RESEARCH ARTICLE

Assessment of Whole-Genome Regression for Type II Diabetes

Ana I. Vazquez[1]*, Yann C. Klimentidis[2], Emily J. Dhurandhar[3], Yogasudha C. Veturi[4], Paulino Paérez-Rodríguez[5]

1 Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, United States of America, **2** Mel and Enid Zuckerman College of Public Health, Division of Epidemiology and Biostatistics, University of Arizona, Tucson, AZ, United States of America, **3** Department of Health Behavior, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, United States of America, **4** Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, United States of America, **5** Colegio de Postgraduados, Montecillo, Edo. de Mexico, Mexico

* avazquez@uab.edu

Citation: Vazquez AI, Klimentidis YC, Dhurandhar EJ, Veturi YC, Paérez-Rodríguez P (2015) Assessment of Whole-Genome Regression for Type II Diabetes. PLoS ONE 10(4): e0123818. doi:10.1371/journal.pone.0123818

Academic Editor: Chuhsing Kate Hsiao, National Taiwan University, TAIWAN

Received: October 9, 2014

Accepted: March 7, 2015

Published: April 17, 2015

Data Availability Statement: All files are available from dbGaP or Database of Genotypes and Phenotypes (accession number pht000040.v4.p7, pht000041.v4.p7 and pht000311.v5.p7)

Funding: The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. Funding for SHARe Affymetrix genotyping

Abstract

Lifestyle and genetic factors play a large role in the development of Type 2 Diabetes (T2D). Despite the important role of genetic factors, genetic information is not incorporated into the clinical assessment of T2D risk. We assessed and compared Whole Genome Regression methods to predict the T2D status of 5,245 subjects from the Framingham Heart Study. For evaluating each method we constructed the following set of regression models: A clinical baseline model (**CBM**) which included non-genetic covariates only. CBM was extended by adding the first two marker-derived principal components and 65 **SNPs** identified by a recent GWAS consortium for T2D (**M-65SNPs**). Subsequently, it was further extended by adding 249,798 genome-wide SNPs from a high-density array. The Bayesian models used to incorporate genome-wide marker information as predictors were: Bayes A, Bayes Cπ, Bayesian LASSO (**BL**), and the Genomic Best Linear Unbiased Prediction (**G-BLUP**). Results included estimates of the genetic variance and heritability, genetic scores for T2D, and predictive ability evaluated in a 10-fold cross-validation. The predictive AUC estimates for CBM and M-65SNPs were: 0.668 and 0.684, respectively. We found evidence of contribution of genetic effects in T2D, as reflected in the genomic heritability estimates (0.492± 0.066). The highest predictive AUC among the genome-wide marker Bayesian models was 0.681 for the Bayesian LASSO. Overall, the improvement in predictive ability was moderate and did not differ greatly among models that included genetic information. Approximately 58% of the total number of genetic variants was found to contribute to the overall genetic variation, indicating a complex genetic architecture for T2D. Our results suggest that the Bayes Cπ and the G-BLUP models with a large set of genome-wide markers could be used for predicting risk to T2D, as an alternative to using high-density arrays when selected markers from large consortiums for a given complex trait or disease are unavailable.

was provided by NHLBI Contract N02-HL-64278. SHARe Illumina genotyping was provided under an agreement between Illumina and Boston University. AIV acknowledges National Institute of Health grant, from National Institute of Diabetes and Digestive and Kidney, Research Supplement to 7-R01-DK-062148-10, for their financial support. YCK acknowledges support from NIH grant K01DK095032. PPR had financial support from NIH Grants R01GM099992 AND R01GM101219. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

While the Human Genome Project provides a detailed description of genetic variation, the causal genes for many diseases are yet to be found. Although some variants have been found to be directly causal and to increase the risk of disease, the significant variants associated with complex traits have been found to explain only a small percentage of the phenotypic variation. This problem has been referred to as the 'missing heritability' of complex traits and diseases [1,2]. However, recent studies show that by using information from hundreds of thousands of loci with Whole Genome Regression (**WGR**), all the heritability in family-based data[3,4] and half of the expected heritability in complex traits can be explained [4–6].

Of late, there is an increasing interest in estimating the variance in complex traits that is explained by molecular markers, with the ultimate goal of obtaining accurate estimates of individual genetic predisposition to disease. With Whole Genome Prediction (**WGP**), genetic risk is modeled using thousands of (small-effect) loci concurrently [5,7]. The foundational idea of these whole-genome methods was published by Meuwissen et al. and it has since brought about a revolution in the animal and plant breeding communities in both academia and industry [8–10]. Modeling genetic risk for human disease by regressing high-density-SNP arrays on phenotypes is feasible with the use of penalized and Bayesian variable selection [11,12] and shrinkage estimation methods [13–15]. Since 2001, WGR has been extensively used [5,6,16–18] to estimate genetic parameters [3,6,19] and, more recently, to predict genetic risk [4,18,20].

In samples of related humans, an increment in the number of markers used by the model increases the genetic variance explained and monotonically increases the prediction accuracy [4]. However, in samples of unrelated humans, while variance explained can reach up to 50% of the genetic heritability [5,7], there is an optimum number of SNPs at which prediction accuracy is maximized [21]. Therefore, in unrelated samples of humans, relatively poor prediction accuracy of disease risk is achieved. The span of linkage disequilibrium (LD) is much shorter in humans compared to domestic agricultural species (e.g., cattle [22]), thus genetic markers cannot correctly estimate genomic relationships and the statistical model cannot separate genetic signal from random variation [21]. Consequently, empirical studies show that while the prediction R-squared in validation samples (for human height) is approximately 0.3 in a family-based sample, it is only about 0.03–0.05 in unrelated individuals [21]. In summary, in family-based samples, the predictive ability is higher if the model is informed by relatives who share large sections of the chromosome with the individuals to be predicted [4,18]. There, a SNP that is distant from a causal locus can still be highly informative of the genetic risk of disease. In short, prediction in related and unrelated individuals are two different problems. Thereby, for complex human traits in unrelated subjects, it may be important to reduce the noise from the genotype by targeting regions of causal loci. Whole Genome Regressions are a large family of methods that can either differentiate genetic regions or weight the entire genome equally. However, how different WGRs work for prediction of unrelated subjects has not been completely addressed yet.

