Integration of Multi-Layer Omic Data for Prediction of Disease Risk in Humans

AI. Vazquez[1*], HW. Wiener, S. Shrestha[1], H. Tiwari, G. de los Campos [1]
[1]University of Alabama at Birmingham, AL, USA; [*]Corresponding author: anainesvs@gmail.com

ABSTRACT: Accurate prediction of disease risk is needed for implementing personalized medicine. Despite important advances in the assessment of genetic risk, our ability to predict disease risk based on information from the genome (e.g., SNPs) remains very limited. Owing to developments in high-throughput technologies integrated omic profiles are becoming increasingly available. These data holds information that can be extremely useful for the assessment of disease risk and progression. However, omic data is high dimensional and complex, and we lack a coherent framework for the integration of multi-layer omic data into risk assessment models. **In this preceding, we discuss extensions of Whole-Genome Regressions that can be used to incorporate integrated omic profiles for the assessment of disease risk**. Some of the models described are evaluated using whole-genome expression profiles for prediction of survival after diagnose of breast cancer.

Keywords: prediction of complex traits; diseases risk; omics integration.

INTRODUCTION

Modern genotyping and sequencing technologies can deliver large volumes of data from multiple omic layers, including the genome (e.g., SNPs, CNV), epigenome (e.g., methylation) and transcriptome (RNA abundance). In recent years, several data sets comprising disease, clinical and omic information have been created and made publicly available through data repositories such as DataBases of Genotypes And Phenotypes (**dbGaP**) or the European Genome-phenome Archive (**EGA**). These data hold extremely useful information that could be used for the development of risk-assessment models.

Since the completion of the human genome project in 2003 (Lander et al., 2001; Venter et al., 2001; Anon, 2003), several Genome Wide Association Studies (**GWAS**) have been conducted; these studies have uncovered un-precedent numbers of variants associated with important human traits and diseases (e.g., www.genome.gov/gwastudies/). These findings have been used to develop risk scores based on either simple or weighted counts of risk-alleles at GWAS-significant loci (e.g., Dominiczak and McBride, 2003; De Jager et al., 2009; Chen et al., 2011). Further, some studies considered statistical learning methods such as naïve Bayes classifier (Okser et al., 2010), support vector machines (Wei et al., 2009), random forest (Bureau et al., 2005), rule induction (Sebastiani and Perls, 2010), and Bayesian networks (Rodin and Boerwinkle, 2005).

However, for most diseases, risk-scores based on GWAS-significant variants explain only a small fraction of the inter-individual differences in genetic risk; a problem referred to as the missing heritability of complex traits and diseases (Maher, 2008). The case of body mass index (**BMI**) illustrates the extent of the problem: despite of BMI being a highly heritable trait, $h^2 \in [0.4,0.6]$, the 15 loci that have been consistently identified to be associated to BMI explain less than 2% of the observed variance on BMI (Loos, 2009).

The "missing heritability" problem has been discussed in the literature *in-extenso* (Maher, 2008; Manolio et al., 2009), and there is a general consensus that an important explanation of the problem resides on the lack of power of standard GWAS: in the majority of these studies a large fraction of small-effect variants do not reach genome-wide significance and their effects remain un-accounted for.

Recent studies have shown that predictive power of risk scores could be increased by considering, variants that have strong but not genome-wide significant association with the trait or disease of interest (Allen et al., 2010). When markers are pre-selected based on stringent p-value cut offs, the estimated proportion of variance explained reflects the predictive power of the selected set of markers. This under-estimates the true proportion of variance that can be potentially explained using all the genomic information available (e.g., common SNPs).

Yang et al., (2010) estimated the total proportion of variance that can be explained by common SNPs--hereinafter referred as to the 'genomic heritability'--using a Whole-Genome Regression (**WGR**) approach where phenotypes are regressed on all available SNPs concurrently. Using the G-BLUP method (VanRaden, 2008; Yang et al., 2010), Yang et al. estimated that 50% of the heritability of human height could be explained. Similar results were obtained by others (Purcell et al., 2009; Speed et al., 2012).

