Beyond Missing Heritability: Prediction of Complex Traits

Robert Makowsky*, Nicholas M. Pajewski¤, Yann C. Klimentidis, Ana I. Vazquez, Christine W. Duarte, David B. Allison, Gustavo de los Campos

Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama, United States of America

Abstract

Despite rapid advances in genomic technology, our ability to account for phenotypic variation using genetic information remains limited for many traits. This has unfortunately resulted in limited application of genetic data towards preventive and personalized medicine, one of the primary impetuses of genome-wide association studies. Recently, a large proportion of the "missing heritability" for human height was statistically explained by modeling thousands of single nucleotide polymorphisms concurrently. However, it is currently unclear how gains in explained genetic variance will translate to the prediction of yet-to-be observed phenotypes. Using data from the Framingham Heart Study, we explore the genomic prediction of human height in training and validation samples while varying the statistical approach used, the number of SNPs included in the model, the validation scheme, and the number of subjects used to train the model. In our training datasets, we are able to explain a large proportion of the variation in height (h^2 up to 0.83, R^2 up to 0.96). However, the proportion of variance accounted for in validation samples is much smaller (ranging from 0.15 to 0.36 depending on the degree of familial information used in the training dataset). While such R^2 values vastly exceed what has been previously reported using a reduced number of pre-selected markers (<0.10), given the heritability of the trait (~ 0.80), substantial room for improvement remains.

Citation: Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, et al. (2011) Beyond Missing Heritability: Prediction of Complex Traits. PLoS Genet 7(4): e1002051. doi:10.1371/journal.pgen.1002051

Editor: Greg Gibson, Georgia Institute of Technology, United States of America

Received October 21, 2010; **Accepted** March 2, 2011; **Published** April 28, 2011

Funding: RM, NMP, and YCK were funded through post-doctoral training grants from the National Heart, Lung, and Blood Institute (T32 HL072757 and T32-HL007457). GdlC was funded by grant number R01 DK076771 from the National Institute of Digestive and Kidney Disorders. DBA was supported by grants R01DK076771 and R01GM077490. 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.

* E-mail: makowsky@uab.edu

¤ Current address: Department of Biostatistical Sciences, Wake Forest University Health Sciences, Winston-Salem, North Carolina, United States of America

Introduction

Few examples exist of findings from Genome Wide Association Studies (GWAS) being applied to preventive and personalized medicine. Despite the success of GWAS in the discovery of many novel disease variants, the variants identified as being statistically significant typically account for minimal fractions of the genetic variance, even for highly heritable traits [1]. This so-called "missing heritability" has prompted a wide array of explanations, ranging from poor modeling (e.g., unaccounted epistatic effects) [2,3], insufficient sample sizes [4], sparse genetic coverage [5], rare variants [6], undetected CNV effects [7], and over-estimated heritability [1,8,9]. While all of these problems (and possibly others [10]) likely contribute to some extent [11], recent articles by Yang *et al.* [12] (hereafter, the Yang Study), and others [13,14] suggest that the primary culprit may be a mismatch between the actual genetic architecture and the statistical techniques applied.

Typically, predictive models from GWAS are constructed using a small number of Single Nucleotide Polymorphisms (SNPs) that have been pre-selected using extremely low p-values derived from single-marker regressions. This approach is most sensible under the assumption that only a few loci affect the trait of interest; however, it performs poorly for complex traits [14,15], which could be subtly affected by many loci [16]. Drawing on methods commonly used in animal breeding [17], the Yang Study built a model for human height (a model trait that has recently received much attention because of its high heritability and relatively reliable phenotyping) with hundreds of thousands of SNPs jointly considered (see Visscher et al. [18] for an expanded commentary on the methodology employed).

Using a Whole Genome Prediction (WGP) method, the authors from the Yang Study estimated that common SNP variation (through Linkage Disequilibrium (LD) with causal polymorphisms) explained 45% of the phenotypic variance, thus accounting for more than 50% of the expected heritability of height, which is reported to be approximately 80% [19,20]. These results suggest that the underlying genetic architecture of human height likely consists of numerous polymorphisms of small effect, resembling the infinitesimal model of quantitative genetics [21,22]. Recent studies suggest similar conclusions for other complex traits, including schizophrenia and bipolar disorder [23], blood lipid levels [24], and body mass index [25], suggesting a broader utility for the approach of WGP methods to account for genetic variance of important complex human traits.

The results of the Yang Study are particularly exciting due to their implications for eventual application to preventive and personalized medicine. However, a remaining question is the extent to which WGP methods improve the prediction of yet-to-be observed phenotypes, given the distinction between proportion of variance accounted for (as a measure of goodness of fit) and predictive accuracy (Figure 1). Heritability estimates can be regarded as measures of goodness of fit (see Materials and

Author Summary

While previous genome-wide association studies have implicated numerous loci associated with complex traits, such loci typically account for a very small proportion of phenotypic variation. However, a recent study using height as a model trait has illustrated that common single nucleotide polymorphisms can explain a large amount of genetic variance when evaluated through whole-genome statistical models. However, it is unclear to what extent higher proportions of explained variance will translate into improved predictive accuracy in future populations. Here we evaluate the predictive ability of whole-genome models for human height while varying the modeling approach, the size of the training population, the validation design, and the number of SNPs. Our results suggest that whole-genome prediction models can yield higher accuracy than what is commonly attained by models based on a few selected SNPs; yet, given the heritability of the trait in question, there exists room for improving prediction accuracy. While gains in predictive accuracy are likely to be small based on more expansive genotyping, our results indicate that more substantial benefits are likely to be gained through larger training populations, as well through the inclusion of related individuals.

Methods for a discussion), yet it is well known that increasing goodness of fit will not necessarily lead to increased predictive accuracy in future samples, due to issues such as over-fitting [26]. In this study, we examine the relationship between estimates of variance accounted for and predictive ability using WGP methods. Using three different statistical approaches and validation designs, we examine how these relationships change as a function of the density of SNPs included, the size of the training sample, and the degree of familial information included in the training sample.