WGRs differ in the priors assigned to the marker effects and in their ability to perform selection and shrinkage of predictors. Some WGRs (e.g., G-BLUP) have an underlying assumption that all predictors have some small effect, with genetic risk being determined by a very large number of variants. This implies that the trait (e.g. human height) has a highly complex genetic architecture [5]. Other 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; examples include Bayes A [16] and the Bayesian LASSO (BL) [15]. Finally, finite-mixture priors assign a certain prior probability for the effects to be equal to zero. These priors—for example, Bayes

Cπ [12,23]—induce variable selection and shrinkage simultaneously and work best for traits whose genetic architectures include regions that do not contribute to genetic risk at all.

In this article we aim to evaluate several Bayesian models, including BL, Bayes A, Bayes Cπ, and G-BLUP that perform differential shrinkage and variable selection. We focus on Type 2 Diabetes Mellitus (**T2D**) since it is the fastest growing chronic disease in the developing world [24]. A complex interaction between lifestyle factors and genetics (h^2 between 0.25–0.70 in family and monozygotic twins) plays a large role in the development of T2D [25–29]. Additionally, for T2D, several studies report highlights of the genetic architecture of T2D by uncovering several SNP variants [30–32], and recently 65 SNPs have been associated with T2D [30]. In our study, we included a benchmark model with these well-established SNPs to evaluate the performance of the Bayesian methods.

Materials and Methods

Ethics Statement

The FHS obtained informed consent from the participants to use their clinical records for research purposes such as this study. Additionally, we obtained the data from dbGap, where data is de-identified before being distributed to other researchers.

Data

Phenotypes. The data set consists of 5,245 participants (2,381 males and 2,864 females) from the Framingham Heart Study, which has collected longitudinal phenotypic information in several generations of families [33,34]. Subjects in this study have been characterized every other year from adulthood to death on risk factors, outcomes of physical exams, and disease status. T2DM was defined as having blood sugar $\geq$ 126 mg/dL, at any exam, or having ever taken anti-diabetic medication. The studies used have the dbGaP (database of Genotypes and Phenotypes) accession number pht000040.v4.p7, pht000041.v4.p7 and pht000311.v5.p7. Participants included in our study belong to the Original cohort (n = 1,498), and the cohort that is comprised of their Offspring (n = 3,747). We excluded subjects from the third generation cohort because their follow up time is still too short. Data and material distributions from this study are made in accordance with the individual consent history of each participant and the current study has been approved by the Internal Review Board of University of Alabama at Birmingham (IRB Protocol Number: X090720002).

Genomic Information. All subjects were genotyped for single nucleotide polymorphisms (SNPs) with the Affymetrix 500K chip. Details on the genotyping method are described at the Framingham SHARe at the NCBI dbGaP website (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id = phs000007.v3p2). We removed markers with minor allele frequency$<$ 0.05 or with more than 10% missing genotypes. After edition, we randomly reduced the platform to approximately half of the SNPs (p = 249,798) to attenuate computational demand. This platform included 20 SNPs of the 65 SNPs previously published. Pre-analysis including and excluding the 65 markers suggests that including or excluding them in WGR does not vary WGR results.

Statistical Methods

The outcome $\boldsymbol{y} = \{y_i\}$ was defined as presence ($y_i = 1$) or absence ($y_i = 0$) of T2D (blood sugar $>$ 126 mg/dL, or having ever taken an anti-diabetic medication) during the follow up time of the FHS. We assessed several models including WGRs with various types of Bayesian approaches that differ in the selection and shrinkage applied to the marker effects. This section is organized as follows: (a) the description of the probit link connecting the response variable (diabetes presence or absence) with a linear predictor (η_i); (b) the sequence of models developed; (c) the Bayesian statistical models evaluated, (d) estimates of genetic effects associated to markers, and (e) model evaluation tools.

a) Probit Link. Let y_i be the random variable that denotes the presence or absence of diabetes and define ϑ_i (i.e. the probability of having diabetes) as $\vartheta_i = P(y_i = 1|\boldsymbol{x}_i)$. It follows that y_i is distributed as a Bernoulli random variable with probability of success given by ϑ_i (ϑ_i) a subject-specific Bernoulli parameter). In probit regression, the probability of success depends on a set of covariates ($\boldsymbol{x}_i{}'s$) and is modeled as,

$$\vartheta_i = \Phi(\eta_i),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and (η_i) is a model dependent linear predictor that will be described next.

b) Sequence of statistical models. *Covariates baseline model* (**CBM**). The CBM included non-genetic covariates only. The linear predictor for CBM is

$$\eta_i = \alpha_0 + \alpha_1 s_i + \alpha_2 c_i + \alpha_3 l_i$$

Where η_i is represented as the sum of an intercept(α_0), plus a regression on the 'fixed effects' of sex (s_i, as dummy variable), cohort (c_i, a dummy variable indicating whether participants were from the Original or Offspring cohort), the age at last contact or death (l_i, ranging from 34 to 104) to control for different exposure times or observational periods, and $\boldsymbol{\alpha} = (\alpha_1,\ldots, \alpha_3)'$, are the corresponding regression coefficients. The sample from FHS includes subjects from two cohorts and each cohort starts at a different year, a few of the measures could have different protocols and a different data collection team. We included cohort information in the models to correct for these factors.

65-SNP Model (**M-65SNP**). The CBM model was first extended by adding 2 marker-derived principal components (*PC1* and *PC2*) and 65 SNPs that have been consistently associated with T2DM. The PCs were derived from 1,000 European-ancestry informative markers reported by [35]. This model is a second benchmark or baseline model to compare with WGR. The linear predictor could then be expressed as,

$$\eta_i = \alpha_0 + \alpha_1 s_i + \alpha_2 c_i + \alpha_3 l_i + \alpha_4 PC1_i + \alpha_5 PC2_i + \sum\nolimits_{j=1}^{65} \gamma_j x_{ij},$$

where α_4 and α_5 are regression coefficients associated with PCs 1 and 2 respectively; x_{ij} is the genotype of the i^{th} individual ($i = 1,\ldots, 5,245$) at the j^{th} marker ($j = 1,\ldots,65$), expressed as the count of one of the two alleles $x_{ij} \in \{0,1,2\}$, for the imputed SNPs $x_{ij} \in [0,2]$ (a real number) and the $\boldsymbol{\gamma} = \{\gamma_i\}$'s are marker effects. When absent in the platform, these SNPs were imputed with IMPUTE2 with 1,092 subjects from the 1,000 Genomes data as reported previously [36–38].