However, while able to account for a large proportion of the genetic variance, prediction accuracy depends also on other factors (Goddard and Hayes, 2007; de los Campos et al., 2012 b). Studies with animal (Goddard and Hayes, 2007; VanRaden, 2008; Vazquez et al., 2010), plant (Crossa et al., 2010; Resende et al., 2012) and human data (Makowsky et al., 2011; de los Campos et al., 2012 a; Vazquez et al., 2012) have shown that WGR can achieve high predictive power when discovery and validation samples are closely related. However, the predictive ability of WGR can be greatly affected by the genetic distance (Habier et al., 2010; Pérez-Cabal et al., 2012). Studies with WGR for prediction of phenotypes of distantly related individuals have shown poor predictive power (e.g., R2 in testing samples of the order of 5% for human height, a trait with heritability of 0.8 and genomic heritability of 0.5; de los Campos et al., 2013b).

Beyond the genome. The integration of multi-layer omic data into risk-assessment methods can be an avenue for advancing our ability to predict disease risk (Berghoff et al., 2013). Chen and co-authors (2012) demonstrated how integrated omic profiles of a person could provide insights into the development of Type 2 diabetes.

Multi-layer omic data is becoming increasingly available. Several GWAS have added information from layers other than the genome (e.g, epigenom, transcriptome). Recently, repositories have been created to deposit and share standardized multi-layer omic data linked to clinical information e.g., The Cancer Genome Atlas, (**TCGA**) (Chin et al., 2011).

Relative to genome-only data (e.g., SNPs) multi-layer omic data has several advantages. First, data from some omic layers (e.g., methylome, transcriptome, metabolome) can account for additive genetic factors and other factors, including non-heritable genetic ones (e.g., effects due to dominance or epistasis) and environmental factors. Accounting for non-heritable factors may not be critical for prediction of breeding values; however, exploiting such signals can have great impacts on the prediction accuracy of yet-to-be phenotypic and disease outcomes.

Second, by being 'biologically closer' to the disease outcome the distribution of effects of some omic layers (e.g., mutations at the tumor cell) may have a distribution of effects easier to deal-with (e.g., large effect-risk-factors may explain a larger proportion of variance in disease risk). Examples of these are signatures for diagnostic of cancer provided by the expression of genes that are known to be differentially expressed in tumor cells (Paik et al., 2004). Third, although redundancies between layers are likely to exist, the information from different omic layers may be complementary. For instance, it has been established that SNPs account only for a fraction of variation of the human genome, and that non-negligible portion of genetic variation can be attributed to structural variations (Forer et al., 2010).

The use of integrated omic profiles for prediction of disease risk is certainly attractive. However, integrating high dimensional data from multiple omic layers into prediction models poses important statistical and computational challenges. Additionally, the development of risk assessment methods lags behind (Palsson and Zengler, 2010). Most of the statistical methods and data analytic tools available focus in finding associations of disease outcomes with risk factors using a one-risk-factor-at-a-time approach. We argue that a perhaps more appropriate approach can be based on whole-genome-multi-layer methods. Therefore, in this article we discuss some approaches for data integration based on extensions of WGRs to multi-layer omic settings. What remains of this article is organized as follows. In the next Section (2) we present a brief review of WGR methods and discuss extensions for multi-layer omic data. In Section (3) we present preliminary results obtained with the application of a whole-transcriptome model for prediction of breast cancer outcomes. Finally, we close our article by providing in Section (4) some concluding remarks.

WHOLE-GENOME MODELS FOR MULTI-LAYER OMIC DATA

In this section we briefly review standard Bayesian WGR methods that are commonly use for prediction using data from the genome (e.g., SNPs), and discuss extensions of these models to accommodate multi-layer omic data.