Results

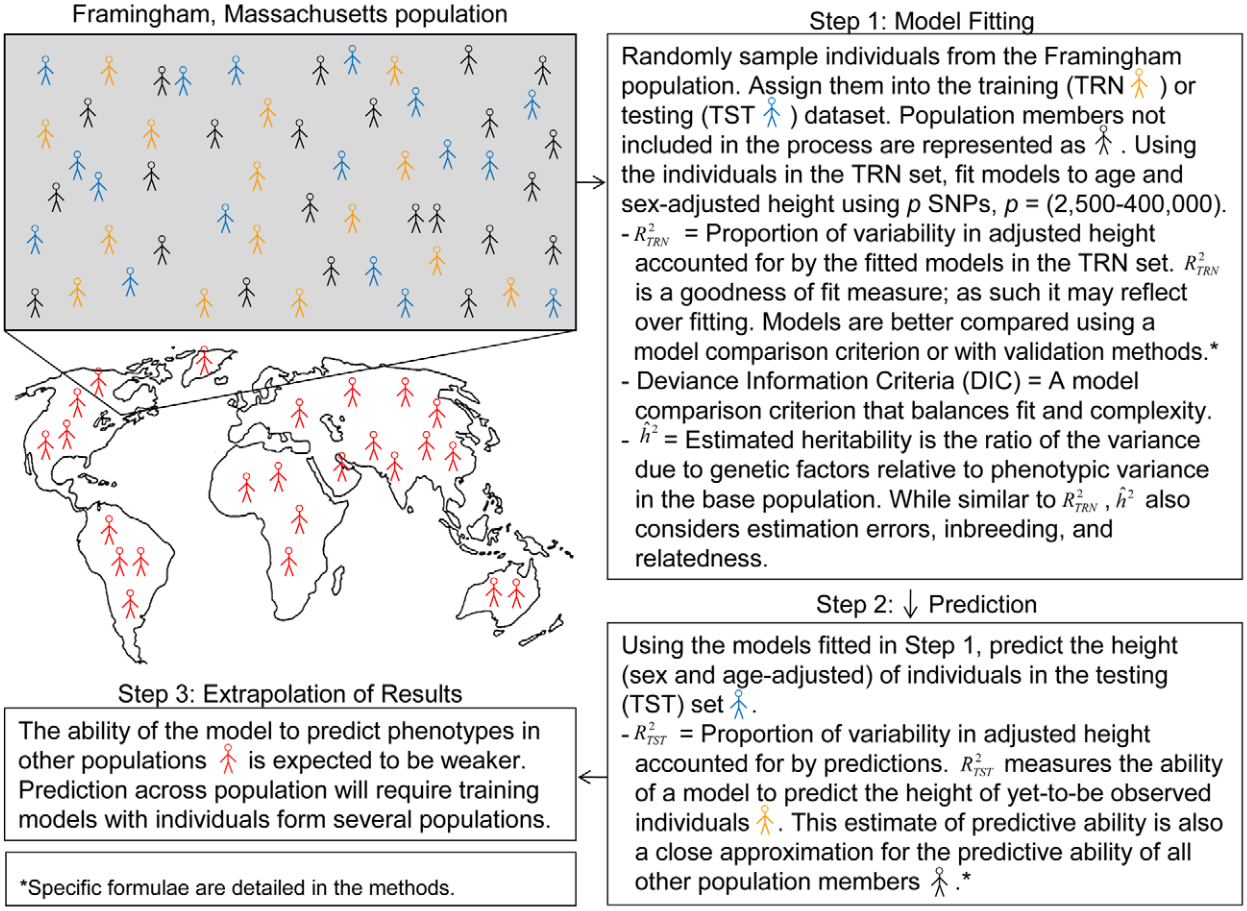
Using data from the Framingham Heart Study [27,28], we built models for the age and sex adjusted height of 5,117 adults using between 2,500 and 400,000 SNPs. Participants included in our analyses were individuals greater than 18 years old from the original (N = 1,493) or the offspring (3,624) cohorts; 2,311 individuals were male and 2,806 were female. Height ranged from 141.6 cm to 198.1 cm with a mean of 167.4 cm (SD = 9.5 cm). Markers were incorporated into statistical models in two ways: (i) regression of adjusted height on marker genotypes via the Bayesian LASSO (BL) [29] ; (ii) Bayesian random effects models using a marker-based (realized) relationship matrix between individuals (G). There are multiple ways to map marker genotypes into G and none is considered generally superior. Here we considered those used by Hayes and Goddard [30] and the Yang Study; the two models are denoted as G^H and G^Y, respectively, producing altogether three separate models. Goodness of fit was evaluated by means of the estimated residual variance and the proportion of variability accounted for by the fitted model in the training (TRN) dataset, R^2_{TRN}. In addition, models were compared based on the estimated heritability, $\hat{h}^2 = \frac{\sigma^2_g}{\sigma^2_p}$ (where σ^2_g is the variance attributed to additive genetic effects and σ^2_p is the total phenotypic variance) and the Deviance Information Criterion (DIC) [31].

Table 1 gives the estimated R^2_{TRN}, $\hat{h}^2$, and DIC by model and number of SNPs. Both R^2_{TRN} and $\hat{h}^2$ increase as more SNPs are included in the model, indicating an improved model fit. With 400,000 SNPs, the R^2_{TRN} statistic indicates that predicted genetic values (see Materials and Methods for a detailed description of terminology) accounted for 95% of the variability in adjusted height (R^2_{TRN}), and the estimated heritability ($\hat{h}^2 \sim 0.83$) is close to what has been previously reported for this trait. Based on the trend observed, any further increases in common SNPs would likely produce a minimal increase in the proportion of accounted variability.

As the number of markers increases, DIC decreases, indicating that information is continually being added to the model. This conforms with expectations under an infinitesimal model where the proportion of variance at Quantitative Trait Loci (QTLs) accounted for by regression on SNPs should increase with marker density [32]. Moreover, for any given number of SNPs, differences in the estimated residual variance, R^2_{TRN}, and heritability estimates across statistical approaches were small. We do not report $\hat{h}^2$ based upon the Bayesian LASSO: while formulae have been proposed to arrive at estimates of genetic variance from estimated marker effects and allele frequencies, they are problematic as they rely on the unrealistic assumption of linkage equilibrium between markers [33]. However, the similarity in R^2_{TRN} across models suggests that the proportion of variance accounted for by the Bayesian LASSO is similar to that of the two other methods.

To evaluate predictive ability, we used three different validation designs. *Approach A*- 10-fold cross-validation (CV) with assignment of individuals to folds at random. Because of the multiple generations present in the Framingham dataset, it is possible for children to be used to predict their parents in this design, which does not correspond to a standard prediction problem. To avoid this situation, we employed *Approach B*- using parents to predict children, we constructed a training dataset (TRN) with 1,493 parents and a testing dataset (TST) comprising offspring (N = 3,624). Because of the structure of the data, the size of the training sample used in Approach B is much smaller than that used in Approach A. Theory and empirical evidence [32] suggest that the accuracy of estimates of genetic values depends on the size of the training sample. To explore how much the size of the training sample affects predictive ability, we devised *Approach C*- randomly split the sample 10 times into TRN (N = 1,493) and TST sets (N = 3,624). Therefore, Approaches B and C differ in the way individuals were assigned to TRN and TST sets but not on the size of the TRN set. While approaches A and C allow for replicate datasets (10 in this study), Approach B is constrained to one replicate. As an aside, replicate datasets yielded highly similar R^2 values, with an average coefficient of variation of <0.5%.