c) Bayesian models for Whole Genome Regressions. Subsequently, we evaluated several WGRs using the CBM as the base and included the genomic effects (u_i) modeled with whole-genome markers. These were comprised of a high density array of p = 249,798 SNPs and were regressed on a function of the phenotype evaluated in this study. SNP effects were included in the models using either Bayes A, Bayes Cπ, Bayesian LASSO (BL), and G-BLUP. The linear predictor for these models could be written in general as,

$$\eta_i = \alpha_0 + \alpha_1 s_i + \alpha_2 c_i + \alpha_3 l_i + \alpha_4 PC1_i + \alpha_5 PC2_i + u_i$$

In addition to the joint conditional probability of the data, given the unknown coefficients, the prior density of the unknowns was flat for $\boldsymbol{\alpha}$, i.e. $p(\boldsymbol{\alpha}) \propto 1$. This yields estimates of effects comparable to those obtained with maximum likelihood. The genomic effect term u_i is different in every one of the Bayesian models evaluated. The definition of u_i and its prior probability completes the Bayesian model. We describe them below for each Bayesian model evaluated.

*Bayes A (**BA**)*. In Bayes A models [16], $u_i = \sum_{j=1}^{p} x_{ij}\beta_{ij}$ and the prior density of the SNPs effects is assumed to follow a t distribution, $T(\beta_j|df_\beta, S_\beta)$, ($j$ = 1,. . .,65), which could be re-expressed as $\int N(\beta_j|0, \sigma^2_{\beta_j})\chi^{-2}(\sigma^2_{\beta_j}|df_{\beta_j}, S_{\beta_j})\partial\sigma^2_{\beta_j}$ where $\sigma^2_{\beta_j}$ is the variance of the marker effects corresponding to the j^{th} position; see [39]. Thus, the conditional distribution of marker effects β_j is normal with mean 0 and variance $\sigma^2_{\beta_j}$, at the next level of hierarchy we assigned a scaled-inverse chi squared distribution to the variance of marker effects. The corresponding hyper-parameters for the scaled-inverse chi-squared distribution were set according to the rules given in [40] and implemented in the BGLR package.

*Bayes C π (**BC**)*. In Bayes C models [12], $u_i = \sum_{j=1}^{p} x_{ij}\beta_{ij}$ and here the prior for the marker effects is a two component mixture. One of the components is a point of mass at zero and the other component is a normal distribution. The prior for the marker effect for this model can be written as,

$$\mathrm{p}(\beta_j|\pi, \sigma^2_\beta) = \pi \times 1(\beta_j = 0) + (1 - \pi) \times \mathrm{N}(\beta_j|0, \sigma^2_\beta)$$

where π is the proportion of markers with non-null effects and the prior assigned to π was *Beta* (p_0, π_0) (see[41]). We assigned a scaled-inverse chi-squared distribution to σ^2_β, the corresponding hyper parameters were set using the rules given in de los Campos et al. (2013).

*Bayesian LASSO (**B-LASSO**)*. In the Bayesian LASSO [15], $u_i = \sum_{j=1}^{p} x_{ij}\beta_{ij}$ and the prior density of the SNPs effects can be expressed as $N(\beta_j|0, \tau^2_j)$, where the prior distribution of τ^2_j is exponential, i.e. $Exp(\tau^2_j\lambda^2)$ and the prior density assigned to λ^2 is a gamma distribution, $G(\lambda^2 \mid \delta_1, \delta_2)$, with the hyper-parameter rate set to 0.0001 and shape 0.55; (For further details on priors for this model see [42]).

G-BLUP. In this model [43], $\boldsymbol{u} = \{u_i\}$ is a random effect in the regression which distribution is $N(\boldsymbol{u}|\boldsymbol{0}, \boldsymbol{G}\sigma^2_u)$ where $\boldsymbol{G} = \{G_{ii'}\}$ is an $n \times n$ matrix of relationships based on the p SNP genotypes such that,

$$G_{ii'} = \frac{1}{n} \times \frac{\sum_{j=1}^{p}(x_{ij} - 2q_j)(x_{i'j} - 2q_j)}{2\sum_{j=1}^{p} q_j(1 - q_j)},$$

where q_j is the estimated j^{th} allele frequency; and σ^2_u is an 'additive' genetic variance parameter. We assigned a scaled-inverse chi-squared distribution to σ^2_β, the corresponding hyper parameters were set using the rules given in de los Campos et al. (2013). The marker effects were obtained with the equivalent Bayesian Ridge Regression model, as described elsewhere [40].

The parameters of the above-described model were estimated in a Bayesian framework using the BGLR package[41,44] in R [45]. Priors used were relatively non influential [41]. We used 40,000 MCMC iterations with 15,000 samples taken as burn in. Convergence was assessed by visual inspection of the trace plots, e.g. S1 Fig and S2 Fig.

Genomic Heritability

In the models described above, narrow sense heritability in the liability scale is defined as the ratio of the genetic variance to the total variance. The residual variance is fixed at one, thus the narrow sense heritability is $h^2_G = \frac{\sigma^2_u}{\sigma^2_u + 1}$, since the residual variance is set to one as defined elsewhere for binary traits[46,47]. The genomic heritability is interpreted as the proportion of inter-individual differences at risk for T2D that can be explained by regression on common SNPs.

Model Evaluation

The covariates included in all models were selected based on significance. This evaluation was done with the generalized linear model (*glm* function) from the R base package [45]. Models were compared based on effect estimates from published GWAS. We present distribution of effects and scatter plots of these estimates for each model. Additionally, we assessed the models' prediction accuracy using a 10-fold cross-validation. Since Framingham is a family based study, we randomized and assigned entire families, according to the pedigree, to folds such that when the model is trained, neither the subject to be predicted nor the subjects in the same fold —which include all subjects in one family—are used to fit the predictive model. The testing sets of the cross-validation yielded predictions of risk scores $\{\hat{\eta}_i\}$, which were derived without using the i^{th} observation or any relative of the i^{th} observation. AUC was computed using the pairs of points that included the presence/absence of diabetes and the risk score was predicted using cross validation $\{y_i, \hat{\eta}_i\}$. Since realization of diabetes (y_i) is a binary response (0/1), it is more appropriate to report results in terms of false positive rate and Area Under the Receiver Operating Characteristic Curve (**AUC**, see [48]). We estimated the former statistics using the R package ROCR [49].