The type of data we are considering consists of a phenotypic outcome y_i $(i = 1, \dots, n)$, e.g., concentration of plasma triglycerides, or a marker for a disease (e.g., levels of blood glucose indicating hyperglycemia), and a set of predictors, including: (a) non-genetic covariates, x_{ij} $(j = 1, \dots, p_F)$, and covariates from two or more omic layers. Considering two omic layers suffices to introduce our models. We denote them, $W = \{w_{ij}\}_{j=1}^{j=p_w}$ and $Z = \{z_{ij}\}_{j=1}^{j=p_z}$. For instance, $w_{ij} \in \{0,1,2\}$ may represent the genotype of the *i*th individual at the *j*th SNP, and z_{ij} may be the measure of gene expression at the *j*th gene on the *i*th individual. The phenotypic outcome may be quantitative or categorical. For ease of presentation we describe our models for a quantitative trait; the modifications needed to handle binary and censored outcomes would be extended as described elsewhere (Gianola and Foulley, 1983; de los Campos et al., 2012 a).

Baseline Model

The baseline model include the fixed effects of non-genetic risk covariates (e.g., sex, treatment) and a WGR on SNPs (Meuwissen et al., 2001; see detailed review in de los Campos et al., 2013a), which takes the form:

$$y_i = \mu + \sum_{j=1}^{j=p_F} x_{ij}\beta_j + \sum_{j=1}^{j=p_w} w_{ij}\alpha_{wj} + \varepsilon_i \qquad \textbf{[1]}$$
$$= \eta_i + \varepsilon_i$$

where μ is an intercept, β_j is the effect of the *j*th covariate, $\alpha_w = \{\alpha_{wj}\}_{j=1}^{j=p_w}$ are marker effects, ε_i are *iid* (independent and identically distributed) normal residuals with mean zero and variance σ_ε^2 and η_i is the linear predictor of the regression. The conditional distribution of the data given the parameters is

$$p(y|\beta, \sigma_\varepsilon^2) = \prod_{i=1}^{i=n} N(y_i|\eta_i, \sigma_\varepsilon^2) \qquad \textbf{[2]}$$

where $N(y_i|\eta_i, \sigma_\varepsilon^2)$ denotes a normal density with mean η_i and variance σ_ε^2.

Inferences are based on the posterior distribution of the parameters given the data, which is proportional to the likelihood [2], times the prior distribution. The intercept and the fixed effects $\{\beta_j\}$ are assigned un-informative (i.e., flat) priors. The residual variance is assumed to follow a scaled-inverse chi-square density, $\chi^{-2}(\sigma_\varepsilon^2|df_\varepsilon, S_\varepsilon)$ and marker effects are assigned *iid* informative priors, $\alpha_{wj} \sim^{iid} p(\alpha_{wj}|\Omega_w)$; therefore, in the baseline model the posterior density becomes,

$$p(\mu, \beta, \sigma_\varepsilon^2, \alpha_w|y) \propto$$
$$\prod_{i=1}^{i=n} N(y_i|\eta_i, \sigma_\varepsilon^2)\chi^{-2}(\sigma_\varepsilon^2|df_\varepsilon, S_\varepsilon)\left\{\prod_{j=1}^{j=p_w} p(\alpha_{wj}|\Omega_w)\right\}$$

The choice of the prior distribution assigned to marker effects, $p(\alpha_{wj}|\Omega_w)$, and the values of the hyper-parameters, Ω_w, will determine whether the model performs shrinkage of estimates of marker effects, variable selection or a combination of both. Commonly used priors include (*i*) the Gaussian prior which induces shrinkage that is either homogeneous across markers (if genotypes were standardized) or proportional to minor allele frequency (if the markers are not standardized). This prior is used to model genetic risk to highly complex traits e.g., human height (Yang et al., 2010). (*ii*) Priors from the thick tailed family (e.g., the scaled-t or the double-exponential) have, relative to the Gaussian prior, higher mass at zero and thicker tails. These priors assume that most predictors (genetic risk factor) have very small effect and a few have large effects. These types of priors are used in models Bayes A (Meuwissen et al., 2001) and the Bayesian Lasso (Park and Casella, 2008). We may also encounter traits and diseases for which some genetic regions may not contribute to genetic risk at all. Accommodating these types of genetic architectures requires using (*iii*) finite-mixture priors that assign a non-null prior probability for the effects to be equal to zero. These priors induce variable selection and shrinkage simultaneously. To achieve this, the most common practice is to use two-component mixture priors formed by combining a spike, that can be either a finite point of mass or a very sharp distribution centered at zero, and a relatively flat slab. The slab can be any density, two commonly used distributions for the slab are the Gaussian (Ishwaran and Rao, 2005) or a distribution from the thick-tailed family (Bayes B, Meuwissen et al., 2001).