Table 2 displays the estimated R^2 evaluated in validation (TST) samples (R^2_{TST}) by model, validation design, and number of SNPs. Within all validation designs, differences between models were very small. To better visualize the relationship between R^2, $\hat{h}^2$, and the number of SNPs, we average the results across modeling techniques (Figure 2). Predictive accuracy increased with the number of SNPs, reaching an R^2_{TST} of 25% in the 10-fold CV design when 400,000 SNPs were used. In the other two validation designs (approaches B and C), R^2_{TST} is considerably smaller than in the 10-fold CV, reaching a maximum R^2_{TST} of 13% (15%) in the 2-generation and random training-testing designs, respectively. The 10-fold CV uses larger relative training datasets than approaches B and C, which can affect prediction accuracy in at least two ways. First, using larger training datasets is expected to increase accuracy, even with nominally un-related individuals [32]. Concurrently, when the size of the training dataset is increased, the likelihood of having multiple close relatives included in the training data also increases, and, as we discuss below, for a


Framingham, Massachusetts population
Step 1: Model Fitting
Step 2: Prediction
Step 3: Extrapolation of Results

Randomly sample individuals from the Framingham population. Assign them into the training (TRN 🧍) or testing (TST 🧍) dataset. Population members not included in the process are represented as 🧍. Using the individuals in the TRN set, fit models to age and sex-adjusted height using *p* SNPs, *p* = (2,500-400,000).

- R^2_{TRN} = Proportion of variability in adjusted height accounted for by the fitted models in the TRN set. R^2_{TRN} is a goodness of fit measure; as such it may reflect over fitting. Models are better compared using a model comparison criterion or with validation methods.*
- Deviance Information Criteria (DIC) = A model comparison criterion that balances fit and complexity.
- $\hat{h}^2$ = Estimated heritability is the ratio of the variance due to genetic factors relative to phenotypic variance in the base population. While similar to R^2_{TRN}, $\hat{h}^2$ also considers estimation errors, inbreeding, and relatedness.

Using the models fitted in Step 1, predict the height (sex and age-adjusted) of individuals in the testing (TST) set 🧍.

- R^2_{TST} = Proportion of variability in adjusted height accounted for by predictions. R^2_{TST} measures the ability of a model to predict the height of yet-to-be observed individuals 🧍. This estimate of predictive ability is also a close approximation for the predictive ability of all other population members 🧍.*

The ability of the model to predict phenotypes in other populations 🧍 is expected to be weaker. Prediction across population will require training models with individuals form several populations.

*Specific formulae are detailed in the methods.

Figure 1. A simplified representation of assessment of goodness of fit in a training dataset and of predictive ability across a population: an example with the Framingham population.

doi:10.1371/journal.pgen.1002051.g001

fixed sample size, prediction accuracy increases with the number of close relatives used to train the model. Unfortunately, the CV designs we evaluate do not allow exact separation of the relative effect of sample size from that of other contributing factors.

The predictive accuracy of WGP methods is known to depend on how closely related individuals in the training and validation samples are to each other [34–36]. The Framingham Heart Study dataset contains varying degrees of familial relationships (e.g., parents, offspring, and siblings) and provides the opportunity to study how prediction accuracy is affected by including familial members in the training population. To demonstrate this effect, for every individual in the 10-fold CV testing datasets, we calculated the number of close relatives (parents, full sibs, half sibs and offspring) present in the training dataset used to derive its prediction. This was calculated as follows: let s_{ij} be an index which takes the values of 1 if individuals (i,j) are either full sibs or a parent-offspring pair, 0.5 if (i,j) is a half-sib pair, or 0 otherwise. Using this system, a score was calculated as $s_i = \sum_{j=1}^{N} s_{ij}\delta(i,j)$ where $\delta(i,j)$ equals one if individual i is in the testing population and individuals j is in the training population, and zero otherwise. Using this score we classified individuals into four groups ($s_i = 0$, $0 < s_i \leq 1$, $1 < s_i \leq 2$, $s_i > 2$) and calculated the average R^2_{TST} within each group after pooling the groups across CV folds.

Figure 3 depicts the relationship between the number of close relatives in the training population, the number of SNPs, and R^2_{TST} averaged across the three modeling techniques (see Table S1 for exact performance values). As expected, when the number of close relatives in the training dataset increases, the predictive ability increases. The relative increase in predictive ability with increasing SNP density is dependent upon the number of close relatives included in the model, with more drastic increases in predictive ability observed when more than two close relatives are included within the training dataset. When 400,000 SNPs are included, the average R^2_{TST} is 0.154, 0.267, 0.322, and 0.363 when $s_i = 0$, $0 < s_i \leq 1$, $1 < s_i \leq 2$, and $s_i > 2$, respectively.

Discussion

Our results are concordant with the Yang Study, demonstrating that much of the variance in human height can be accounted for using WGP methods based on common SNPs. However, there are a number of differences between our studies that warrant consideration. First, we focused on prediction accuracy and several factors that may affect it, while the Yang Study focused on estimating the proportion of variance in human height that can be explained by common SNPs. While we report heritability estimates, we stress that our estimates of $\hat{h}^2$ are not comparable to the $\hat{h}^2 = 0.45$ reported by the Yang study because, unlike the Yang Study, we did not restrict our sample to be composed of nominally unrelated individuals. While removing related individuals may allow estimation of genetic variance solely attributable to common SNPs through LD with causative polymorphisms, the use of exclusively un-related individuals may harm a model's ability to separate genetic signal from non-genetic components [36] and therefore measures of prediction

Table 1. R-squared statistic measured in the data used to train the model (R^2_{TRN}), estimated posterior mean of heritability ($\hat{h}^2$), and Deviance Information Criterion (DIC) by model and number of SNPs (where K = 1,000).

Number of SNPs	Bayesian Lasso[1]		Genomic Relationship G^Y			Genomic Relationship G^H		
	R^2_{TRN}	DIC	R^2_{TRN}	$\hat{h}^2$	DIC	R^2_{TRN}	$\hat{h}^2$	DIC
2.5K	0.33	32,920	0.36	0.21	32,883	0.34	0.26	32,912
5.0K	0.47	32,666	0.49	0.31	32,605	0.48	0.37	32,642
10K	0.65	32,106	0.69	0.47	31,950	0.66	0.52	32,081
20K	0.79	31,359	0.82	0.60	31,124	0.79	0.65	31,365
40K	0.87	30,564	0.89	0.70	30,201	0.87	0.74	30,564
80K	0.92	29,629	0.93	0.77	29,220	0.92	0.80	29,685
160K	-	-	0.95	0.79	28,925	0.93	0.81	29,416
400K	-	-	0.96	0.81	28,444	0.94	0.83	29,017

Estimates were obtained by fitting models to height adjusted by sex and age and using all available data (N = 5,117).

