Estimation of genomic prediction accuracy from reference populations with varying degrees of relationship

S. Hong Lee*, Sam Clark, and Julius H.J. van der Werf*

School of Environmental and Rural Science, University of New England, NSW 2351, Australia

Running title: Estimation of genomic prediction accuracy

Key words: Genomic prediction; Prediction accuracy; Effective number of chromosome segments; Genomic prediction design; varying degrees of relationship

*Correspondence:

S. Hong Lee, School of Environmental and Rural Science, University of New England, NSW 2351, Australia

Tel: +61 2 6773 3665

Email: hong.lee@une.edu.au

Or

Julius H.J. van der Werf, School of Environmental and Rural Science, University of New England, NSW 2351, Australia

Tel: +61 2 6773 2092

Email: Julius.vanderwerf@une.edu.au

ABSTRACT

We present a theoretical framework for genomic prediction accuracy when the reference data consists of information sources with varying degrees of relationship to the target individuals. A reference set can contain both close and distant relatives as well as 'unrelated' individuals from the wider population, assuming they all come from the same homogeneous population. The various sources of information were modeled as different populations with different effective population sizes (N_e). With a similar amount of data available for each source, we show that close relatives can have a substantially larger effect on genomic prediction accuracy than lesser related individuals. However, the number of individuals from the wider population can be far greater than that of close relatives. We validate our theory with analysis of real data, and illustrate that the variation in genomic relationships with the target, rather than the variation in genomic relationship as a deviation for the expected relationship, is a predictor of the information content of the reference set and information from pedigree relationships is then naturally included in the prediction framework. Both the effective number of chromosome segments (M_e) and N_e are considered to be a function of the data used for prediction rather than being population parameters. We illustrate that when prediction also relies on closer relatives, there is less improvement in prediction accuracy with an increase in training data or marker panel density. We release software that can estimate the expected prediction accuracy and power when combining different reference sources with various degrees of relationship to the target, which is useful when planning genomic prediction (i.e. before collecting data) in animal, plant and human genetics.

INTRODUCTION

Genomic prediction of (additive) genetic effects and phenotypes is emerging in a wide range of fields including animal and plant breeding, risk prediction in human medicine and forensics[1-4]. Genomic prediction requires modeling of the association between genome-wide single nucleotide polymorphisms (SNPs) and phenotypes. The success of genomic prediction is measured by its accuracy, i.e. how reliable a future phenotype of target individuals can be predicted.

Genomic prediction requires a reference population of individuals having information on both genotype and phenotype. The accuracy of genomic prediction depends on various parameters, including sample size of the reference and its genetic structure. An important parameter in relation to the latter is the effective size of the population. The effective population size is a predictor of the effective number of chromosome segments that are represented in the population[5-7]. Theoretical predictions have usually considered a homogeneous population of individuals that are essentially unrelated. However, in most practical applications, the reference population used for genomic predictions possibly consists of many sub-groups with individuals having a variety of relatedness to the target individual, e.g. direct relatives, more distant relatives, and individuals that are considered unrelated, but still part of the same population as the target individual. It is relevant to assess the contribution of these various sources to prediction accuracy before actually conducting an experiment.

A number of studies have shown that genomic predictions are more accurate if the genomic relationship between the proband and the reference population is higher, both in humans[8-11] and in other species[12-14]. Habier et al (2013)[15] distinguished

between three types of information in genomic prediction; linkage disequilibrium, additive-genetic relationships and co-segregation of QTL predicted from markers genotypes with a pedigree. They argued that it would be useful to understand how these sources contribute to the accuracy of genomic predictions, especially when designing reference populations for breeding programs. They show these contributions via simulated examples but did not provide methods that allow simple predictions for their contribution to accuracy. Pszczola et al. (2012)[16] showed that the relationship between the reference population and the proband should be maximized to achieve an optimal design using a simulation study. However, they also did not attempt to derive the expected prediction accuracy from an optimal design in advance. Hayes et al. (2009)[17] considered the influence of direct relatives on genomic prediction. They followed the same approach as the general theory, i.e. by considering the number of independently segregating chromosome segments within families. They showed the accuracy of genomic prediction from varying sizes of the first and second degree of relatives, but did not consider the information from combined sources[18]. It should also be noted that those studies that derived genomic prediction accuracy from theory using effective number of chromosome segments (M_e)[5,6,19-21], did not consider the correlation between relatedness at different chromosomes, therefore overestimating M_e and underestimating whole-genome prediction accuracy[7].

Wientjes et al (2016)[22] proposed a simple selection index approach to combine information from different populations. They considered a genetic correlation between genetic effects expressed in different populations. We propose to use the same approach to combine different sources of information from within a population, where the different cohorts have a different degree of relationship with the target

individual. To predict the accuracy, we derive the number of effective chromosome segments from a hypothetical N_e associated with each subset, and we show that that combining such subsets using selection index theory gives the same result as using a prediction from an M_e derived from the variation in genomic relationships between the reference data and the target. Prediction accuracy is derived from variation in genomic relationship rather than the variation in genomic relationship as a deviation for the expected relationship among members of the reference set, as was proposed by Goddard et al (2011) and also applied by Wientjes et al (2016). This approach leads to a theoretical concept useful for assessing the accuracy of genomic predictions in advance, and we illustrate this with examples based on real data.

MATERIALS AND METHODS

Predicting genomic selection accuracy

The accuracy of genomic breeding values (GBV) or (genomic profile score in the context of human risk prediction[23]) based on genome-wide SNP genotypes can be predicted from theory[5-7,24], assuming that prediction is based on a reference population with phenotypes and genotypes for the same genome-wide SNPs that are linked to quantitative trait loci (QTL). The accuracy depends on i) the proportion of genetic variance at QTL captured by markers and ii) the accuracy of estimating marker effects. The proportion of genetic variance at QTL captured by markers (b) depends on linkage disequilibrium (LD) between markers and QTL, which in turn depends on the number of markers (M) and the number of 'effective chromosome segments' (M_e)[5], that is

b = M / (M_e + M)[5].

Various forms of prediction of M_e have been presented[5,6,21] that were however inconsistent to each other, and without considering the correlation between chromosomes. Recently, we presented a prediction formula with the form[7]

$$M_e = \frac{N_{chr}}{[\ln(2N_eL+1)+2N_eL(\ln(2N_eL+1)-1)]/(4N_e^2L^2)+(1/3N_e)\cdot(N_{chr}-1)} \quad (1)$$

where N_e = effective population size; L = average chromosome length; N_{chr} = number of chromosomes. This formula accounts for mutation, and that without considering mutation should be referred to equation (10) in Lee et al. (2017)[7]. The accuracy of the genomic prediction of a phenotype can be written as[5]

$$r_{y,\hat{g}} = h\cdot r_{g,\hat{g}} = h\cdot\sqrt{\frac{bh^2}{bh^2+M_e/N}} = \sqrt{\frac{bh^4}{bh^2+M_e/N}} \quad (2)$$

where $r_{y,\hat{g}}$ is the correlation coefficient between the true phenotypes (y) and estimated GBV, h^2 is the heritability of the trait, and N is the number of phenotypic observations. Other measures for genomic prediction accuracy, particularly for human risk prediction, such as the area under the receiver operating characteristic curve or odds ratio of case-control status contrasting the higher or lower risk group are described elsewhere[7].