Each of the prior densities above-described has one or more hyper-parameters ($\Omega_{.}$). These parameters (e.g. the variance of the normal density, or the scale and degrees of freedom of the scaled-t density, or the mixing proportions of a two-component mixture prior) control the extent of shrinkage and the propensity of the model to induce variable selection. These parameters can strongly influence inferences; therefore, the preferred approach consists of estimating these hyper-parameters from data (de los Campos et al., 2013 a; Gianola, 2013). In a fully Bayesian setting this is done by assigning a prior density to them. Considering this, the joint posterior of a model with fixed effects and one sets of random effects describing risk conferred by W becomes:

$$p(\mu,\beta,\sigma_\varepsilon^2,\alpha_w|y) \propto \prod_{i=1}^{i=n} N(y_i|\eta_i,\sigma_\varepsilon^2) \times \chi^{-2}(\sigma_\varepsilon^2|df_\varepsilon,S_\varepsilon) \prod_{j=1}^{j=p_w} p(\alpha_{wj}|\Omega_w)\ p(\Omega_w) \quad \textbf{[3]}$$

Additive Omic Model

An extension of the model above-described can be obtained by expanding η_i with addition of information from a second omic layer (Z) as follows,

$$\eta_i = \mu + \sum_{j=1}^{j=p_F} x_{ij}\beta_j + \sum_{j=1}^{j=p_w} w_{ij}\alpha_{wj} + \sum_{j=1}^{j=p_z} z_{ij}\alpha_{zj} \textbf{[4]}$$

above $\alpha_w = \{\alpha_{wj}\}_{j=1}^{j=p_W}$ and $\alpha_z = \{\alpha_{zj}\}_{j=1}^{j=p_Z}$ represent regression coefficients representing the main effects of predictors in W and Z on the risk score. Each omic set may have a different prior assigned to α_w and α_z and also have separate hyper-parameters reflecting a different distribution of the effects of each omic set. Thus [3] would can expand as follows:

$$p(\mu,\beta,\sigma_\varepsilon^2,\alpha_w,\alpha_z,\Omega_w,\Omega_z|y) \propto \prod_{i=1}^{i=n} N(y_i|\eta_i,\sigma_\varepsilon^2)\chi^{-2}(\sigma_\varepsilon^2|df_\varepsilon,S_\varepsilon) \times \prod_{j=1}^{j=p_W} p(\alpha_{wj}|\Omega_w)\ p(\Omega_w)\ \prod_{j=1}^{j=p_Z} p(\alpha_{zj}|\Omega_z)\ p(\Omega_z) \quad \textbf{[5]}$$

where $p(\alpha_{.j}|\Omega_{.})$ denotes the prior density assigned to $\alpha_{.j}$, and $p(\Omega_{.})$ the prior distribution of the hyper-parameters.

Accounting for Interactions

The models described so far are additive, since the effect of any given predictor does not depend on other predictors. However, the effects of some predictors maybe modulated by other predictors. For instance, the effects of genes on gene expression and ultimately on phenotypes may be modulated by methylation. Different types of interactions that could extend the model in expressions **[4]** and **[5]**, as follows:

Case 1. A first type of interactions includes a major factor (e.g., treatment A or B) and high dimensional predictors (e.g., gene expression levels). For instance, the response to a cancer treatment may be modulated by gene expression profiles at the tumor cell. This interaction can be accommodated by adding to equation [4] a new set of random effects contrasts between the major factor and each of the small-risk factors. Thus, if the k^{th} fixed effect, x_{ik}, interacts with predictors in the set W, the set of contrasts will be $W_{x_k} = \{x_{ik} \times w_{i1}, x_{ik} \times w_{i2} \dots, x_{ik} \times w_{ip_w}\}$. The new term will be $x_{ik} \times \sum_{j=1}^{j=p_w} w_{ij}\gamma_{wx_{kj}}$ and the prior density described in [5] will also include following term: $\prod_{j=1}^{j=p_w} p(\gamma_{wx_kj}|\Omega_{wx_k})p(\Omega_{wx_k})$.

Case 2. A second case is the interaction between two high-dimensional sets. If the number of predictors in the set is $p_{.}$, the number of contrasts for all possible 1st order interactions is $p_{.} \times (p_{.} - 1)/2$. Adding in the models all these possible contrasts is unfeasible. Relevant interactions could be found with search algorithms. A review of existing methods for GWAS can be found in (Cordell, 2009). However, search algorithms may fail to encounter a reasonably good model in a set of models too large. For instance, stepwise methods have problems modeling response surfaces in high complexity models (Friedman and Stuetzle, 1981). Alternatively, non-parametric models, such us neural networks and kernel methods (**RKHS**, e.g., Gianola et al., 2006), capture departures from linearity. Empirical evidence have suggested that kernel methods, particularly kernel averaging, (de los Campos et al., 2010) can be very effective (Heslot et al., 2012). To extend models described in [4] and [5] incorporating high dimensional interactions could be achieved by adding a random effect with co-variance structure given by the kernel. The BGLR package (de los Campos and Perez, 2013) allows inclusion of both parametric and non-parametric components in the same model.

Case 3**,** involves interactions between predictors in two high dimensional sets. For examples, we may want to model interactions between predictors in W and those in Z. Here we face the problem of the number of terms in two high dimensional input sets. For instance, we may want to model first order interactions between predictors the total number of contrasts needed, which is equal to the dimension of each of the sets (e.g., $p_W \times p_Z$) can be extremely large. Again, here at least two strategies are possible, one is to search for interactions using model search algorithms, and the other is to accommodate departures from the additive model using semi-parametric methods such as RKHS. The kernel function in this case would depend on a weighted sum of the distance between individuals in set W and set Z; and the tuning of bandwidth parameters will need to be addressed.

AN APPLICATION EXAMPLE

In this section we present a simple example designed to assess the potential benefits of using a whole-genome approach in risk-assessment models for prediction of breast cancer (**BC**) outcomes. Advances in early detection and in adjuvant therapy have reduced mortality. However, adjuvant therapy has important undesirable side effects on treated patients, including permanent infertility, heart damage, cognitive impairment, and increased probability of developing other types of cancers (Eifel et al., 2001). There is a great deal of variability in the aggressiveness of BC tumors, and it has been estimated that in 60% of patients BC will not recur (Weigelt et al., 2005). However, due to lack of models that can accurately predict BC progression, approximately 80% of BC patients are treated with adjuvant therapy, leading to important undesirable health effects and even deaths attributable to harmful long-term effects of the treatment.

Gene expression levels at the tumor cell are predictive of cancer sub-type and this information can be used to inform medical decisions. Existing methods include dimension reduction approaches (Sørlie et al., 2001) or variable selection methods (Paik et al., 2004; Glas et al., 2006), all applied to expression profiles of a relatively small set of large-effect genes. Despite important advances in the use of gene expression profiles for classification of BC tumors (Sørlie et al., 2001) a large proportion of inter-individual differences in BC progression remain unaccounted for, making it difficult to implement a personalized approach to the BC treatment. In the example presented here, we developed models to assess the potential benefits of moving from prediction models based on a small set of genes towards the use of whole-transcriptome profiles.