Results

Descriptive Statistics

Participants in the two cohorts that were included in this study were born between 1890 and 1968, and the age at either last contact or death was 74±12 (mean ± standard deviation). The total number of T2D cases was 939 out of the 5,245 subjects; since FHS is an ongoing study, many study participants do not have diabetes yet. The incidence of diabetes in participants with last contact at age <65 was 8%, while it was 26% for participants with last contact at age > 83 years. The 939 cases showed a first record of diabetes at 63±24 years of age [50]. The proportion of subjects in the population with diabetes between cohorts was 30.2% in the Original cohort and 13.0% in the Offspring cohort (paternal and offspring generations, respectively). This difference reflects the shorter observational time of the Offspring cohort, whose subjects have an age at the last contact time (or death time) of 69±10, whereas this age is 87±8 for the Original cohort. The proportion of people that had diabetes was 20.5% in males (45% of the sample) and 15.7% in females (55% of the sample). This difference in proportions is in accordance with what has been observed in the literature where males have a higher incidence of diabetes [50]. In the study we also included principal components derived from ethnicity informative SNPs to account for population structure; Fig 1 shows a scatter plot with the ethnicity-informed marker-derived PC 1 and 2.

Estimates of Fixed Effects and Odd Ratios for CBM

The effects estimated in CBM are on a liability scale. The estimated effect of sex (being female) was -0.26± 0.04 (p-value<1e-6), with an odds ratio of 0.77, implying a decreased risk of developing diabetes in females, relative to males. The estimated effect of the Offspring cohort was -0.40±0.06 (p-value<1e-6), with an odds ratio of 0.68, implying a lower risk of diabetes for subjects on the Offspring cohort relative to subjects of the Original cohort. This estimate is probably affected by the shorter observational period of the members of the Offspring cohort, even though the age at the last contact was included in the study to correct for

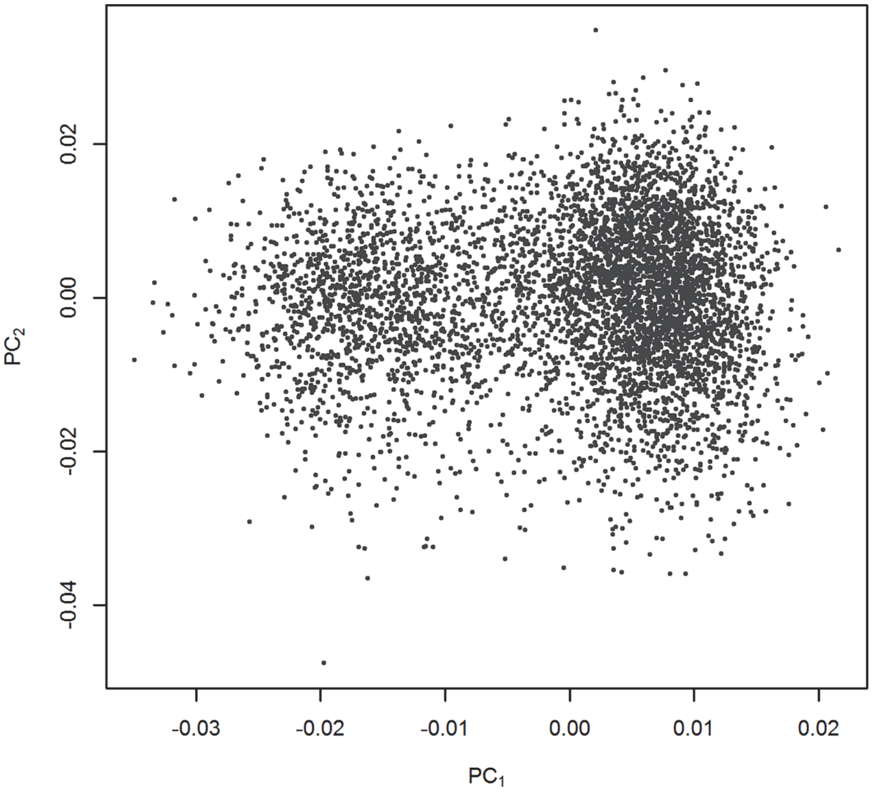

Fig 1. Principal Components 1 and 2, derived from 1,000 ethnicity informative SNPs for European origin.

doi:10.1371/journal.pone.0123818.g001

different censoring times. The estimated effect for the age at the last contact was 0.016±0.002 (p-value <1e-6) with an odds ratio of 1.02, implying that there is a slightly higher risk of observing diabetes per year of exposure. Finally, PCs 1 and 2 were non-significant with odds ratios of 1.25±1.89 (p-value = 0.57) and 1.18±2.06 (p-value = 0.57), respectively. However, in order to correct for population stratification, we retained the principal components in the model.

SNP Effects, Genetic Effects and Genetic Parameters

Odds ratio for the 65 SNPs obtained with the M-65SNP model and odds previously reported in the literature [30] have a correlation of 0.42. Genetic scores were calculated for all subjects with the different models. The genetic risk score derived with the G-BLUP model ranged from -0.980 to 1.766 and was centered at -0.003±0.498 (mean ± standard deviation). The genetic risk score with G-BLUP, and M65-SNPs had a positive but weak association of 0.23. The much higher amount of SNPs between the M-65SNPs and the G-BLUP brings different information to the prediction of genetic scores.

Fig 2 shows the estimated SNP effect with G-BLUP model (re parameterized as a Bayesian Ridge Regression) and the dots highlight where the effects of the 65 SNPs of the M-65SNPs model are. It can be observed that the highly significant markers were not always the SNPs with the greatest effect. Models BC and BLASSO produced estimated effects of the SNP with similar shrinkage than one observed in Fig 2; BA however had convergence problems in the full data analysis.