M_e and genomic relationship

After collecting genotypic information of the reference data and the target individual, it is possible to obtain an empirical M_e from a genomic relationship matrix (GRM). In this derivation, the elements in the GRM are $G_{Tj} = \sum_{m=1}^{M} x_{Tm}x_{jm} / M$ where x_{Tm} and x_{jm} are the standardised genotype coefficients (mean 0 and variance 1) for the target individual (T) and jth individual in the reference data at the mth locus. It is possible to construct a GRM for each locus, and the elements in the GRM at the mth locus are $G_{Tj(m)} = G_{T*(m)} = x_{Tm}x_{jm}$ where * denotes the set of all reference individuals. Then, the variance of the mean of $G_{T*(m)}$ across all M_i SNPs in a single chromosome is

$$\begin{aligned}
\operatorname{var}\left(\sum_{m=1}^{M_i} G_{T*(m)} / M_i\right) &= \frac{1}{M_i^2}\left[\sum_{m=1}^{M_i}\sum_{l=1}^{M_i} \operatorname{cov}(G_{T*(l)}, G_{T*(m)})\right] \\
&= \frac{1}{M_i^2}\left[\sum_{m=1}^{M_i}\sum_{l=1}^{M_i} \operatorname{cov}(x_{T(l)}x_{*(l)}, x_{T(m)}x_{*(m)})\right] \\
&= \frac{1}{M_i^2}\left[\sum_{m=1}^{M_i}\sum_{l=1}^{M_i} \operatorname{cov}(x_{*(l)}, x_{*(m)}) \cdot (x_{T(l)} \cdot x_{T(m)})\right] \\
&= \frac{1}{M_i^2}\left[\sum_{m=1}^{M_i}\sum_{l=1}^{M_i} r_{m,l}^2\right] \\
&= \frac{1}{M_{e(i)}}
\end{aligned} \tag{3}$$

where $\mathrm{cov}(x_{*(l)}, x_{*(m)}) = x_{T(l)} \cdot x_{T(m)} = r_{m,l}$, which is a correlation between the m^{th} and l^{th} SNP-genotype, because of var(x) = 1 and mean(x) = 0, i.e. the genotype coefficients are standardized in the population, M_i is the number of SNPs in the ith chromosome and $M_{e(i)}$ is the effective number of chromosome segments for the ith chromosome, which is the inverse of the expectation of the squared correlation between SNPs[6,7]. When considering multiple chromosomes, the covariance of the pairwise relationship between two chromosomes is not negligible[7]. Assuming equal length and number of SNPs for N_{chr} chromosomes, M_e for the whole genome can be written as

$$\frac{1}{M_e} = \left[\mathrm{var}\left(\sum_{m=1}^{M_i} G_{T*(m)} / M_i \right) + \left(\frac{N_{chr} - 1}{3N_e} \right) \right] \cdot \frac{1}{N_{chr}} = \mathrm{var}\left(\sum_{m=1}^{M} G_{T*(m)} / M \right) \qquad (4)$$

where N_{chr} is the number of chromosomes.

In Goddard et al. (2011)[5], their theoretical derivation had to assume a homogeneous population of individuals that are essentially unrelated. However, here we show that the assumption about unrelated individuals can be relaxed so that any random samples from the population can be used for Eq. (4), irrespective of they are related or not.

Effective population size in a reference data set

One of critical parameters to determine the accuracy of genomic prediction is the effective population size (N_e). It is not very common to represent a reference population by a single value of N_e when it consists of several cohorts of individuals with different relationships to the target individual. The only study using a single value for M_e representing a reference set consisting of two population is by Wientjes et al. (2016)[22]. Here, we introduce a novel concept based on the relationship between N_e, M_e and var(G_{T*}), which can assign a value of N_e for a reference population

consisting of several cohorts. For any subset of the reference data set, there are realized relationships with the target sample. From Eq. (4), a value of M_e, which is the inverse of the variance of the genomic relationships between the proband and the reference sample, can be assigned to the reference data. Then, a single value of N_e, which is a function of M_e from Eq. (1), can be obtained for the reference data. The effective population size of the reference set is therefore a parameter specific to the data used. It can be smaller, but also larger than the actual effective size that is often contributed to the population from which the data is derived, depending on whether closer or more distant individuals are chosen for the reference set.

Based on this concept of N_e, reflecting information content of the reference sample in relation to the target sample, we consider three information sources consisting of 1) close relatives of the proband, e.g. N_e = 10, 2) distant relatives or individuals from the local area of the proband, e.g. N_e = 100 and 3) a wider population sample of individuals that are not related to the proband, e.g. N_e = 1,000.

The GBV can be estimated based on each of these information sources, and the accuracy of the estimation can be calculated as above, e.g. $r_{g,\hat{g}(i)}$ from Eq. (1) where i represents the ith information source. It is also possible to estimate GBV based on combined data of all three information sources. Assuming a random sample from the same population for each source, the accuracy of the GBV based on the combined data set can then be calculated using standard selection index theory as

$$r_{g,\hat{g}} = \sqrt{\mathbf{g}'\mathbf{P}^{-1}\mathbf{g}} = \sqrt{\begin{bmatrix} r^2_{g,\hat{g}(1)} \\ r^2_{g,\hat{g}(2)} \\ r^2_{g,\hat{g}(3)} \end{bmatrix}' \begin{bmatrix} r^2_{g,\hat{g}(1)} & r^2_{g,\hat{g}(1)} \cdot r^2_{g,\hat{g}(2)} & r^2_{g,\hat{g}(1)} \cdot r^2_{g,\hat{g}(3)} \\ r^2_{g,\hat{g}(2)} \cdot r^2_{g,\hat{g}(1)} & r^2_{g,\hat{g}(2)} & r^2_{g,\hat{g}(2)} \cdot r^2_{g,\hat{g}(3)} \\ r^2_{g,\hat{g}(3)} \cdot r^2_{g,\hat{g}(1)} & r^2_{g,\hat{g}(3)} \cdot r^2_{g,\hat{g}(2)} & r^2_{g,\hat{g}(3)} \end{bmatrix}^{-1} \begin{bmatrix} r^2_{g,\hat{g}(1)} \\ r^2_{g,\hat{g}(2)} \\ r^2_{g,\hat{g}(3)} \end{bmatrix}} \quad (5)$$

where **g** is the vector with covariances between each of the GBV and the true breeding value, and **P** is the variance-covariance matrix of the set of GBV. The accuracy of the GBV based on the combined data set can also be estimated based on the weighted M_e from the three information sources. Assuming a random sample from the same population for each source, the weighted M_e can be obtained as

$$M_{e(weighted)} = \frac{1}{\sum_{k=1}^{N_{sub-sample}} \mathrm{var}(G_{T*})_k \, p_k} = \frac{1}{\sum_{k=1}^{N_{sub-sample}} \frac{p_k}{M_{e(k)}}} \quad (6)$$

where p_k is the proportion of the sample size over the total sample for each information source. The accuracy of the GBV based on the weighted M_e is identical with that using standard selection index theory above (Eq. (5)).

Following Wientjes et al. (2016)[22] we can further generalize for a case where genetic correlations among multiple reference populations and those between reference populations and the target are not one. Equation (5) can be generalized as

$$r_{g,\hat{g}} = \sqrt{\mathbf{g}'\mathbf{P}^{-1}\mathbf{g}} = \sqrt{\begin{bmatrix} r^2_{G_{1,T}} r^2_{g,\hat{g}(1)} \\ \vdots \\ r^2_{G_{k,T}} r^2_{g,\hat{g}(k)} \end{bmatrix}' \begin{bmatrix} r^2_{G_{1,T}} r^2_{g,\hat{g}(1)} & \cdots & r^2_{g,\hat{g}(1)} r^2_{g,\hat{g}(k)} r_{G_{1,k}} r_{G_{1,T}} r_{G_{k,T}} \\ \vdots & \ddots & \vdots \\ r^2_{g,\hat{g}(k)} r^2_{g,\hat{g}(1)} r_{G_{k,1}} r_{G_{k,T}} r_{G_{1,T}} & \cdots & r^2_{G_{k,T}} r^2_{g,\hat{g}(k)} \end{bmatrix}^{-1} \begin{bmatrix} r^2_{G_{1,T}} r^2_{g,\hat{g}(1)} \\ \vdots \\ r^2_{G_{k,T}} r^2_{g,\hat{g}(k)} \end{bmatrix}} \quad (7)$$

where $r_{G_{k,T}}$ is the genetic correlation between the k^{th} reference population and the target set, and similarly, $r_{G_{i,j}}$ is the genetic correlation between the i^{th} and j^{th} reference population ($i = j = 1 \sim$ k).

Simulation

In a simulation, a stochastic gene-dropping method [25,26] was used to simulate 4,000 SNPs for each of 30 chromosomes, each of length L=1 Morgan with N_e = 50, 500 and

1000 for 50, 500 and 1000 generations, respectively. Recombination and mutations were modelled according to the genetic distance between SNPs and the mutation rate of 1e-08 per site per generation[27]. In the final generation, we constructed a genomic relationship matrix for a random set of 3000 individuals. Among the 3000 individuals, we randomly selected 1000 individuals as target data and 2000 individual as reference data, and estimated variance of the genomic relationships between the target and reference data to validate Eq. (3).