Data

Data (N=186) was obtained from the TCGA and the outcome predicted was five-year survival (yes=157; no=29) of non-metastatic breast cancer patients, having estrogen receptor positive tumors at stage I or II. There is an FDA-approved diagnostic assay (Oncotype DX, Genomic Health Inc, Redwood City, CA, Cronin et al., 2004; Paik et al., 2004) based on the expression profiles of 21 genes, which after QC edition, we found 17 of the genes in the Oncotype assay. Later, we assessed the benefits of including the expression profiles of the genes in Oncotype diagnostic assay versus 17,899 genes that are not included in the standard Oncotype assay. In addition to gene expression profiles we also included age at diagnosis and ethnicity (Caucasian, African American, Hispanic and Others) of each patient.

Models

We begin by fitting baseline models based on covariates only (**COV** included age at diagnostic and ethnicity as the only predictors). This model was expanded by adding the random effects of the genes in the Oncotype panel (**ONCO**). Finally, we expanded the ONCO model by adding the random effects of 17,899 genes not included in the Oncotype panel (**WGGE**,=whole-genome gene expression). All models were fitted using the BGLR package (de los Campos and Perez, 2013). The response was treated as binary, using the probit link implemented in BGLR and random effects were assigned *iid* normal priors with separate variance components for the genes in the Oncotype assay and the rest of the genes not included in that panel.

Analysis

Variance components were estimated with the models above described to the full data set. Subsequently we assessed prediction accuracy of the model with the Area Under the Curve in ten-fold cross-validation, (**AUC-CV**) with random assignment of cases and controls to folds.

Results

Results are given in Table 1, when the oncogene expression profiles were included in the model (ONCO) they explained 9.4% of the variance of risk un-explained by demographics. This result confirms the association between the genes in the Oncotype test with BC progression, but, at the same time, it suggests that a large proportion of the variance in risk remains un-explained. When we included the GE profiles of all the genes available (WGGE model) the proportion of variance in risk explained increased to 27.4 suggesting that indeed a sizable proportion of the variance of risk that was not explained by demographics and genes in Oncotype DX could be explained by integrating in the model genome-wide GE profiles. Importantly, our point estimates suggest that information from genes not included in the Oncotype test explained twice as much (18.9%) than those in the Oncotype (8.7%). Note, however, that all the credibility regions (in square brackets in the table) are wide, reflecting the small sample size used. Clearly, these results need to be confirmed with larger sample size. The AUC-CV of the COV model was 0.71, adding GE of genes in the Oncotype increased AUC-CV by a very small amount, finally when all GE profiles were jointly considered AUC-CV increased from 0.71 to 0.75.

Table 1. Results from preliminary data analysis

Model	Variance Component (%)[1]		AUC
	Oncogenes[2]	Other[3]	
COV	---	---	0.71
ONCO	9.4 [3.3;22.4]	---	0.72
WGGE	8.7 [3.0;21.0]	18.9 [5.0;46.9]	0.75

1: % of the variance in risk (liability scale), after accounting for by demographics and [95% credibility region].
2: Oncogenes: Genes in the Oncotype array (17)
3: Other: expression of genes (17,899) not in Oncogenes group.
AUC: Area under the curve; COV: Models with covariates only; ONCO: Extends COV model by accommodating the genes in the Oncotype panel; WGGE: extends COV by incorporating whole genome gene expression.

Discussion

Estimates of variance components suggest that a sizable proportion of inter-individual differences in risk that were not explained by demographics and GE profiles of genes in the Oncotype could be captured by adding to the model genome-wide RNA sequencing data. Prediction accuracy also increased when WGGE profiles were considered. The increase in AUC-CV was modest and this is likely due to small sample size. We expect that with larger sample size AUC-CV of both ONCO and WGGE will increase markedly, especially WGGE.

CONCLUDING REMARKS

The availability of multi-layer omic data has increased in recent years and is expected to increase in years to come. This data has potential to advance our ability to predict health outcomes. However, the development of statistical methods for integration of multi-layer omic data into risk assessment methods lags behind. The majority of the methods used are either based on a limited number of risk factors on dimension reduction approaches. We argue that higher prediction accuracy could be obtained with full integration of whole-genome-multi-layer omic profiles into risk assessment models.

