, model goodness of fit, model complexity and prediction accuracy (Case Study II).

| Model                  | Whole Data Analysis | 200-Cross-Validations |
|------------------------|---------------------|-----------------------|
|                        | Estimated Variance [90% Posterior Confidence Region] | AUC model in column > the one in row (4). All Luminals |  |
|                        | Log Likelihood (2) | Deviance Information Criteria (DIC) | Average Cross-Validation AUC (3) | AUC model in column > the one in row (4). Lymph node negative |
|                        | Effective number of Parameters (pD) | M10 | M11 | M10 | M11 |
| Model                  | Oncotype-DX | Whole-Genome Gene Expression (WGGE) | M9 (COV) | x | - | - | -59.7 | 4.4 | 123.7 | .703<sup>a</sup> (.026) | .96 | >.99 | .689<sup>a</sup> (.052) | .43 | .99 |
|                        | Oncotype-DX | Whole-Genome Gene Expression (WGGE) | M10 (COV+ONCO) | x | X | .027 [.003;.056] | - | -45.3 | 4.2 | 94.9 | .725<sup>b</sup> (.033) | - | .99 | .685<sup>a</sup> (.055) | - | >.99 |
|                        | Oncotype-DX | Whole-Genome Gene Expression (WGGE) | M11(COV+WGGE) | X | X | .439 [.083;.931] | - | -37.7 | 9.2 | 84.6 | .774<sup>c</sup> (.031) | - | - | .755<sup>c</sup> (.039) | - | - |

(1): Age and race (African American Yes/No); (2): estimated posterior mean of the log-likelihood; (3): average over 200 ten-fold cross-validations. (4): Proportion of times that the model in column had AUC greater than the model in row (in 200 ten-fold cross validations). Same letter in the superscript of the Average AUC indicates that the models are no different (empirical p-value<0.05).
Table 3. Parameter estimates, model goodness of fit, model complexity and prediction accuracy (Case Study III).

| Set      | Model      | Factors Included | Variance [90% Posterior Confidence Region] | Log Likelihood(5) | Effective number of Parameters (pD) | Deviance Information Criteria (DIC) | Average Cross-Validation AUC(6) | Proportion of times that the model in the column had AUC greater than the model in the row |
|----------|------------|-----------------|---------------------------------------------|-------------------|-------------------------------------|------------------------------------|---------------------------------|----------------------------------------------------------------------------------|
|          |            |                 |                                             |                   |                                     |                                    |                                 |                                                                                   |
| Set 1 (n=270) | M12: COV   | X               | -                                           | -125.5            | 8.1                                 | 259.0                              | 0.699 ± 0.009                    | < .01 | > .99 |
|          | M13: CNV   | X               | 0.637 [0.155; 1.124]                        | -112.1            | 26.8                                | 250.9                              | 0.653 ± 0.012                    | -      | > .99 |
|          | M14: COV+CNV | X X             | 0.398 [0.070; 0.736]                        | -110.5            | 24.5                                | 245.6                              | 0.714 ± 0.009                    | -      | -      |
| Set 2 (n=199) | M15: COV   | X               | -                                           | -88.7             | 8.4                                 | 185.7                              | 0.667 ± 0.013                    | 0.60 | > .99 |
|          | M16: METH  | X               | 0.652 [0.086; 1.261]                        | -76.6             | 18.7                                | 171.8                              | 0.672 ± 0.017                    | -      | .76   |
|          | M17: COV+METH | X X             | 0.402 [0.032; 0.739]                        | -78.9             | 18.5                                | 176.3                              | 0.684 ± 0.013                    | -      | -      |
| Set 3 (n=167) | M18: COV   | X               | -                                           | -71.2             | 8.2                                 | 150.6                              | 0.747 ± 0.011                    | < .01 | .29   |
|          | M19: miRNA | X               | 0.338 [0.072; 0.615]                        | -75.2             | 13.5                                | 163.8                              | 0.623 ± 0.018                    | -      | > .99 |
|          | M20: COV+ miRNA | X X             | 0.179 [0.029; 0.324]                        | -67.3             | 13.8                                | 148.5                              | 0.744 ± 0.011                    | -      | -      |

(1): Age, African American Yes/No, Lobular (Y/N), Cancer subtype and Stage,(2): Copy-number variants,(3): methylation,(4): Whole-Genome RNA-seq,(5): estimated posterior mean of the log-likelihood; (6): average over 200 ten-fold cross-validations. Same letter in the superscript of the Average AUC indicates that the models are no different (empirical p-value<0.05).
Table 4. Parameter estimates, model goodness of fit, model complexity and prediction accuracy (Case Study IV).

| Models Components | Whole Data Analysis | 200-Cross-Validations |
|-------------------|---------------------|-----------------------|
|                   | Estimated Variance  | Log Likelihood        | Effective number of Parameters (pD) | Deviance Information Criteria (DIC) | Average Cross-Validation AUC | Proportion of times that the model in column had AUC greater than the model in row |
|                   | [90% Posterior Confidence Region] |                     |                           |                                    |                           |                                      |
|                   | METH                | WGGE                 | METH×WGGE                | METH×WGGE                        |                           |                                      |
| COV               | X                   | -                    | -                        | -                                 | -85.7                     | 6.4                                   | 177.9                              | .724<sup>a</sup> (0.001) | >.99 | >.99 |
| COV+ METH+WGGE   | X X X               | .162 [0.075; 0.440] | .220 [0.090; 0.690] | -                                 | -73.9                     | 17.6                                  | 165.4                              | .754<sup>b</sup> (0.004) | -    | .40  |
| COV+ METH×WGGE   | X X X X             | .101 [0.046; 0.272] | .138 [0.055; 0.474] | .101 [0.044; 0.329] | -69.9                     | 20.2                                  | 159.9                              | .753<sup>b</sup> (0.005) | -    | -    |

(1): Age, African American Yes/No, Lobular (Y/N), and Tumor Subtype; (2): Methylation; (3): Whole-Genome RNA-seq; (4): Methylation-by-WGGE; (5): estimated posterior mean of the log-likelihood; (6): average over 200 ten-fold cross-validations. Same letter in the superscript of the Average AUC indicates that the models are no different (empirical p-value<0.05).