[1] For the Bayesian LASSO, due to high memory requirements, only models including up to 80K markers were considered. This model does not include a genetic variance parameter, therefore it does not yield a direct estimate of heritability. For this reason heritability is not reported for this model.

doi:10.1371/journal.pgen.1002051.t001

accuracy derived from such approach may under-estimate the predictive power of common SNPs. In addition, we focused on adult height (≥18 years old), while the Yang Study included individuals ≥16 years of age, which may induce added non-genetic variability as some teenagers will still be growing at that age. Finally, there likely are differences between the Framingham population and the Australian population used in the Yang study.

In all validation designs, we found that predictive ability increased with the number of SNPs, suggesting that a large number of SNPs are needed to capture genetic variance at QTLs. These results are similar to findings in the animal breeding literature for infinitesimal traits [37,38]. Our results also suggest a diminishing rate of return, with the difference in predictive ability between 80,000 and 400,000 SNPs being only ~6% in the 10-fold CV. However, the number of markers at which this "plateau" occurs is likely to depend on multiple conditions such as the extent of LD in the population and the number of individuals in the training data. Indeed, other studies using populations with smaller effective population sizes (N_e), and therefore larger LD spans, have reported high accuracy with much sparser coverage [37,38].

A recent study [39] reported a decrease in predictive ability of human height for models with p-value inclusion thresholds greater than 5×10^{-3}; suggesting that prediction accuracy may be harmed by including a large number of markers in a predictive model. However, an important difference between this study and ours is that in the former, marker effects were estimated using a fixed effects model while we use a Bayesian mixed model framework where all unknowns are modeled as random effects. Unlike the fixed effects approach, the Bayesian mixed model framework induces a shrinkage of estimates which, to some extent, controls over-fitting and seems to prevent a reduction in predictive ability in models with $p \gg n$.

Importantly, we found no drastic differences between any of the statistical methods we considered. This is not surprising given that all three methods are based on an underlying additive model and that height likely conforms to an infinitesimal architecture. Moreover, these results are in agreement with findings reported in the animal breeding literature [40] which report small differences in predictive ability between contrasting methods. However, this conclusion may not apply to traits with simpler architecture, e.g., traits where major associated variants explain a substantial proportion of genetic variance. In these cases, models using marker-specific shrinkage of estimates such as the BL may outperform models such as G^H or G^Y where all markers are equally weighted.

Theoretical [32,41] and empirical studies [37,38] demonstrate that prediction accuracy increases monotonically with the size of the training population. Our results showed the same pattern, with a ~70% increase in predictive ability when the size of the training dataset was increased from 1,493 to 4,506. A practical question resulting from this is how many individuals are needed to attain a certain predictive accuracy. The answer to such question depends on several factors such as trait heritability, marker density, N_e, the genetic architecture of the trait, and the degree of propinquity between individuals whose phenotypic outcomes are to be predicted and those used to train the model. For nominally unrelated individuals under an infinitesimal model for a trait with $h^2 = 0.8$, Goddard and Hayes [41] report that for effective population sizes of 100 or 1,000, achieving a correlation between predicted and true genetic values of 0.7, or equivalently, an R^2 between predicted and realized height of about 0.39 (calculated as $0.\ 7^2 \times 0.8$), requires training samples of approximately 4,000 and 50,000 individuals, respectively. However, as our results illustrate, prediction accuracy can be increased substantially by using information from related individuals.

Simulation [34] and empirical studies [35,36] in animal breeding have suggested that the prediction accuracy of WGP methods depends on familial relationships between individuals in the training and validation samples. This was confirmed by our analysis: in the 10-fold CV with 400K SNPs, the R^2_{TST} of individuals whose prediction was derived without using information from close relatives in the training dataset ($R^2_{TST} =$ 0.15) is much smaller than that obtained when direct relatives were included in the training dataset ($R^2_{TST} =$ 0.27, 0.32, and 0.36, for individuals with$0 < s_i \leq 1$, $1 < s_i \leq 2$, and $s_i > 2$ respectively). This occurs because WGP methods exploit genetic similarity across individuals and because recent family history plays a central role in determining genetic similarity. In light of this observation, one may wonder: does the use of genetic markers simply recapitulate pedigree-relationships? Several studies in animal and plant breeding have demonstrated the superiority of WGP over pedigree methods [40,42–44] suggesting that markers convey more information than that provided by pedigrees. In particular, molecular markers can account for similarity/differences due to common ancestry not traced by the pedigree, and, more

Table 2. R-squared between predicted and observed values (R^2_{TST}) estimated using different number of SNPs (where K = 1,000), models, and validation designs.

Number of SNPs	10-Fold CV[1]			2-Generations design[2]			Training-Testing Random[3]		
	BL	G^Y	G^H	BL	G^Y	G^H	BL	G^Y	G^H
2.5K	.097	.102	.098	.054	.035	.035	.064	.035	.033
5.0K	.126	.130	.129	.066	.058	.061	.080	.059	.057
10K	.166	.174	.168	.087	.088	.093	.099	.094	.088
20K	.200	.204	.199	.106	.111	.115	.119	.119	.114
40K	.217	.221	.216	.117	.118	.123	.128	.131	.126
80K	.236	.237	.236	.124	.126	.129	.138	.139	.137
160K	-	.240	.240	-	.130	.132	-	.142	.141
400K	-	.247	.249	-	.133	.133	-	.146	.145

BL = Bayesian LASSO, G^H = Goddard-Hayes, and G^Y = Yang study (see Materials and Methods for elucidation).

[1] 10-fold cross validation, where the training set comprised 4,605–4,606 individuals.

[2] Models were trained using the original cohort (N = 1,493) and predictive ability was assessed in the Offspring cohort (N = 3,624).

[3] Data was assigned at random to a training set (N = 1,493) and predictive ability was evaluated in the remaining individuals (N = 3,624). This was repeated 10 times; each time individuals were randomly assigned into training/testing sets. Results are averaged across the ten replicates.

doi:10.1371/journal.pgen.1002051.t002

importantly, markers can account for differences due to Mendelian segregation. Relative to plant or animal breeding populations, the level of inbreeding in humans is smaller, with the quality of pedigree information typically being poorer, if it is even available. Therefore, the benefits of using markers relative to pedigree information for prediction could be even larger in humans.