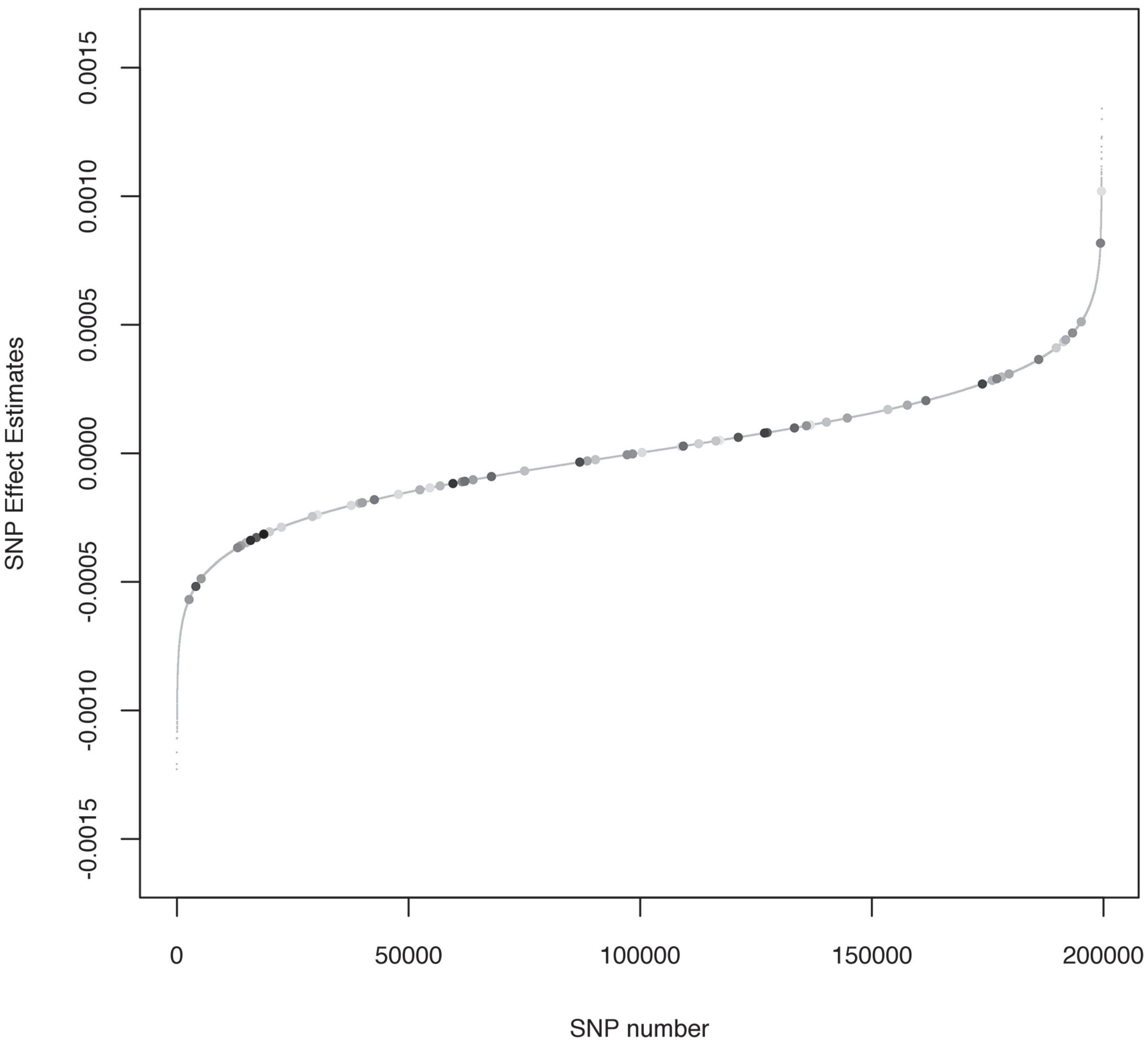

Fig 2. SNP estimated effects ordered by effect for G-BLUP. This is re-parameterized with a Bayesian Ridge Regression. Dots show the effects of the 65 SNPs are and are on a gray scale; the darker the dot, the more significant is its association with the response.

doi:10.1371/journal.pone.0123818.g002

Fig 3 shows the relationship between the probability of having diabetes for healthy (a) and diabetic (b) participants. Estimates in the scatter plots were obtained from the G-BLUP model and M-65SNPs. Other WGR (not shown) presented a strong association with the probabilities estimated with the G-BLUP. The line is a slope of one, indicating the hypothetical situation in which both models estimate the same probability of having diabetes for a person. M-65SNPs is based on fixed-effect estimates. For this model, we observe that each SNP has stronger effect than the same SNP in the G-BLUP model because a fixed effect is a value close to the conditional mean of all the samples with that SNP (jointly with deviations given by other factors); while the RKHS predicts random effects where the SNP effect is pushed towards zero. The resulting estimates for a person that has all the high-risk SNPs (of the 65) accumulate into a very large estimated overall risk, while in the RKHS, the estimated effects of these SNPs are mitigated and predicted risk is thus lower.

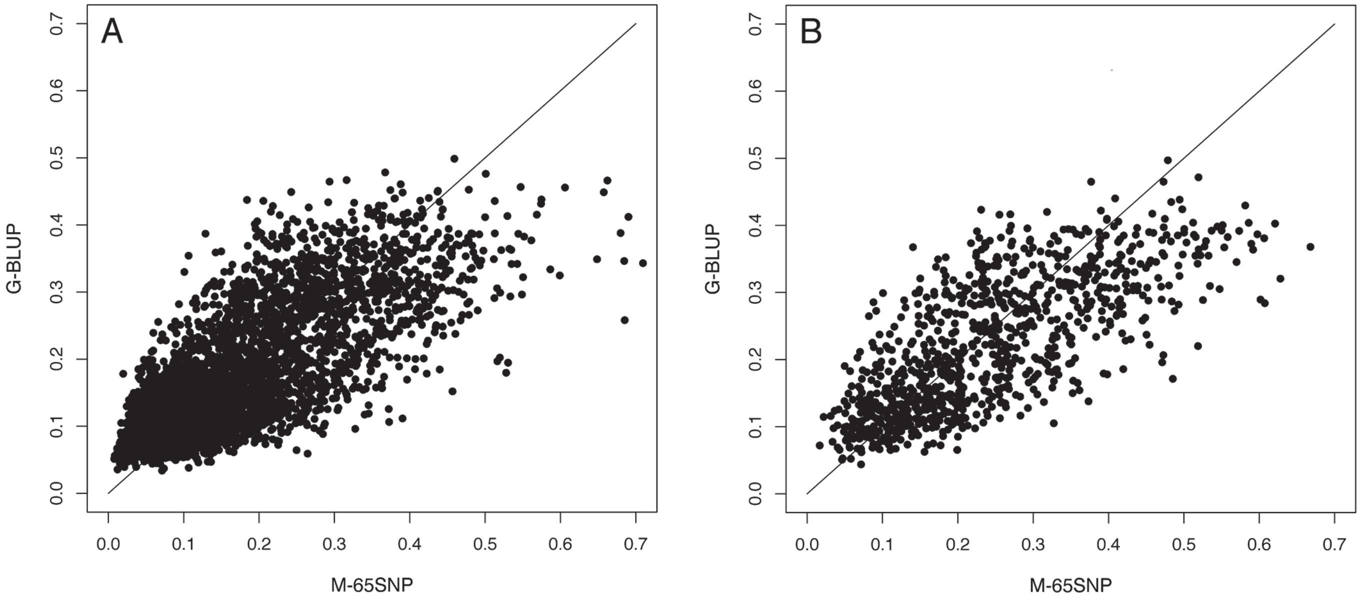

Fig 3. Probability of diabetes for M-65SNP and G-BLUP. These are classified by the presence or absence of diabetes: a) healthy and b) diabetic people.