Evaluation of the formulas

For each of the three information sources contributing to genomic prediction we varied values for N_e, sample size in reference data and marker density. We compared the expected accuracy of GBV from the sample of N_e =1000 with predictions that additionally included information from the sample of N_e =100 and N_e =10. The total number in the reference population was kept equal between the comparisons.

Real data analysis

We used publicly available data from the Framingham heart study (phs000007.v26.p10.c1)[28]. There were 6950 individuals genotyped for 500,568 SNPs. Stringent quality control for genotype data and phenotype adjustment for confounders were applied to the data (the details can be found in Lee et al. (2016)[7]). The quality control included SNP call rate > 0.95, individual call rate > 0.95, HWE p-value > 0.0001, MAF > 0.01 and individual population outliers < 6 SD from the first and second principal components (PC). After QC, 6920 individuals and 389,265 SNPs remained. Among them, 4243 individuals were phenotyped for height and body mass index (BMI).

We made three different information sources to form the reference data that were tested in 100 replicated analyses (Table 1). Initially, we randomly selected 800 individuals out of 4243 phenotyped individuals as a target data set. For reference data set #1, we selected 50% of individuals that were highly related (> relatedness of 0.3) to the 800 target individuals ($N_1 = 617 \pm 19$). For reference data set #2, we selected 80% of moderately related individuals (> relatedness of 0.1) of the 800 target individuals ($N_2 = 1254 \pm 30$). For reference data set #3, we took the rest of the individuals that were not selected for reference data set #1 and #2 ($N_3 = 1572 \pm 33$). There was no overlap sample between target data set and reference data sets #1, #2 and #3.

Using the real genotype data, the genomic relationships between the reference and target sample were constructed. Empirical M_e was estimated from equation (3) for reference #1, 2 and 3, and that for combined data. We took a median rather than mean because the distribution of variance of the genomic relationship between target and reference sample was skewed. The correlation between the true phenotypes (that were not used in the analyses) and estimated GBV in the target data set was estimated for the combined data set, which was used as the genomic prediction accuracy ($r_{y,\hat{g}}$).

Phenotypes were adjusted for birth year, sex, and the first 10 PCs were used to control non-genetic confounding effects, e.g. population stratification.

RESULTS

In the simulation study, as shown in Figure 1A, 1B and 1C, the expected (from Eq. (1)) and empirically observed M_e from the simulated genotyped data (using Eq. (4)) are in good agreement, however, they are considerably lower than the expectation

from the previous formulas[5,6,21], which confirms the result from Lee et al. (2016)[7]. It is noted that Eq. (4) is still valid in the subset with a smaller N_e = 50 that has a significant proportion of high related individuals, indicating that the assumption about unrelated individuals (made in Godard et al. (2011)[5]) can be relaxed.

In the evaluation of the formulas, we first tested how the prediction accuracy was changed with varying marker density, using formula (1) and (2) and b = M / (M_e + M) (Figure 2). For N_e=10,000, the accuracy gradually increased with marker density, but the slope became flat when using the number of SNPs exceeded 100,000 (Figure 2A). For N_e=1,000, the accuracy did not increase with marker density as long as the number of SNPs was higher than 50,000 (Figure 2B). For N_e=100, there was no improvement of the accuracy if the number of SNPs was more than 10,000 (Figure 2C). This would be expected because the proportion of genetic variance at QTL captured by markers (b = M / (M_e + M) approached one when the number of SNPs (M) was more than 100,000, 50,000 and 10,000 for N_e = 10,000, 1000 and 100, respectively (Figure 3), as M_e was equal to 21,248, 2,313 and 254 for these three cases.

Next, we quantified the contribution of each information source when varying sample size in the reference data using formula (1), (2) and (4) (Figure 4). It was assumed that the number of SNPs was sufficient to capture most of causal variants (e.g. > 50,000). When adding 100 individuals of N_e=100 or N_e=10 to the reference sample with N_e=1000, the accuracy was slightly or substantially improved (Figure 4A). The improvement was larger when adding more individuals (500) (Figure 4B). Results showed that an information source of a smaller N_e was more important when the

samples sizes of each information source were the same. When the total number of reference data was increased, the importance of adding an information source of a smaller N_e was relatively decreased (Figure 4). When heritability was higher, overall accuracy was increased, and the relative contribution from an information source of a smaller N_e, i.e. the close relatives, was reduced (Figure 5).

Figure 6 confirms again that the smaller N_e, the better the prediction accuracy when using each information source separately. However, the sample sizes can be also varied across the information sources, as there are generally a lot fewer close relatives than individuals from the wider population. In Figure 6A, the accuracy at a sample size of 100 for N_e=10 was 0.73, which was lower than that of a sample size of 1,000 for N_e=100 (0.81) or that of a sample size of 20,000 for N_e=1,000 (0.83). With a higher heritability, the result is similar in that the 20,000 records in the information source of N_e=1000 gave a better accuracy than the 100 records of close relatives (N_e=10).

In real situations, the most common and desirable design may combine all of the information sources to maximize the prediction accuracy. We plotted the accuracy using a composite design consisting of N_e=1000 + N_e=100 (N=500) + N_e=10 (N=50), compared to that using N_e=1000 (Figure 7). The accuracy for a composite design was substantially increased especially when the total number of reference sample is low (Figure 7).

Figure 8 illustrates the real data analyses. The median of empirically estimated M_e from the inverse of the variance of the genomic relationship (Eq. 3) over 100

replicates was 2254 (SD=50), 3989 (SD=104) and 28848 (SD=920) for reference #1, #2 and #3, respectively (Table 1). Empirically estimated M_e based on the combined data was 4836 (SD=106) while expected M_e was 5309 (SD=88), approximately confirming Eq. (4). The (small) difference between empirical observation and expectation was probably due to skewed distribution of the variance of the genomic relationships.

Given M_e, N and h^2, the expected accuracy (from Eq. (2)) agreed well with the observed accuracy when using Framingham data (Figure 9). The reported heritabilities, h^2=0.8 [29-31] for height and h^2=0.46[32,33] for BMI, were used.

Finally, we quantified the importance of marker density using the real data. In agreement with Figure 2, the prediction accuracy is not much decreased even with 50,000 SNPs that were randomly selected from 389,265 SNPs (Figure 10).

DISCUSSION

This work shows a simple approach for modeling genomic prediction in a reference data set that contained several subpopulations that differ in relatedness to the target set, and by modeling these subpopulations as having different effective population size. The model allows assessing the prediction accuracy before actually conducting an experiment so that designing genomic prediction can be precise and effective in animal, plant and human genetics. For example, it can address a question how much

the prediction accuracy can be increased by adding 10,000 (conventionally) unrelated individuals into the current experiment consisting of 100 relatives in the reference data. The value for N_e in equation (1) can be approximated based on prior knowledge of a population, and the relatedness of the sample with the target, possibly supported by some genotype information that maybe available on cohorts, or samples thereof. Prediction in advance indeed relies on arbitrary modelling a number of cohorts, but it would be a useful exercise, as illustrated in the results when considering marker density and various sizes of the subsets of the training data. The theory is also useful for an animal breeder to predict the value of genotyped animals in an own herd versus those in a wider references population consisting of a larger number of more distantly related individuals.

The genotypic and phenotypic information of close and distant relatives of the proband can be effectively used as a part of the unified reference panel that also include a large number of individuals that are not related to the predicted subject to improve the accuracy further as illustrated in Figure 7. For a random sample from the same homogenous population, e.g. within the same breed or ethnicity, an optimal design should consist of both close and distant relatives and unrelated individuals, e.g. a composite design, to maximise the prediction accuracy (Figure 7). That is, the composite design takes advantage of effective information from smaller number of relatives while it also use information from a greater number of unrelated individual.

Using equation (1) and (4), we showed that the prediction accuracy derived for a population with unrelated individuals turns out to be higher, compared to previous quantifications that overestimated M_e for a larger number of chromosomes[5,6,21,34].

Using the same theory, we also showed that the information from close relatives could increase the accuracy even further, especially for smaller reference populations (Figures 3 – 6). It is important to note that the assumption about using unrelated individuals in estimating empirical M_e from genomic relationship[5] is not strictly necessary and can be relaxed (Eq. (3) and (4)). The theory and empirical observation from simulation study agreed well (Figure 1) even when using a population with a smaller effective population size (N_e=50) that consisted of a significant proportion of high relatedness.