Clearly, there exists a redundancy between the information conveyed by the pedigree and that provided by markers. However, this redundancy is not complete and there may be benefits to incorporating pedigree and marker information in the model. For example, Vazquez et al. (2010) used data from US Holsteins to quantify the prediction accuracy using pedigree-based predictions, marker based WGP, and predictions combining pedigree and markers. The study confirmed the superiority of marker-based models (with a correlation of 0.42 for pedigree-based predictions and 0.649 for the marker-based predictions in CV) and found that, when more than 10,000 markers were available (for a Holstein sample), combining pedigree and molecular marker data was no better than using marker data only. This suggests that dense markers are able to capture genetic similarity due to recent family history as well as other sources of genetic similarity not described by pedigrees. Therefore, we speculate that the largely incomplete pedigrees of most humans will provide little to no additional information for the prediction of complex traits, especially given the high density of markers typically available.

A pertinent question is whether a WGP model fitted to one population can be used to predict phenotypes in a distantly related population; this remains, so far, an un-answered question [14]. The prediction accuracy of WGP methods depends on the patterns of LD between markers and QTLs; these are likely to change across populations and therefore it is reasonable to expect relatively poor prediction accuracy across populations. This does not represent a failure of the methodology per se, but instead a feature that needs to be considered when applying these methods for prediction.

Population structure, admixture, or other population features can lead to spurious associations and affect prediction accuracy; therefore accounting for these features has been an important focus for GWAS analyses [45]. A pertinent question is the extent to which structure and other forms of genetic diversity are accounted for by WGP methods. An important difference between WGP methods and standard single-marker regressions is that, when all markers are jointly modeled, population structure, admixture, familial relationships, genetic differences between full-sibs within a family, and genetic relationships between nominally un-related individuals are all implicitly accounted for to the extent that whole-genome markers describe them. Indeed, regressing a phenotype simultaneously on a set of whole-genome markers is equivalent to regressing the phenotype on all marker-derived principal components, with a degree of shrinkage in the estimated effect for each component that is proportional to its associated squared-singular value [46]. The Framingham population consists of individuals from various European ethnic backgrounds and height is typically correlated with northern European ancestry; therefore, population stratification is likely contributing to prediction accuracy [47]. Conversely, the patterns of LD between markers and QTL may be different across sub-populations and this may hinder predictive ability, especially when the sub-populations were separated for many generations [48]. The exact nature of this tradeoff is difficult to establish and constitutes an important area of future exploration.

In conclusion, WGP methods provide a promising approach for the prediction of complex traits. The results of the Yang Study and those reported in this study both support this conclusion: they account for a larger proportion of the expected genetic variance and, as our study indicates, are able to predict yet-to-be observed phenotypes with greater success. Yet, it is apparent that predictive ability depends to a large part upon how many close relatives are included while training the model, and there is an apparent need for improving the accuracy of predictions of nominally unrelated individuals. Therefore, while whole-genome prediction of complex human traits can yield more accurate predictions than those based on models using a reduced number of markers, prediction of such traits remains difficult and significant room for improvement exists.

Materials and Methods

Genotyping and Quality Control

Subjects were genotyped using the Affymetrix GeneChip Human Mapping 500K Array Set. For details on genotyping, see http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v3.p2. SNPs with call rates less than 90% and with a minor allele frequency (MAF) less than 3% were excluded. The remaining missing genotypes were imputed by sampling from a

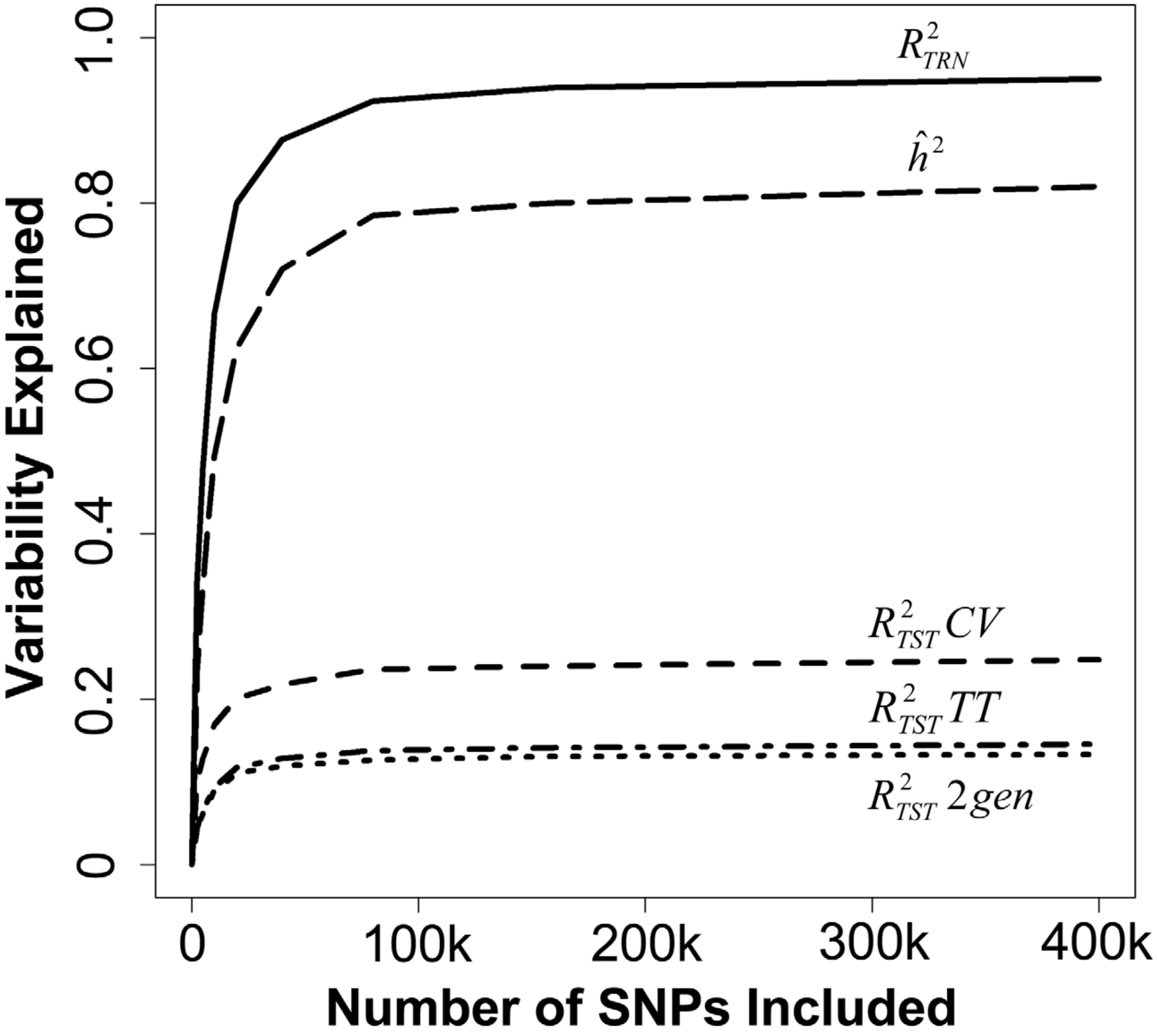


Figure 2. We averaged the estimates of R^2_{TRN} (measured in the training data), $\hat{h}^2$, $R^2_{TST}CV$ (measured in a 10 fold cross validation), $R^2_{TST}2gen$ (measured in a 2 generation validation), and $R^2_{TST}TT$ (measured in a replicated Training-Testing validation) over the three modeling techniques (BL, G^H, G^Y) and showed their relationship to the number of SNPs included in the model.
doi:10.1371/journal.pgen.1002051.g002