Whole-genome regression models were originally developed for prediction of genetic values using information from the genome (SNPs). These methods can be extended to accommodate additive effects of multi-layer omic data. However, due to the high-dimensional nature of omic data, accommodating interactions between risk factors represent a major challenge. Here we have discussed some parametric and semi-parametric approaches that can be used to handle some of those challenges.

Some omic profiles (e.g., the transcriptome) vary across space (tissue) and time; therefore, predictions of outcomes of diseases that are tissue-specific such as cancer are likely to have the most potential. The example presented in this article, although preliminary, suggests that integration of whole-genome profiles of gene expression can lead to risk assessment models with predictive power above and beyond only considering the expression profiles of large-effect genes.

Although multiple methodological and empirical questions remain open, we believe that the integration of multi-layer omic profiles into risk assessment models represents a promising approach and one that merits more research efforts.

ACKNOWLEDGMENTS

The authors thank The Cancer Genome Atlas Network for data access. AIV acknowledges financial from the School of Public Health at the University of Alabama at Birmingham, Back of The Envelope awards. GDLC acknowledges financial support from NIH grants GM099992 and GM101219.

LITERATURE CITED

Allen HL, Estrada K, Lettre G, et al. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature Letter, doi:10.1038/nature09410:1–7.

Anon. 2003. Human genome finally complete. BBC. Available at from: http://news.bbc.co.uk/2/hi/science/nature/2940601.stm

Berghoff BA, Konzer A, Mank NN, et al. 2013. Integrative "omics"-approach discovers dynamic and regulatory features of bacterial stress responses. PLoS genetics 9:e1003576.

Bureau A, Dupuis J, Falls K, et al. 2005. Identifying SNPs predictive of phenotype using random forests. Genetic Epidemiology 28:171–182.

de los Campos G, Gianola D, Rosa GJM, et al. 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genetics Research 92:295–308.

de los Campos G, Hickey JM, Pong-Wong R, et al. 2013a. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–345.

de los Campos G, Klimentidis YC, Vazquez AI, et al. 2012a. Prediction of Expected Years of Life Using Whole-Genome Markers. PloS one 7:e40964.

de los Campos G, Perez P. 2013. BGLR: Bayesian Generalized Linear Regression. Available from: http://cran.at.r-project.org/web/packages/BGLR/index.html

de los Campos G, Vazquez AI, Fernando R, et al. 2013b. Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor. PLoS Genetics 9:e1003608.

Chen H, Poon A, Yeung C, et al. 2011. A Genetic Risk Score Combining Ten Psoriasis Risk Loci Improves Disease Prediction. PLoS ONE 6:e19454.

Chen R, Mias GI, Li-Pook-Than J, et al. 2012. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 148:1293–1307.

Chin L, Hahn WC, Getz G, et al. 2011. Making sense of cancer genomic data. Genes Dev 25:534–555.

Cordell HJ. 2009. Detecting gene–gene interactions that underlie human diseases. Nat Rev Genet 10:392–404.

Cronin M, Pho M, Dutta D, et al. 2004. Measurement of gene expression in archival paraffin-embedded tissues: development and performance of a 92-gene reverse transcriptase-polymerase chain reaction assay. Am J Pathol 164:35–42.

Crossa J, de los Campos G, Perez P, et al. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713.

de los Campos, Vazquez A, Sorensen D. 2012b. Prediction of Complex Human Traits Using Genomic BLUP. Poultry Round Table.

Dominiczak AF, McBride MW. 2003. Genetics of common polygenic stroke. Nat Genet 35:116–117.

Eifel P, Axelson JA, Costa J, et al. 2001. National Institutes of Health Consensus Development Conference Statement: adjuvant therapy for breast cancer, November 1-3, 2000. Journal of the National Cancer Institute 93:979.

Forer L, Schönherr S, Weissensteiner H, et al. 2010. CONAN: copy number variation analysis software for genome-wide association studies. BMC bioinformatics 11:318.