Previous studies related to genomic prediction accuracy have suggested that M_e can be derived from the variation in the differences between realized and expected relationships[6,22], i.e. $\mathbf{D} = \mathbf{G} - \mathbf{A}$ where **G** is a genomic relationship matrix and **A** is a numerator relationship matrix based on pedigree. Those studies validated their results also with simulation. If the individuals used in the training set have a low expected relationship to the target individuals, then there is not much difference between the variations in **D** versus **G**. However, when some closer relatives are used, var(**G**) is larger than var(**D**) and M_e is therefore smaller. Note that non-random sampling of individuals used for the training set can cause a difference between the N_e of the population that was simulated, and the N_e of the data set that was used for prediction.

We have not tested the theory for multi-breed reference populations, i.e. those that are heterogeneous in the sense of consisting of populations from different genetic background, i.e. different breeds or ethnicities, each with different minor allele frequencies, different LD structure and different effects for causal variants. Wientjes et al. (2016)[22] explicitly addressed the problem of different effects for causal variants

(i.e. genetic correlation less than one) when combining data from two populations. Individuals from different populations share genomic relationships that are lower than those among members within each population. Evidence in literature suggests low prediction accuracies when using information from different breeds or populations, which could be viewed as predicting from populations with very large N_e. Moreover, we have not considered historical population dynamics such as bottleneck and admixture, but assumed a constant N_e over the historical generations, which leads to simplifications that make the formulae tractable and easy to derive. Further work is required to extend the theory accounting for admixture populations and historical population dynamics.

We have shown an improved theory for the prediction of the effective number of chromosome segments, which is a key parameter in genomic prediction accuracy[7]. The theory accounts for the correlation between relationships at different chromosomes and as a result the effective number of chromosome segments is smaller than predicted from previous theory[5,6,21]. As a result, the increase of the genomic prediction accuracy appears to be less reliant on higher marker density unless N_e is very large (e.g. > 10,000) (Figure 2), compared to what have been quantified by previous theory[5,6,21]. The previous theory overestimates M_e (mostly due to neglecting correlation between chromosomes), therefore underestimates the proportion of genetic variance at QTL captured by markers. Little improvement of prediction accuracy with increasing SNP marker density has been empirically observed in a number of studies[35-37]. This may also have important implication in genomic prediction as to designing marker density in animal, plant and human genetics.

The ability to quantify the accuracy in relation to various degrees of relationships (e.g. close relatives, distant relatives, local or extensive population sample) is important for predicting outcomes of genomic prediction for specific designs. This study has addressed this question, and the theory has been implemented in MTG2 software (https://sites.google.com/site/honglee0707/mtg2). Therefore, a user can know the expected prediction accuracy and the power[38] before designing an experiment of genomic prediction. Our approach can be applied both before and after collecting the data.

ACKNOWLEDGEMENTS

This research is supported by the Australian National Health and Medical Research Council (APP1080157), the Australian Research Council (DP160102126, FT160100229) and the Australian Sheep Industry Cooperative Research Centre. 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. The authors acknowledge useful discussion with Han Mulder that contributed to a clearer paper.

DISCLOSURE DECLARATION

The authors declare no competing financial interests.

Table 1. The sample size (N) and empirically observed M_e in each of three different reference data sets and combined data set in the Framingham data analysis.

	N_i (standard deviation)	M_e (standard deviation)
Reference #1	N_1 = 617 (19)	2254 (50)
Reference #2	N_2 = 1254 (30)	3989 (104)
Reference #3	N_3 = 1572 (33)	28848 (920)
Combined all	N_{all} = 3443 (0)	4836 (106)

The values were averaged over 100 replicates.

References

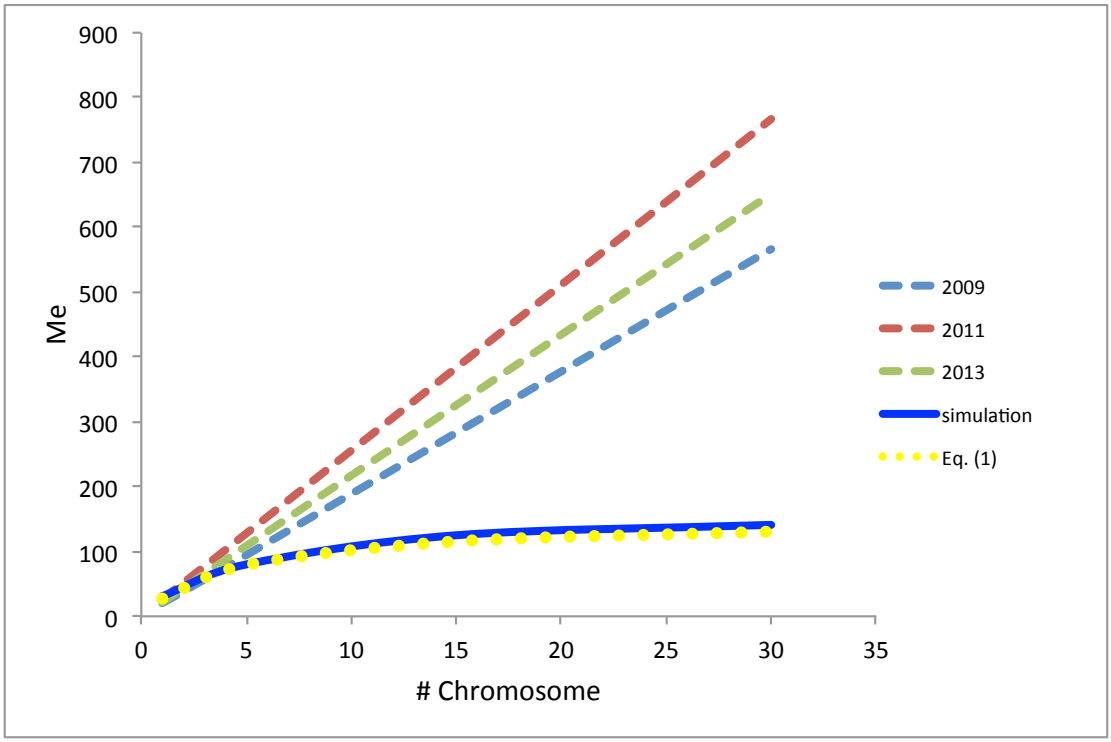
1. Meuwissen, T., Hayes, B. & Goddard, M. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819 - 1829 (2001).
2. Wray, N.R., Goddard, M.E. & Visscher, P.M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**, 1520-1528 (2007).
3. Collins , F.S. & Varmus , H. A New Initiative on Precision Medicine. *New England Journal of Medicine* **372**, 793-795 (2015).
4. Jannink, J.-L., Lorenz, A.J. & Iwata, H. Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics* **9**, 166-177 (2010).
5. Goddard, M.E., Hayes, B.J. & Meuwissen, T.H.E. Using the genomic relationship matrix to predict the accuracy of genomic selection. *Journal of Animal Breeding and Genetics* **128**, 409-421 (2011).
6. Goddard, M.E. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**, 245-257 (2009).
7. Lee, S.H., Weerasinghe, W.M.S.P., Wray, N., Goddard, M. & Van der Werf, J. Using information of relatives in genomic prediction to apply effective stratified medicine. *Scientific Reports* **7**, 42091 (2017).
8. Tucker, G. *et al.* Two-Variance-Component Model Improves Genetic Prediction in Family Datasets. *The American Journal of Human Genetics* **97**, 677-690.
9. de los Campos, G., Vazquez, A.I., Fernando, R., Klimentidis, Y.C. & Sorensen, D. Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor. *PLoS Genet* **9**, e1003608 (2013).
10. Makowsky, R. *et al.* Beyond Missing Heritability: Prediction of Complex Traits. *PLoS Genet* **7**, e1002051 (2011).
11. Aulchenko, Y.S. *et al.* Predicting human height by Victorian and genomic methods. *Eur J Hum Genet* **17**, 1070-5 (2009).
12. Lee, S., van der Werf, J., Hayes, B., Goddard, M. & Visscher, P. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLOS Genet* **4**, e1000231 (2008).
13. Clark, S.A., Hickey, J.M., Daetwyler, H.D. & van der Werf, J.H. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genet Sel Evol* **44**, 4 (2012).
14. Legarra, A., Robert-Granie, C., Manfredi, E. & Elsen, J.M. Performance of genomic selection in mice. *Genetics* **180**, 611-8 (2008).
15. Habier, D., Fernando, R.L. & Garrick, D.J. Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. *Genetics* **194**, 597-607 (2013).
16. Pszczola, M., Strabel, T., Mulder, H.A. & Calus, M.P. Reliability of direct genomic values for animals with different relationships within and to the reference population. *J Dairy Sci* **95**, 389-400 (2012).
17. Hayes, B., Visscher, P. & Goddard, M. Increased accuracy of artificial selection by using the realized relationship matrix. *Genet Res* **91**, 47 - 60 (2009).
18. van der Werf, J.H.J., Clark, S.A. & Lee, S.H. Predicting genomic selection accuracy from heterogeneous sources. in *Association for the Advancement of*

Animal Breeding and Genetics Conference Vol. 21 161 (AAABG, Lorne, Australia, 2015).