Binomial distribution using the empirical MAF estimate under the assumption of Hardy-Weinberg Equilibrium.

Genome-Wide Models for Human Height

In all models, age and sex-adjusted height of individual i, y_i, was expressed as $y_i = \mu + g_i + \varepsilon_i$ where: μ is an effect common to all individuals, g_i is a genetic value (i.e., a component of phenotypes that can be attributed to genetic factors), and ε_i is a model residual which captures all factors affecting the response not captured by g_i. The conditional distribution of the data is:

$$p(\mathbf{y}|\mu,\mathbf{g},\sigma^2_\varepsilon) = \prod_{i=1}^{n} N(y_i|\mu + g_i,\sigma^2_\varepsilon) \quad (1)$$

where, $\mathbf{y} = \{y_i\}$, μ is an effect common to all individuals, $\mathbf{g} = \{g_i\}$ is a vector of genetic values, and $N(y_i|\mu + g_i,\sigma^2_\varepsilon)$ is a normal density for the random variable, y_i, centered at $\mu + g_i$, with variance σ^2_ε.

All models were implemented in a Bayesian framework with inferences based on the posterior distribution of the unknowns given the data. Models differed in the number of markers used and the way they were incorporated into g_i. In the first group of models, genetic values were assumed to be multivariate normal:

$$p(\mathbf{g}|\mathbf{G},\sigma^2_g) = N(\mathbf{g}|0,\mathbf{G},\sigma^2_g) \quad (2)$$

where $\mathbf{g} = \{g_i\}$, $\mathbf{G} = \{G_{ij}\}$ is a relationship matrix between individuals i,j computed from marker genotypes and σ^2_g is an additive variance parameter. This approach has been used in many applications for modeling infinitesimal additive effects using molecular markers [12,30,49–51]. We focus on those used by Hayes and Goddard [30] (G^H) and the Yang Study (G^Y) to generate $\mathbf{G}$ from the marker data. In method G^Y, relationships are standardized so that the average diagonal value equals one. In order to make the genetic variance parameters comparable, this

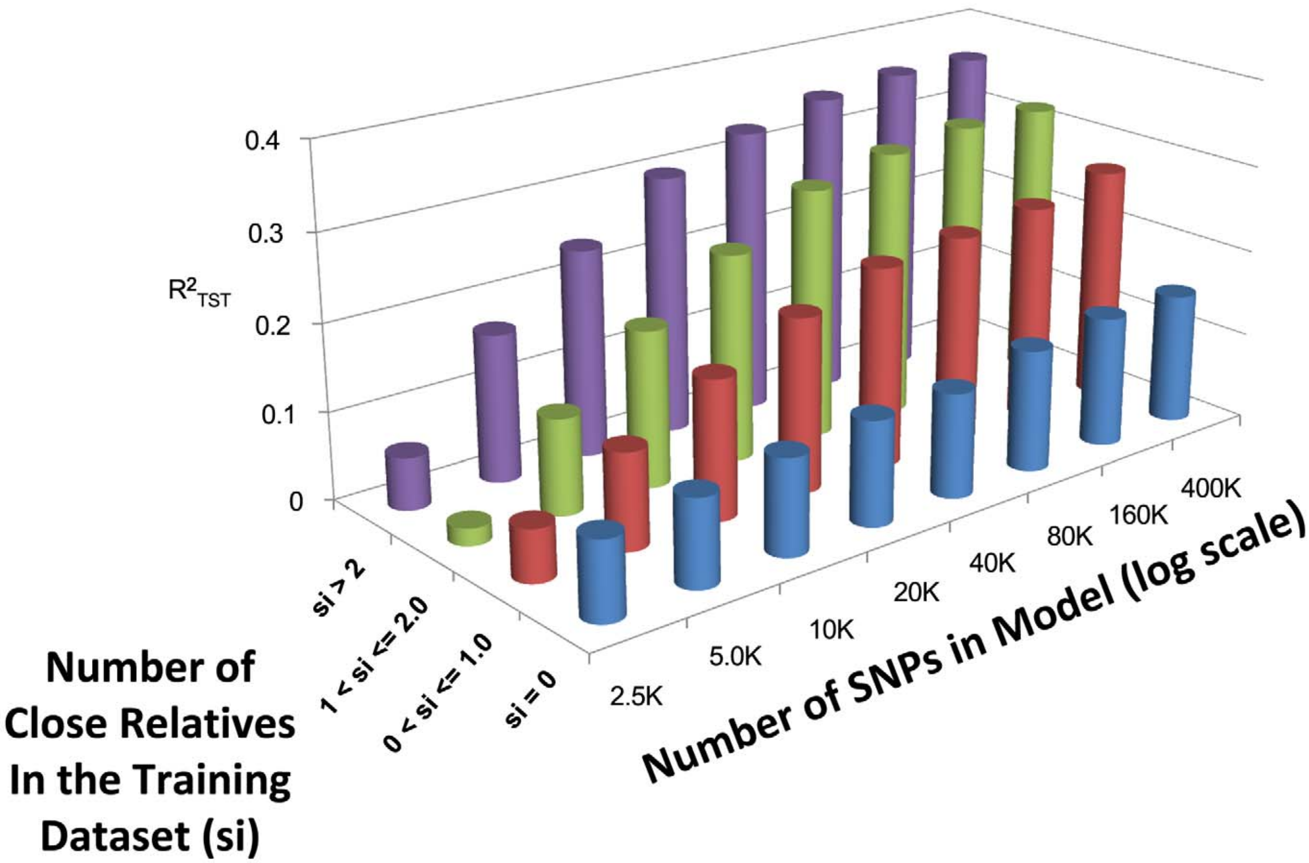


Figure 3. Averaged (across the three different models) estimates of $R^2_{TST}CV$ (measured in a 10 fold cross validation) while varying the number of close relatives (s_i) in the training dataset with 2.5K to 400K SNPs.

doi:10.1371/journal.pgen.1002051.g003

standardization was also applied to G^H by dividing the entries of $\mathbf{G}$ by the average diagonal value.