doi:10.1371/journal.pone.0123818.g003

Bayes Cπ estimates. The estimated proportion of SNPs with a genetic signal (1- π) was quite large at 0.579± 0.114. Thus, this result suggests that more than 50% of the total number of SNPs contributed to the overall genetic variation, although the individual contribution of most of these SNPs is likely to be a miniscule proportion of the total genetic variation.

Genomic Heritability of Type 2 Diabetes. The G-BLUP model results in an estimate of genetic variance associated with common SNPs in the full training sample of 0.92 ± 0.26. Considering that the residual variance was fixed at 1, the estimate of genomic heritability for Type 2 Diabetes was 0.492±0.066. Previous studies demonstrated that in family datasets genomic heritability from G-BLUP yielded similar results [3,21].

Predictive Ability of the Models

Models were also evaluated for their predictive ability using a 10-fold cross-validation. We randomized families and assigned entire families to folds. The number of families per fold and the size of the families can be seen in S1 Table. The number of subjects per fold was between 364 and 727 individuals, contained within 147 to 181 families.

Table 1 shows the predictive AUC of the models (in the testing sets of the cross validation). The predictive ability of all the models that included genetic information improved with

Table 1. Area Under the Receiver Operating Characteristic Curve (AUC) for the CBM, M-65SNP, BRR, BA, BC, B-LASSO and G-BLUP.

Model	AUC-CV (Mean ± S.D.)
CBM	0.668 ± 0.025
M-65SNP	0.684 ± 0.041
BA	0.678 ± 0.027
BC	0.680 ± 0.027
B-LASSO	0.681 ± 0.027
G-BLUP	0.678 ± 0.027

doi:10.1371/journal.pone.0123818.t001

respect to the model only with covariates. However, the improvement was moderate and there were not important differences between models.

Discussion

Diabetes increases morbidity and mortality significantly, even when the condition is treated [51]. Diabetes is expected to rise from 171 million in 2001 (globally) to 366 million by 2030 [52]. Thus, it is imperative that we understand the underlying causes of diabetes, enhance our ability to identify those at risk, and mitigate that risk. We found that both genetic and non-genetic factors are associated with diabetes. In the Framingham study, gender is associated with developing T2D and has an odds ratio of 0.77 for females. Age also clearly matters, as evidenced in our study by the odds ratio of 0.68 for the Offspring cohort (that is younger) and the odds ratio of 1.02 per year of exposure for the effect of age at last contact. The Framingham cohorts are primarily Caucasian and are of European descent [53] (see Fig 1), and T2D susceptibility did not vary with ancestry in this population. This is contrary to findings for conditions such as skin cancer, which may be influenced by differences in ancestry even among those of European descent [54] and may indicate that risk alleles for T2D among those of European descent are relatively homogeneous. Although diabetes prevalence is not consistent across ethnicities, we found no evidence of origin differences within Caucasians. Genetic factors also appear to play a substantial role in T2D susceptibility. In our paper we estimated a genomic heritability which is within the range of what has been reported in the literature. Additionally, there are several reports of uncovered SNP variants associated with diabetes, and in our study we confirmed evidence of association between individual genetic score and T2D.

The predictive ability of the WGR models was moderately improved in comparison to the model based only on clinical covariates. It is known that prediction accuracy will greatly depend on the genetic distance between training and testing sets [55–57]. Several animal and plant studies [17,43,55,58]; [59,60] and some human studies where training and validation samples are closely related [18,20,61] have shown that WGRs can achieve high predictive power and sometimes even produce better predictions than those based on pedigrees [54]. For unrelated individuals, theoretical formulae for the prediction accuracy of G-BLUP [62,63] suggest that achieving reasonably good accuracy of estimates of effects for dense marker panels will require using extremely large samples. According to [62] the prediction accuracy of a WGRs depends on two main factors: (a) the proportion of variance that can be explained by regression on markers, and (b) the accuracy of effects estimates. As more markers are added to a model, genomic heritability increases; however, the more markers we include in the model the lower the accuracy of estimates of individual effects. Consequently, for any given sample size, adding large numbers of markers to the regression may increase the estimated genomic heritability but will not necessarily result in higher predictive power.

In our study, most models evaluated did not differ in the shrinkage of the SNP estimates or in the prediction accuracy. In agreement with our results in humans, Daetwyler and coauthors in a review paper report that there are not major differences in the predictive ability of WGR methods in several animal and plant studies [22]. Most traits of interest are highly complex and the benefit of heavy tailed distribution or mixtures is expressed more in traits where few genes explain a sizable proportion of the genetic variance [40]. Our results suggest a highly complex genetic architecture of T2D; in this situation there are no markers that could improve the modeling by capturing a greater signal since there is no greater signal to be captured. Models that are able to do selection selected approximately 50% of the markers. However, our result may also be affected by population substructure in the training sample since the training set has families. In these samples of related subjects, causal SNPs will be in long regions of high

LD, whereas in samples with unrelated subjects these regions will be shorter; thus, within that region any SNP would be an equally good predictor to capture the effect of the causal SNP. Consequently, this 'more complex' architecture mitigates differences between the models. Having training sets with families is equivalent to reducing the sample size of the training set, which we know is essential to achieve good prediction accuracy[43,62]. Still, among all the methods, the Bayesian LASSO had slightly thicker tails, allowing for more markers to have higher effects, while also achieving higher prediction accuracy.

Nevertheless, the AUC and performance achieved with 65 known markers was the same as the one achieved with Bayes Cπ and the G-BLUP. In the case of diabetes, BMI, height, and several other complex traits or diseases, mega consortiums are pulling thousands of genotyped subjects and their phenotypes to find the SNPs highly associated with the phenotype. However, for several phenotypes these resources are not available, and will probably never be (e.g., rare diseases), and significant SNP markers are unknown. Hence it is worth using the G-BLUP or Bayes Cπ, both of which achieved the same performance relative to the model using information of the well-established SNPs without using any prior information from large GWAS consortiums.