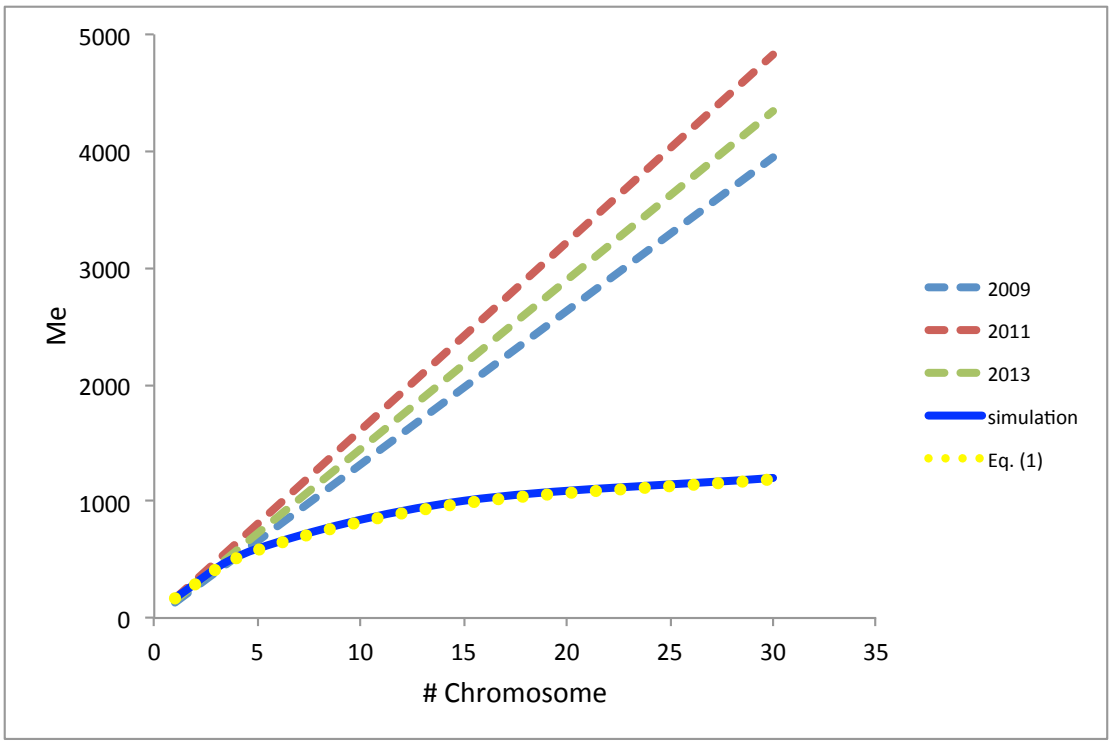
19. Wientjes, Y.C., Veerkamp, R.F. & Calus, M.P. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics* **193**, 621-31 (2013).
20. Rabier, C.-E., Barre, P., Asp, T., Charmet, G. & Mangin, B. On the Accuracy of Genomic Selection. *PLoS ONE* **11**, e0156086 (2016).
21. Meuwissen, T., Hayes, B. & Goddard, M. Accelerating Improvement of Livestock with Genomic Selection. *Annual Review of Animal Biosciences* **1**, 221-237 (2013).
22. Wientjes, Y.C.J., Bijma, P., Veerkamp, R.F. & Calus, M.P.L. An Equation to Predict the Accuracy of Genomic Values by Combining Data from Multiple Traits, Populations, or Environments. *Genetics* **202**, 799-823 (2016).
23. Wray, N.R. *et al.* Research Review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry* **55**, 1068-1087 (2014).
24. Daetwyler, H.D., Villanueva, B. & Woolliams, J.A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLOS ONE* **3**, e3395 (2008).
25. MacCluer, J.W., VandeBerg, J.L., Read, B. & Ryder, O.A. Pedigree analysis by computer simulation. *Zoo Biology* **5**, 147-160 (1986).
26. Lee, S. & van der Werf, J. The efficiency of designs for fine-mapping of quantitative trait loci using combined linkage disequilibrium and linkage. *Genetics Selection Evolution* **36**, 145-161 (2004).
27. Roach, J.C. *et al.* Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science* **328**, 636-639 (2010).
28. Splansky, G.L. *et al.* The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *Am J Epidemiol* **165**, 1328-35 (2007).
29. Silventoinen, K. *et al.* Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res* **6**, 399-408 (2003).
30. Macgregor, S., Cornes, B.K., Martin, N.G. & Visscher, P.M. Bias, precision and heritability of self-reported and clinically measured height in Australian twins. *Hum Genet* **120**, 571-80 (2006).
31. Visscher, P.M. *et al.* Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet* **2**, e41 (2006).
32. Elks, C.E. *et al.* Variability in the heritability of body mass index: a systematic review and meta-regression. *Front Endocrinol (Lausanne)* **3**, 29 (2012).
33. Wilson, J.G. *et al.* Study design for genetic analysis in the Jackson Heart Study. *Ethn Dis* **15**, S6-30-37 (2005).
34. Goddard, M.E. & Hayes, B.J. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet* **10**, 381-391 (2009).
35. Su, G. *et al.* Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. *Journal of Dairy Science* **95**, 4657-4665 (2012).

36. VanRaden, Paul M., O'Connell, Jeffrey R., Wiggans, G.R. & Weigel, K.A. Genomic evaluations with many more genotypes. *Genetics Selection Evolution* **43**, 1-11 (2011).
37. Moghaddar, N., Swan, A.A. & van der Werf, J.H.J. Accuracy of genomic prediction for Merino wool traits using high-density marker genotypes. *Proc. Assoc. Advmt. Breed. Genet.* **21**, 165-168 (2015).
38. Lee, S.H. & Wray, N.R. Novel genetic analysis for case-control genome-wide association studies: quantification of power and genomic prediction accuracy. *PLOS ONE* **8**, e71494 (2013).

A.


900
800
700
600
500
400
300
200
100
0
Me
0
5
10
15
20
25
30
35
Chromosome
2009
2011
2013
simulation
Eq. (1)

B.


5000
4000
3000
2000
1000
0
Me
0
5
10
15
20
25
30
35
Chromosome
2009
2011
2013
simulation
Eq. (1)

C.