To estimate the remaining model parameters, we utilized a Bayesian approach by assigning prior distributions to $\Theta=\left\{\mu,\sigma^2_\varepsilon,\sigma^2_g\right\}$. We assigned a flat uniform prior to μ, with conjugate scaled inverse chi-square priors used for σ^2_ε and σ^2_g, implying a joint posterior density proportional to:

$$p(\mu,\mathbf{g},\sigma^2_\varepsilon,\sigma^2_g|\mathbf{y})\propto\left\{\prod_{i=1}^{n}N(y_i|\mu+g_i,\sigma^2_\varepsilon)\right\} \\ N(\mathbf{g}|\mathbf{0},\mathbf{A}\sigma^2_g)\chi^{-2}(\sigma^2_\varepsilon|df_\varepsilon,S_\varepsilon)\chi^{-2}(\sigma^2_g|df_g,S_g) \tag{3}$$

Samples from the posterior distribution of the above model were obtained using a Gibbs sampler implemented in the R-language (http://www.R-project.org). We specified the hyper-parameters in [3] as $\left\{df_\varepsilon=5,S_\varepsilon=60,df_g=5,S_g=60\right\}$. These values give a prior expectation of the variance of genetic values and of model residuals that are equal to approximately one half of the sample variance of adjusted height. With 5 degrees of freedom, priors have finite mean and variance, and a relatively small influence on inference.

In a third model, genetic values were described as a linear regression on marker covariates: $g_i=\sum_{l=1}^{L}X_{il}\beta_l$. Here, β_l is the additive effect of the l^{th} marker. Marker effects were inferred using the Bayesian LASSO (BL) of Park and Casella [29]. This model has been used successfully to model complex traits in genetic applications [37,43,52]. This leads to the joint posterior distribution density:

$$p(\mu,\boldsymbol{\beta},\sigma^2_\varepsilon,\boldsymbol{\tau}^2,\gamma|\mathbf{y})\propto\left\{\prod_{i=1}^{n}N(y_i|\mu+\sum_{l=1}^{L}X_{il}\beta_l,\sigma^2_\varepsilon)\right\} \\ \left\{\prod_{l=1}^{L}N(\beta_l|0,\sigma^2_\varepsilon\tau^2_l)Exp(\tau^2_l|\lambda^2)\right\}\times\chi^{-2}(\sigma^2_\varepsilon|df_\varepsilon,S_\varepsilon)G(\lambda^2|\delta,\gamma) \tag{4}$$

where $N\left(\beta_l|0,\sigma^2_\varepsilon\tau^2_l\right)$ denotes a normal prior assigned to β_l centered at zero and with prior variance equal to $\sigma^2_\varepsilon\tau^2_j$, $Exp\left(\tau^2_l|\lambda^2\right)$ is an exponential prior assigned to the τ^2_l's , and $G\left(\lambda^2|\delta,\gamma\right)$ is a Gamma prior assigned to the regularization parameter λ^2. This model was fitted using the BLR package [53] in R. The use of SNP-specific conditional prior variances, $Var\left(\beta_j|\tau^2_j,\sigma^2_\varepsilon\right)=\tau^2_j\sigma^2_\varepsilon$, allows for SNP-specific shrinkage of the estimates of effects. This contrasts with models G^H and G^Y in which all markers are equally weighted. The joint posterior distribution given by [4] is indexed by several hyper-parameters. In our application, those hyper-parameters were: $\left\{df_\varepsilon=5,S_\varepsilon=60,\gamma=0.55,\delta=1\times10^{-6}\right\}$. These values give a prior expectation of the residual variance that is about one half of the sample variance of adjusted age and a relatively flat prior density over a wide range of the regularization parameter λ. We applied the above-mentioned models using subsets of evenly-spaced SNPs, ranging from 2,500 to 400,000. Due to limitations in

RAM-memory, the maximum number of SNPs considered for the BL (method 3) was 80,000.

Heritability and R-Squared

Heritability, $\hat{h}^2=\frac{\sigma_g^2}{\sigma_p^2}=\frac{\sigma_g^2}{\sigma_g^2+\sigma_\varepsilon^2}$, is defined as the ratio of the variance due to additive genetic factors, σ_g^2, relative to the phenotypic variance, $\sigma_g^2+\sigma_\varepsilon^2$, in the base population (in a pedigree-model, this is the population from where the founders were sampled, which is assumed to be comprised of un-related individuals). This is also the squared correlation between genetic values and phenotypes, and the proportion of variance accounted for by genetic factors, both in the base population [54]. Heritability *estimates* ($\hat{h}^2$) are commonly obtained by replacing population parameters with estimates derived using Restricted Maximum Likelihood or Bayesian procedures.

The R^2 statistic is the ratio between the variance accounted for by a model relative to the sample variance of the response. That is: $R^2=1-\frac{Var(\hat{\varepsilon}_i)}{Var(y_i)}$ where $Var(\hat{\varepsilon}_i)$ is the sample variance of predictive residuals derived from a model and $Var(y_i)$ is the sample variance of phenotypes. The R^2 statistic is related to $\hat{h}^2$. However, R^2 measures the proportion of variance accounted for by predicted genetic values in the sample, while $\hat{h}^2$ estimates the proportion of phenotypic variance accounted by true genetic values in the base population. Fundamentally, R^2 ignores inbreeding, relationships between individuals in the sample and estimation errors; therefore, it is not a consistent estimate of heritability [54,55].

The R^2 statistic is sometimes evaluated in the same dataset that was used to derive predictions, which tend to over-estimate predictive ability. A better assessment of the ability of a model to predict future data can be obtained using validation methods [26]. We therefore distinguish two R-squared measures: $R^2_{TRN}=1-\frac{Var(\hat{\varepsilon}_i)}{Var(y_i)}$ and $R^2_{TST}=1-\frac{Var(\tilde{\varepsilon}_i)}{Var(y_i)}$ where: $\hat{\varepsilon}_i$ denotes a prediction error derived when all available data, including the i^{th} observation, was used to fit the model, and $\tilde{\varepsilon}_i$ denotes a prediction error derived when the validation set containing the i^{th} observation was not used to fit the model, respectively. Therefore, R^2_{TRN} measures goodness of fit between the training data and the model while R^2_{TST} measures the ability of the model to predict future observations.