Friedman JH, Stuetzle W. 1981. Projection Pursuit Regression. Journal of the American Statistical Association 76:817–823.

Gianola D, Fernando RL, Stella A. 2006. Genomic-Assisted Prediction of Genetic Value With Semiparametric Procedures. Genetics 173:1761–1776.

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

Gianola D. 2013. Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. Genetics 194: 573-596.

Glas AM, Floore A, Delahaye LJ, et al. 2006. Converting a breast cancer microarray signature into a high-throughput diagnostic test. BMC Genomics 7:278.

Goddard ME, Hayes BJ. 2007. Genomic selection. Journal of Animal Breeding and Genetics 124:323–330.

Habier D, Tetens J, Seefried FR, et al. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics Selection Evolution 42:5.

Heslot N, Yang H-P, Sorrells ME, et al. 2012. Genomic Selection in Plant Breeding: A Comparison of Models. Crop Science 52:146.

Ishwaran H, Rao JS. 2005. Spike and slab variable selection: frequentist and Bayesian strategies. The Annals of Statistics 33:730–773.

De Jager PL, Chibnik LB, Cui J, et al. 2009. Integration of genetic risk factors into a clinical algorithm for multiple sclerosis susceptibility: a weighted genetic risk score. The Lancet Neurology 8:1111–1119.

Lander ES, Linton LM, Birren B, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Loos RJF. 2009. Recent progress in the genetics of common obesity. Br J Clin Pharmacol 68:811–829.

Maher B. 2008. The case of the missing heritability. Nature 456:18–21.

Makowsky R, Pajewski NM, Klimentidis YC, et al. 2011. Beyond Missing Heritability: Prediction of Complex Traits. PLoS Genet 7:e1002051.

Manolio TA, Collins FS, Cox NJ, et al. 2009. Finding the missing heritability of complex diseases. Nature 461:747–753.

Meuwissen TH, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Okser S, Lehtimäki T, Elo LL, et al. 2010. Genetic variants and their interactions in the prediction of increased pre-clinical carotid atherosclerosis: the cardiovascular risk in young Finns study. PLoS genetics 6:e1001146.

Paik S, Shak S, Tang G, et al. 2004. A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. New England Journal of Medicine 351:2817–2826.

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

Pérez-Cabal MA, Vazquez AI, Gianola D, et al. 2012. Accuracy of genome-enabled prediction in a dairy cattle population using different cross-validation layouts. Frontiers in Genetics 3.

Purcell SM, Wray NR, Stone JL, et al. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460:748–752.

Resende MD, Resende MF, Sansaloni CP, et al. 2012. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytologist 194:116–128.

Rodin AS, Boerwinkle E. 2005. Mining genetic epidemiology data with Bayesian networks I: Bayesian networks and example application (plasma apoE levels). Bioinformatics 21:3273–3278.

Sebastiani P, Perls TT. 2010. Prediction models that include genetic data. Circulation: Cardiovascular Genetics 3:1–2.

Sørlie T, Perou CM, Tibshirani R, et al. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences 98:10869–10874.

Speed D, Hemani G, Johnson MR, et al. 2012. Improved Heritability Estimation from Genome-wide SNPs. The American Journal of Human Genetics 91:1011–1021.

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

Vazquez AI, de los Campos G, Klimentidis YC, et al. 2012. A Comprehensive Genetic Approach for Improving Prediction of Skin Cancer Risk in Humans, Genetics. 192: 1493–1502.

Vazquez AI, Rosa GJM, Weigel KA, et al. 2010. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. Journal of Dairy Science 93:5942–5949.

Venter JC, Adams MD, Myers EW, et al. 2001. The sequence of the human genome. science 291:1304–1351.

Wei Z, Wang K, Qu H-Q, et al. 2009. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS genetics 5:e1000678.

Weigelt B, Peterse JL, van't Veer LJ. 2005. Breast cancer metastasis: markers and models. Nature reviews cancer 5:591–602.

Yang J, Benyamin B, McEvoy BP, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42:565–569.