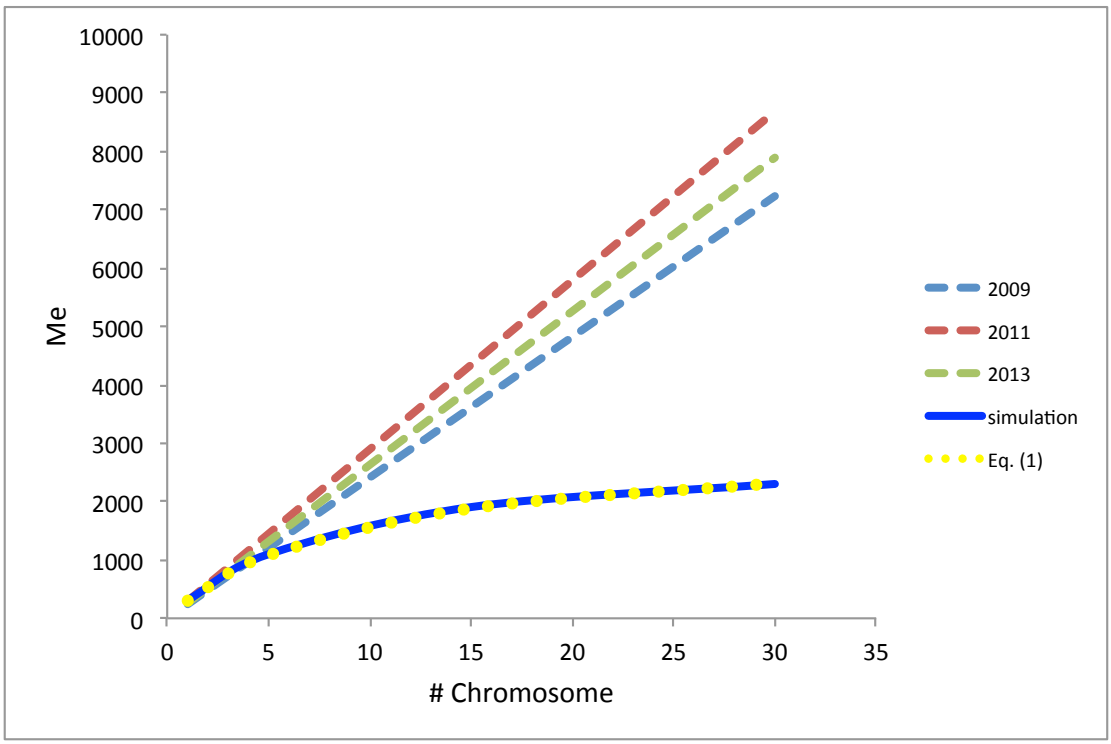
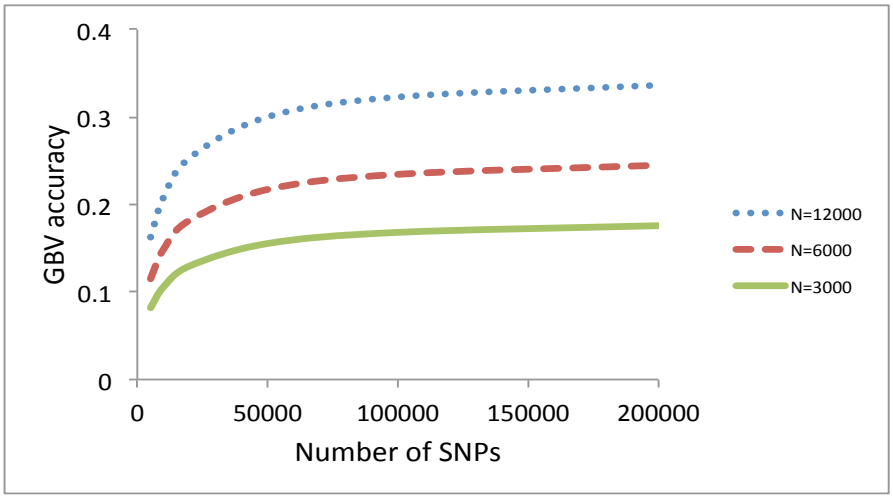
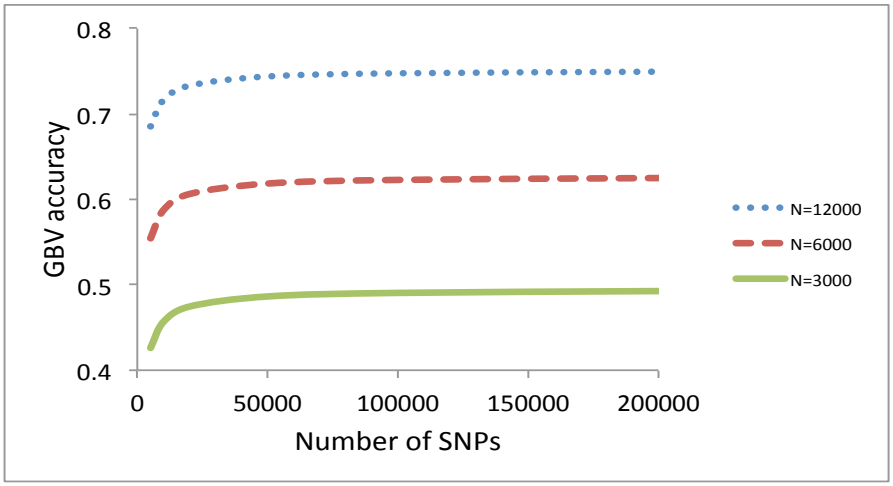


Figure 1. Expected effective number of chromosome segments (M_e) from previous studies in 2009[6], 2011[5] and 2013[21] and from Eq. (1) in this study, compared to empirically observed from M_e simulation when varying the number of chromosomes each with 1 Morgan long. Effective population size was used as N_e =50 **(A)**, 500 **(B)** and 1000 **(C)**. This confirms the result from Lee et al. (2016)[7].

A.



B.



C.

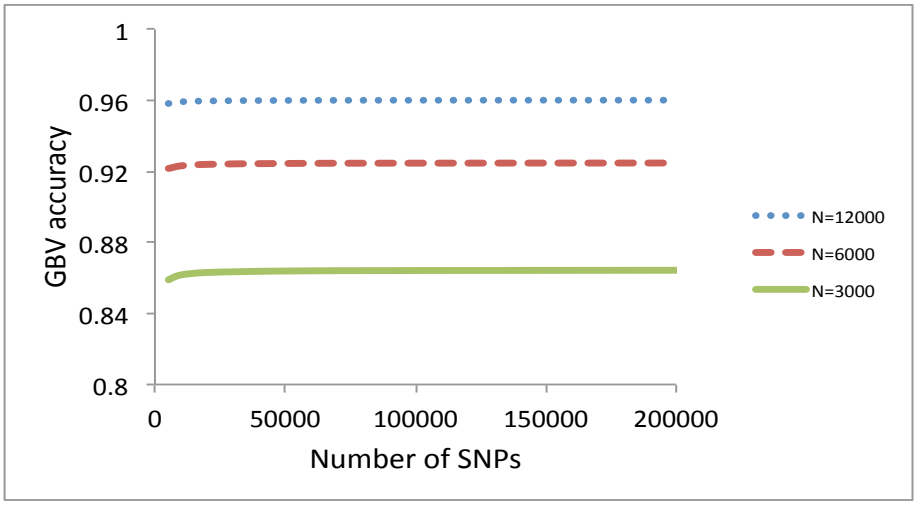


Figure 2. Accuracy of GBV when varying the number of SNPs for N_e = 10,000 **(A)**, 1000 **(B)** and 100 **(C)**. The sample size in the reference data was N=12,000, 6000 or 3000. The heritability was 0.25.