Supporting Information

Table S1 R-squared between predicted and observed values (R^2_{TST}) estimated using different number of SNPs with different numbers of relatives in the training populations averaged across validation designs.
(DOC)

Acknowledgments

We would like to thank Kyle Grimes and Marshall Abrams for helpful comments on the writing, Vinodh Srinivasasainagendra for help obtaining the data, and constructive comments received from four anonymous reviewers.

Author Contributions

Conceived and designed the experiments: RM NMP GdlC. Analyzed the data: GdlC. Contributed reagents/materials/analysis tools: YCK AIV DBA. Wrote the paper: RM NMP YCK AIV CWD DBA GdlC.

References

1. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.
2. Clarke AJ, Cooper DN (2010) GWAS: heritability missing in action? Eur J Hum Genet 18: 859–861.
3. Huebinger RM, Garner HR, Barber RC (2010) Pathway genetic load allows simultaneous evaluation of multiple genetic associations. Burns 36: 787–792.
4. Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, et al. (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet 42: 570–575.
5. Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, et al. (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet 2: e41. doi:10.1371/journal.pgen.0020041.
6. Bodmer W, Tomlinson I (2010) Rare genetic variants and the risk of cancer. Curr Opin Genet Dev 20: 262–267.
7. Forer L, Schönherr S, Weissensteiner H, Haider F, Kluckner T, et al. (2010) CONAN: copy number variation analysis software for genome-wide association studies. BMC Bioinformatics 11: 318.
8. Maher B (2008) The case of the missing heritibility. Nature 456: 18–21.
9. Dominiczak AF, McBride MW (2003) Genetics of common ploygenic stroke. Nat Genet 35: 116–117.
10. Goriely A, Wilkie AOM (2010) Missing heritability: paternal age effect mutations and selfish spermatogonia. Nat Rev Genet 11: 589–589.
11. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11: 446–450.
12. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565–569.
13. Gibson G (2010) Hints of hidden heritability in GWAS. Nat Genet 42: 558–560.
14. de los Campos G, Gianola D, Allison DB (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet 11: 880–886.
15. Paynter NP, Chasman DI, Pare G, Buring JE, Cook NR, et al. (2010) Association between a literature-based genetic risk score and cardiovascular events in women. JAMA 303: 631–637.
16. Hill WG (2010) Understanding and using quantitative genetic variation. Phil Trans R Soc B 365: 73–85.
17. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.
18. Visscher PM, Yang K, Goddard ME (2010) A commentary on 'Common SNPs explain a large proportion of the heritability for human height' by Yang et al. Twin Res Hum Genet 13: 517–524.
19. Silventoinen K, Sammalisto S, Perola M, Boomsma DI, Cornes BK, et al. (2003) Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Res 6: 399–408.
20. Macgregor S, Cornes B, Martin N, Visscher P (2006) Bias, precision and heritability of self-reported and clinically measured height in Australian twins. Hum Genet 120: 571–580.
21. Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. Phil Trans R Soc Edinb 52: 399–433.
22. Wright S (1921) Systems of mating. I–V. Genetics 6.
23. International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, et al. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460: 748–752.
24. Teslovich T, Musunuru K, Smith A, Edmondson A, Stylianou I, et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466: 707–713.
25. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, et al. (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet 42: 937–948.
26. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: Data mining, inference, and prediction. New York: Springer-Verlag.
27. Dawber TR, Meadors GF, Moore FE (1951) Epidemiological approaches to heart disease: the Framingham Study. Am J Public Health 41: 279–286.
28. Dawber TR, Kannel WB, Lyell LP (1963) An approach to longitudinal studies in a community: the Framingham Study. Ann N Y Acad Sci 107: 539–556.
29. Park T, Casella G (2008) The Bayesian Lasso. J Am Stat Assoc 103: 681–686.
30. Hayes BJ, Goddard ME (2008) Prediction of breeding values using marker-derived relationship matrices. J Anim Sci 86: 2089–2092.
31. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. J Roy Stat Soc Ser B (Stat Method) 64: 583–639.
32. Goddard M (2009) Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136: 245–257.
33. Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R (2009) Additive genetic variability and the Bayesian alphabet. Genetics 183: 347–363.

34. Habier D, Fernando RL, Dekkers JCM (2007) The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics 177: 2389–2397.
35. Habier D, Tetens J, Seefried F-R, Lichtner P, Thaller G (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol 21: 5.
36. Perez-Cabal MA, Vazquez AI, Gianola D, Rosa GJM, Weigel KA (2010) Accuracy of genomic predictions in USA Holstein cattle from different training-testing designs. Proceedings of the 9th World Congress on Genetics Applied to Livestock Production # 563 and book of abstracts, p 150 August 1–6, Leipzig, Germany.
37. Weigel KA, de los Campos G, González-Recio O, Naya H, Wu XL, et al. (2009) Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J Dairy Sci 92: 5248–5257.
38. Vazquez AI, Rosa GJM, Weigel KA, de los Campos G, Gianola D, et al. (2010) Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. J Dairy Sci 93: 5942–5949.
39. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832–838.
40. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci 92: 433–443.
41. Goddard ME, Hayes BJ (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet 10: 381–391.
42. Crossa J, de los Campos G, Perez P, Gianola D, Burgueno J, et al. (2010) Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. Genetics 186: 713–724.
43. de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, et al. (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182: 375–385.
44. Van Raden PM, Wiggans GR, Van Tassell CP, Sonstegard TS, Schenkel FS (2009) Benefits from cooperation in genomics. Interbull Bulletin 39: 67–72.
45. Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11: 459–463.
46. de los Campos G, Gianola D, Rosa GJM, Weigel KA, Crossa J (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genetics Res 92: 295–308.
47. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, et al. (2005) Demonstrating stratification in a European American population. Nat Genet 37: 868–872.
48. de Roos APW, Hayes BJ, Goddard ME (2009) Reliability of Genomic Predictions Across Multiple Populations. Genetics 183: 1545–1553.
49. Lynch M, Ritland K (1999) Estimation of pairwise relatedness with molecular markers. Genetics 152: 1753–1766.
50. Eding H, Meuwissen THE (2001) Marker-based estimates of between and within population kinships for the conservation of genetic diversity. J Anim Breed Genet 118: 141–159.
51. Van Raden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, et al. (2009) Reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92: 16–24.
52. Yi N, Xu S (2008) Bayesian LASSO for quantitative trait loci mapping. Genetics 179: 1045–1055.
53. de los Campos G, Perez P (2010) BLR: Bayesian linear regression. R package version 1.1. http://www.R-project.org/.
54. Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31: 423–447.
55. Hadfield JD, Wilson Alastair J, Garant D, Sheldon Ben C, Kruuk Loeske EB (2010) The Misuse of BLUP in Ecology and Evolution. Am Nat 175: 116–125.