In summary, we found evidence of genetic effects in T2D, which reflects in the heritability estimates; additionally, existing genetic variation can be captured by high-density markers. Results from different WGR methods did not differ. We can find similar reports in the animal science literature; the improvement in AUC using cross validation was positive, albeit poor. Finally, the AUC from 65 well-known variants affecting diabetes was similar to that obtained from including all variants. However, these 65 SNPs were found in a large consortium study. Most complex traits and diseases do not have large consortiums data available, and in rare diseases, it is possible that there will never be large amounts of phenotypic and genomic data. WGRs could be important for diseases where no large consortium is available. Bayes LASSO and G-BLUP models are alternative methodologies using dense arrays of markers.

Supporting Information

S1 Fig. Trace plot and density plot of the genetic variance associated to the markers in the dense SNP array.
(TIFF)

S2 Fig. Trace plot and density plot of the genetic variance associated to the markers in the dense SNP array.
(TIFF)

S1 Table. Number of families and subjects assigned to each fold of the 10-fold cross-validation.
(DOCX)

Acknowledgments

The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. Funding for SHARe Affymetrix genotyping was provided by NHLBI Contract N02-HL-64278. SHARe Illumina genotyping was provided under an agreement between Illumina and Boston University. AIV acknowledge National Institute of Health grant, from National Institute of Diabetes and Digestive and Kidney, Research

Supplement to 7-R01-DK-062148-10, for their financial support. YCK acknowledges support from NIH grant K01DK095032. PPR had financial support from NIH Grants R01GM099992 and R01GM101219.

Author Contributions

Conceived and designed the experiments: AIV. Analyzed the data: AIV PPR. Wrote the paper: AIV YCK EJD YCV PPR.

References

1. Maher B. Personal genomes: The case of the missing heritability. Nature. 2008; 456: 18. doi: 10.1038/456018a PMID: 18987709
2. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461: 747–753. doi: 10.1038/nature08494 PMID: 19812666
3. Klimentidis YC, Vazquez AI, de Los Campos G, Allison DB, Dransfield MT, Thannickal VJ. Heritability of pulmonary function estimated from pedigree and whole-genome markers. Front Genet. 2013; 4: 174. doi: 10.3389/fgene.2013.00174 PMID: 24058366
4. Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, et al. Beyond Missing Heritability: Prediction of Complex Traits. PLoS Genet. 2011; 7: e1002051. doi: 10.1371/journal.pgen.1002051 PMID: 21552331
5. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010; 42: 565–569. doi: 10.1038/ng.608 PMID: 20562875
6. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460: 748–752. doi: 10.1038/nature08185 PMID: 19571811
7. De los Campos G, Gianola D, Allison DB. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet. 2010; 11: 880–886. doi: 10.1038/nrg2898 PMID: 21045869
8. Avendano S, Watson, K., Kranis, A. Genomics in Poultry Breeding—from utopias to deliverables. 9th World Congress on Genetics applied to Livestock Production (WCGALP). 2010; Available: http://www.kongressband.de/wcgalp2010/assets/html/0049.htm
9. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci. 2009; 92: 433–443. doi: 10.3168/jds.2008-1646 PMID: 19164653
10. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, et al. Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci. 2009; 92: 16–24. doi: 10.3168/jds.2008-1514 PMID: 19109259
11. Kuo L, Mallick B. Variable selection for regression models. Sankhy Indian J Stat Ser B. 1998; 65–81.
12. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics. 2011; 12: 186. doi: 10.1186/1471-2105-12-186 PMID: 21605355
13. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970; 55–67.
14. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol. 1996; 267–288.
15. Park T, Casella G. The Bayesian lasso. J Am Stat Assoc. 2008; 103: 681–686.
16. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157: 1819–1829. PMID: 11290733
17. Vazquez AI, Rosa GJM, Weigel KA, de los Campos G, Gianola D, Allison DB. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. J Dairy Sci. 2010; 93: 5942–5949. doi: 10.3168/jds.2010-3335 PMID: 21094768
18. Vazquez AI, de los Campos G, Klimentidis YC, Rosa GJM, Gianola D, Yi N, et al. A Comprehensive Genetic Approach for Improving Prediction of Skin Cancer Risk in Humans. Genetics. 2012; Available: http://www.genetics.org/content/early/2012/10/04/genetics.112.141705.short
19. Aslibekyan S, Wiener HW, Wu G, Zhi D, Shrestha S, de Los Campos G, et al. Estimating Proportions of Explained Variance: a Comparison of Whole Genome Subsets. BMC Proceedings. 2014; doi: 10.1186/1753-6561-8-S1-S102

20. De los Campos G, Klimentidis YC, Vazquez AI, Allison DB. Prediction of Expected Years of Life Using Whole-Genome Markers. PloS One. 2012; 7: e40964. doi: 10.1371/journal.pone.0040964 PMID: 22848416

21. De los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D. Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor. Goddard ME, editor. PLoS Genet. 2013; 9: e1003608. doi: 10.1371/journal.pgen.1003608 PMID: 23874214

22. Daetwyler HD, Calus MP, Pong-Wong R, de los Campos G, Hickey JM. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics. 2013; 193: 347–365. doi: 10.1534/genetics.112.147983 PMID: 23222650

23. Ishwaran H, Rao JS. Spike and slab variable selection: frequentist and Bayesian strategies. Ann Stat. 2005; 33: 730–773.

24. Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. Diabetes Res Clin Pract. 2010; 87: 4–14. doi: 10.1016/j.diabres.2009.10.007 PMID: 19896746

25. Zaitlen N, Kraft P, Patterson N, Pasaniuc B, Bhatia G, Pollack S, et al. Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. PLoS Genet. 2013; 9: e1003520. doi: 10.1371/journal.pgen.1003520 PMID: 23737753

26. Newman B, Selby JV, King M-C, Slemenda C, Fabsitz R, Friedman GD. Concordance for type 2 (non-insulin-dependent) diabetes mellitus in male twins. Diabetologia. 1987; 30: 763–768. PMID: 3428496

27. Kaprio J, Tuomilehto J, Koskenvuo M, Romanov K, Reunanen A, Eriksson J, et al. Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. Diabetologia. 1992; 35: 1060–1067. PMID: 1473616

28. Poulsen P, Kyvik KO, Vaag A, Beck-Nielsen H. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study. Diabetologia. 1999; 42: 139–145. PMID: 10064092

29. Mills GW, Avery PJ, McCarthy MI, Hattersley AT, Levy JC, Hitman GA, et al. Heritability estimates for beta cell function and features of the insulin resistance syndrome in UK families with an increased susceptibility to type 2 diabetes. Diabetologia. 2004; 47: 732–738. PMID: 15298351

30. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segre AV, Steinthorsdottir V, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet. 2012; 44: 981–990. doi: 10.1038/ng.2383 PMID: 22885922

31. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet. 2008; 40: 638–645. doi: 10.1038/ng.120 PMID: 18372903

32. Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet. 2010; 42: 579–589. doi: 10.1038/ng.609 PMID: 20581827

33. Dawber TR, Kannel WB, Lyell LP. An approach to longitudinal studies in a community: the Framingham Study. Ann N Y Acad Sci. 1963; 107: 539–556. PMID: 14025561

34. Dawber TR, Meadors GF, Moore FE Jr. Epidemiological approaches to heart disease: the Framingham Study. Am J Public Health. 1951; 41: 279. PMID: 14819398

35. Drineas P, Lewis J, Paschou P. Inferring geographic coordinates of origin for Europeans using small panels of ancestry informative markers. PLoS One. 2010; 5: e11892. doi: 10.1371/journal.pone.0011892 PMID: 20805874

36. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. Department of Statistics, University of Oxford, Oxford, UK; 2009; 5: e1000529. doi: 10.1371/journal.pgen.1000529 PMID: 19543373

37. Howie B, Marchini J, Stephens M, Chakravarti A. Genotype Imputation with Thousands of Genomes. G3 GenesGenomesGenetics. Genetics Society of America; 2011; 1: 457–470. doi: 10.1534/g3.111.001198 PMID: 22384356

38. Consortium GP, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. Nature Publishing Group; 2012; 491: 56–65. doi: 10.1038/nature11632 PMID: 23128226

39. Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R. Additive genetic variability and the Bayesian alphabet. Genetics. 2009; 183: 347–363. doi: 10.1534/genetics.109.103952 PMID: 19620397

40. De los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 2013; 193: 327–345. doi: 10.1534/genetics.112.143313 PMID: 22745228

41. De los Campos G, Perez P. BGLR: Bayesian Generalized Linear Regression [Internet]. 2013. Available: http://cran.at.r-project.org/web/packages/BGLR/index.html

42. De los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, et al. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics. 2009; 182: 375. doi: 10.1534/genetics.109.101501 PMID: 19293140

43. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008; 91: 4414–4423. doi: 10.3168/jds.2007-0980 PMID: 18946147

44. Pérez P, Campos G de los. Genome-Wide Regression & Prediction with the BGLR Statistical Package. Genetics. 2014; genetics.114.164442. doi: 10.1534/genetics.114.164442

45. R Development Core Team R. R: A language and environment for statistical computing. R Found Stat Comput Vienna Austria. 2010; Available: http://www.R-project.org/

46. Vazquez AI, Weigel KA, Gianola D, Bates DM, Perez-Cabal MA, Rosa GJM, et al. Poisson versus threshold models for genetic analysis of clinical mastitis in US Holsteins. J Dairy Sci. 2009; 92: 5239–5247. doi: 10.3168/jds.2009-2085 PMID: 19762842

47. Vazquez AI, Gianola D, Bates D, Weigel KA, Heringstad B. Assessment of Poisson, logit, and linear models for genetic analysis of clinical mastitis in Norwegian Red cows. J Dairy Sci. 2009; 92: 739–748. doi: 10.3168/jds.2008-1325 PMID: 19164686

48. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006; 27: 861–874.

49. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005; 21: 3940. PMID: 16096348

50. Cowie CC, Eberhardt MS. Sociodemographic characteristics of persons with diabetes. Diabetes Am. 1995; 2: 85–116.

51. Garcia MJ, McNamara PM, Gordon T, Kannell WB. Morbidity and mortality in diabetics in the Framingham population: sixteen year follow-up study. Diabetes. 1974; 23: 105–111. PMID: 4359625

52. Wild S, Roglic G, Green A, Sicree R, King H. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. Diabetes Care. 2004; 27: 1047–1053. PMID: 15111519

53. Govindaraju DR, Adrienne Cupples L, Kannel WB, O'Donnell CJ, Atwood LD, et al. Genetics of the Framingham Heart Study population. Adv Genet. 2008; 62: 33–65. doi: 10.1016/S0065-2660(08)00602-0 PMID: 19010253

54. Vazquez AI, de los Campos G, Klimentidis YC, Rosa GJ, Gianola D, Yi N, et al. A Comprehensive Genetic Approach for Improving Prediction of Skin Cancer Risk in Humans. Genetics. 2012; 192: 1493–1502. doi: 10.1534/genetics.112.141705 PMID: 23051645

55. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007; 177: 2389. PMID: 18073436

56. Habier D, Tetens J, Seefried FR, Lichtner P, Thaller G. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol. 2010; 42: 5. doi: 10.1186/1297-9686-42-5 PMID: 20170500

57. Perez-Cabal M, Vazquez A, Gianola D, Rosa G, Weigel K. Accuracy of genomic predictions in USA Holstein cattle from different training-testing designs. Proceedings of the 9 th World Congress on Genetics Applied to Livestock Production. 2010. p. 150.

58. Goddard ME, Hayes BJ. Genomic selection. J Anim Breed Genet. 2007; 124: 323–330. PMID: 18076469

59. Crossa J, de los Campos G, Perez P, Gianola D, Burgueño J, Araus JL, et al. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics. 2010; 186: 713. doi: 10.1534/genetics.110.118521 PMID: 20813882

60. Resende MD, Resende MF, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, et al. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytol. 2012; 194: 116–128. doi: 10.1111/j.1469-8137.2011.04038.x PMID: 22309312

61. Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, et al. Beyond Missing Heritability: Prediction of Complex Traits. PLoS Genet. 2011; 7: e1002051. doi: 10.1371/journal.pgen.1002051 PMID: 21552331

62. Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009; 136: 245–257. doi: 10.1007/s10709-008-9308-0 PMID: 18704696

63. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 2010; 185: 1021–1031. doi: 10.1534/genetics.110.116855 PMID: 20407128