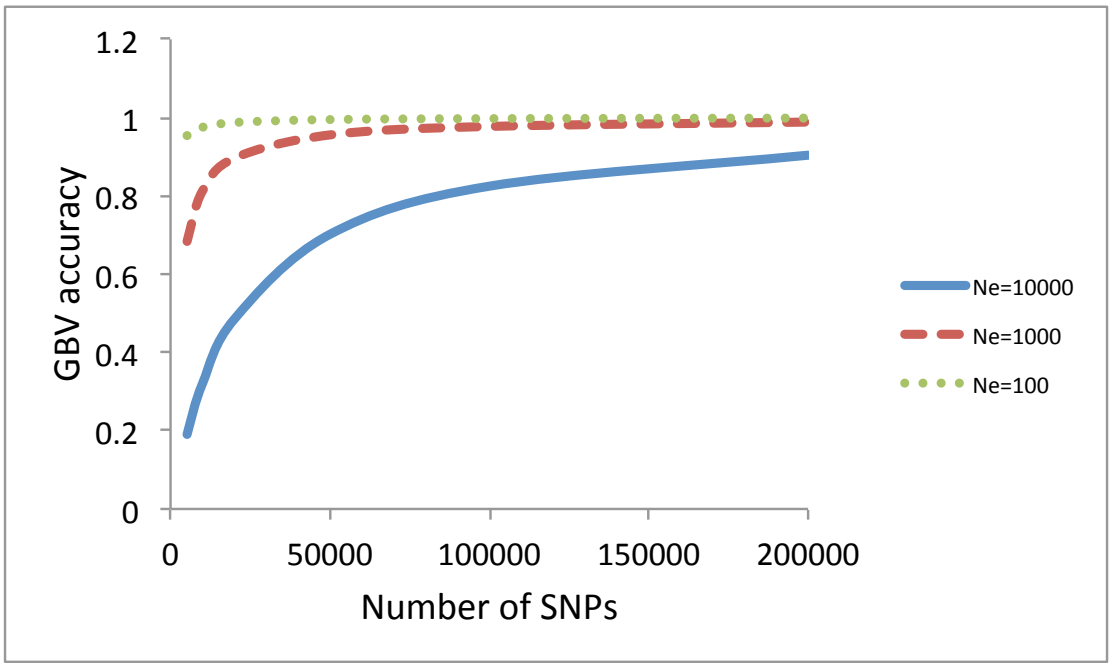
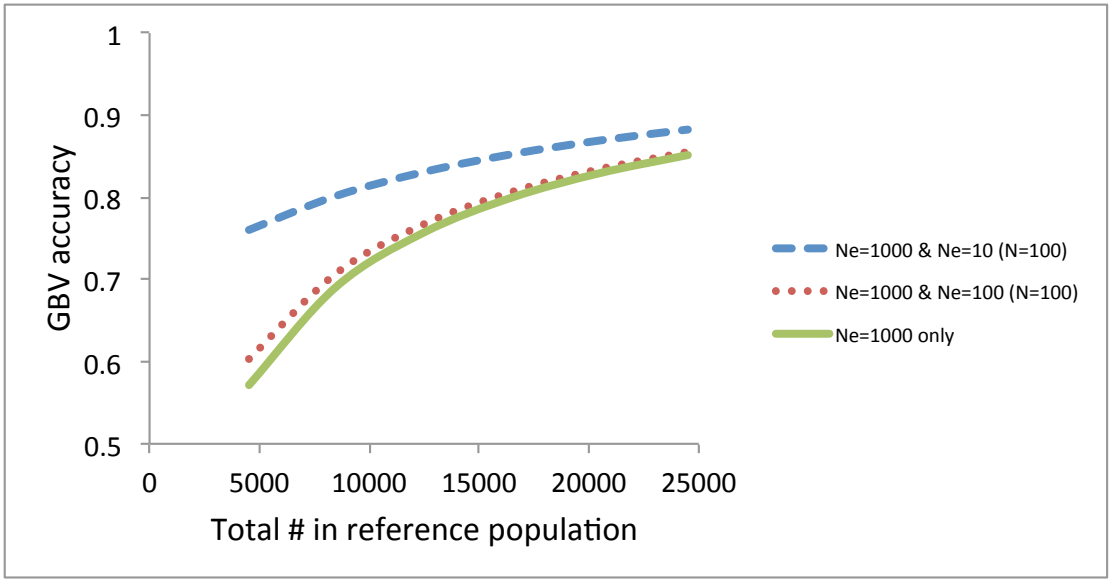


Figure 3. The proportion of genetic variance at QTL captured by markers (b = M / (M_e + M) when varying the number of SNPs for N_e = 10,000, 1000 and 100.

A.



B.

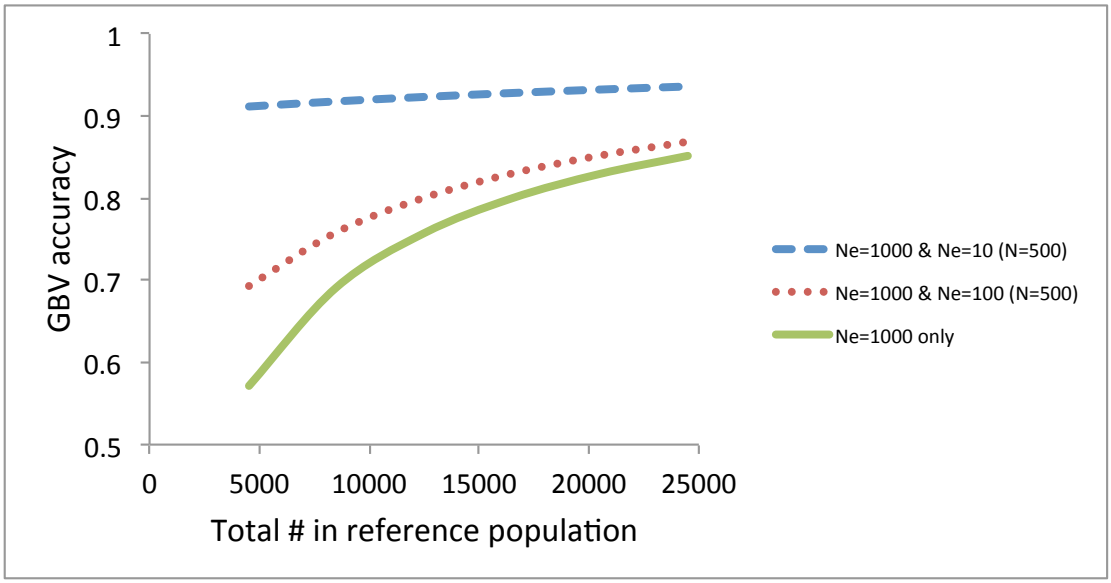
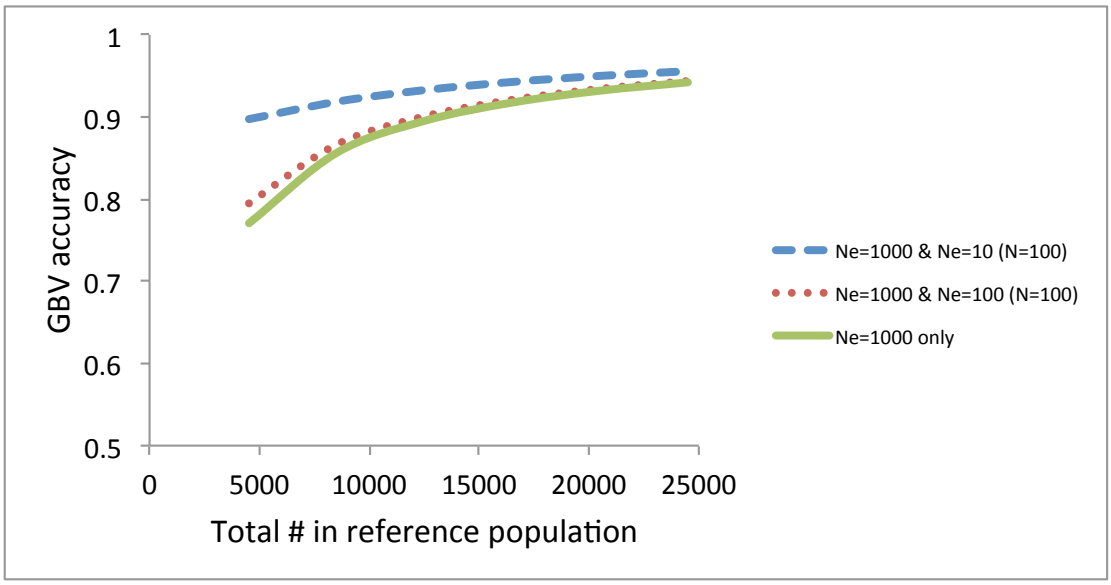


Figure 4. Accuracy of GBV when adding 100 individuals (N=100) **(A)** or 500 individuals (N=500) **(B)** of N_e=100 or N_e=10 to the reference population of N_e=1000. The heritability was 0.25.

A.



B.

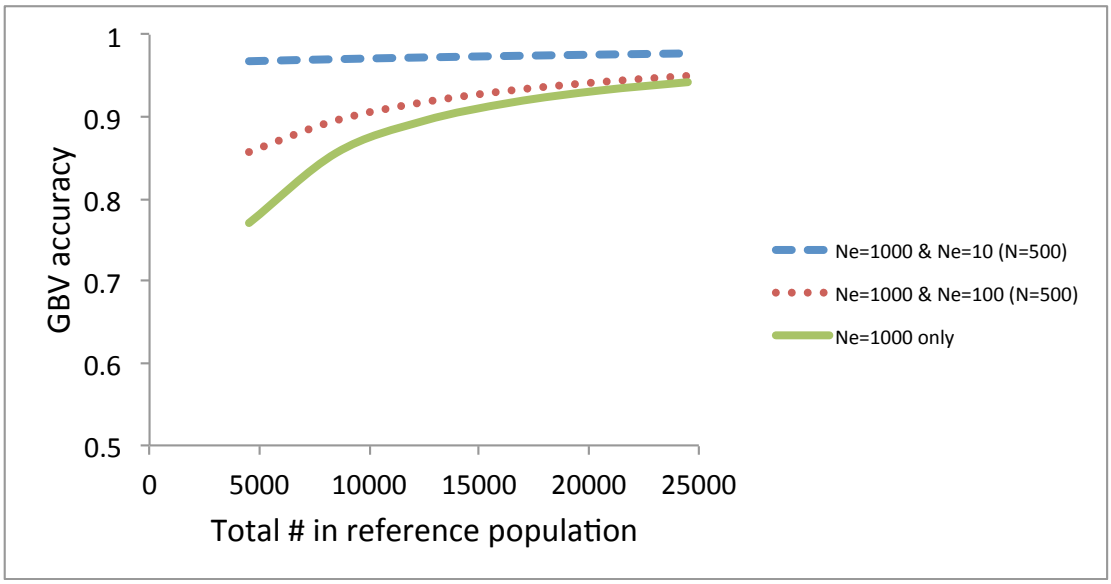
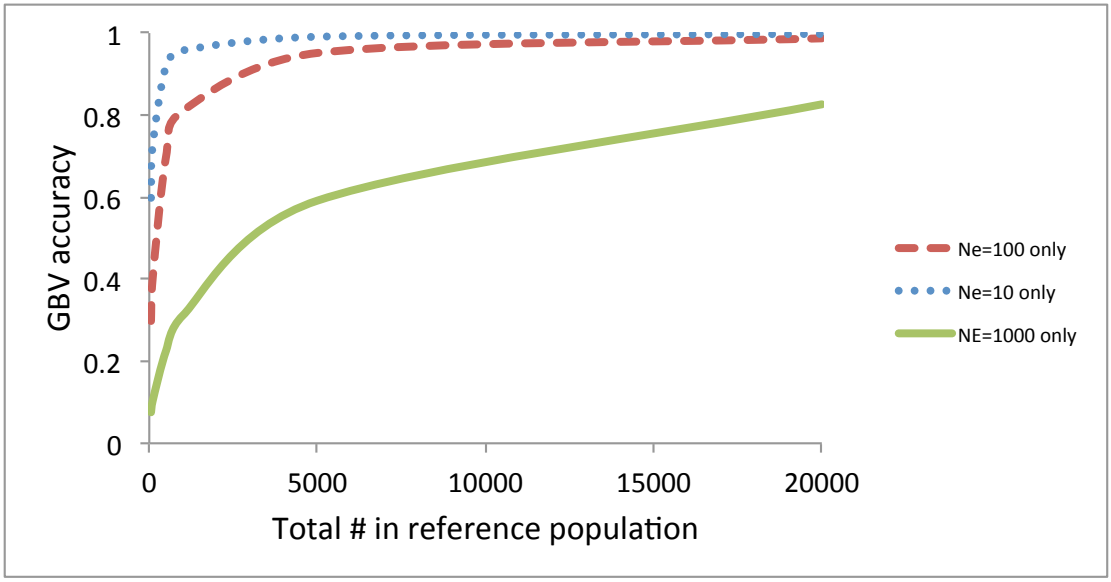


Figure 5. Accuracy of GBV when adding 100 individuals (N=100) **(A)** or 500 individuals (N=500) **(B)** of N_e=100 or N_e=10 to the reference population of N_e=1000. The heritability was 0.25. The heritability was 0.75.

A.



B.

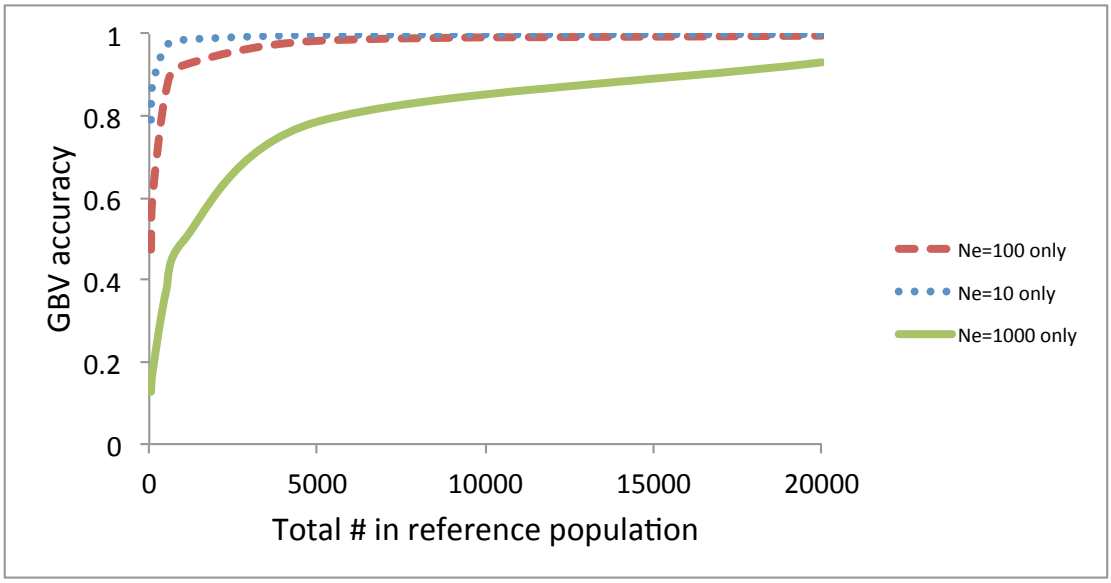
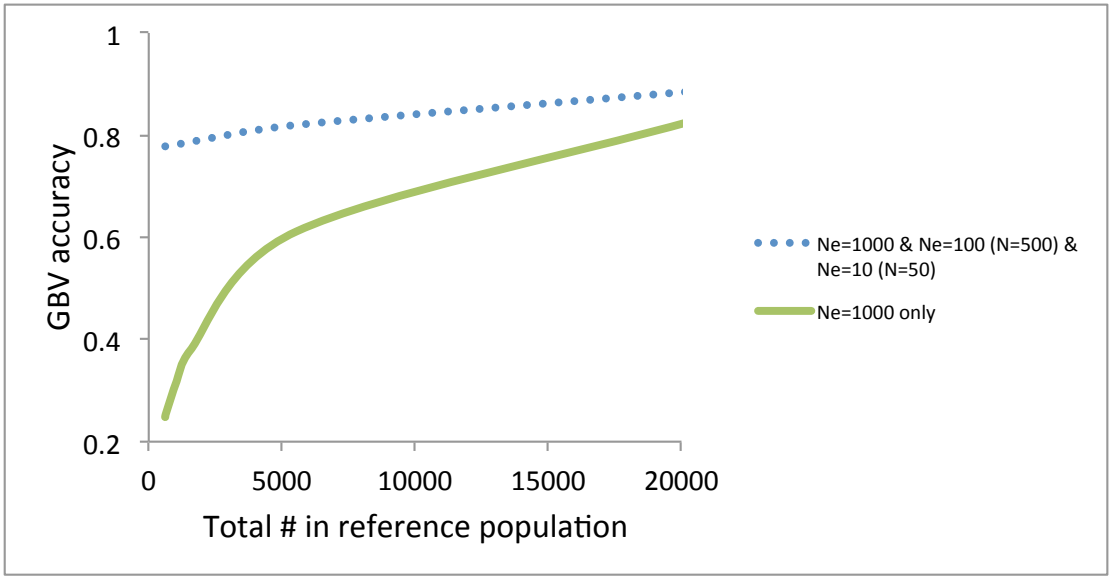


Figure 6. Accuracy of GBV when using N_e=1000 only, N_e=100 only and N_e=10 only with a heritability of 0.25 **(A),** and with a heritability of 0.75 **(B).**
For N_e=10 only, the accuracy at a sample size of 100 was 0.73 **(A)** and 0.88 **(B).** For N_e=100 only, the accuracy at a sample size of 1000 was 0.81 **(A)** and 0.92 **(B).** For N_e=1000 only, the accuracy at a sample size of 20,000 was 0.83 **(A)** and 0.93 **(B).**

A.



B.

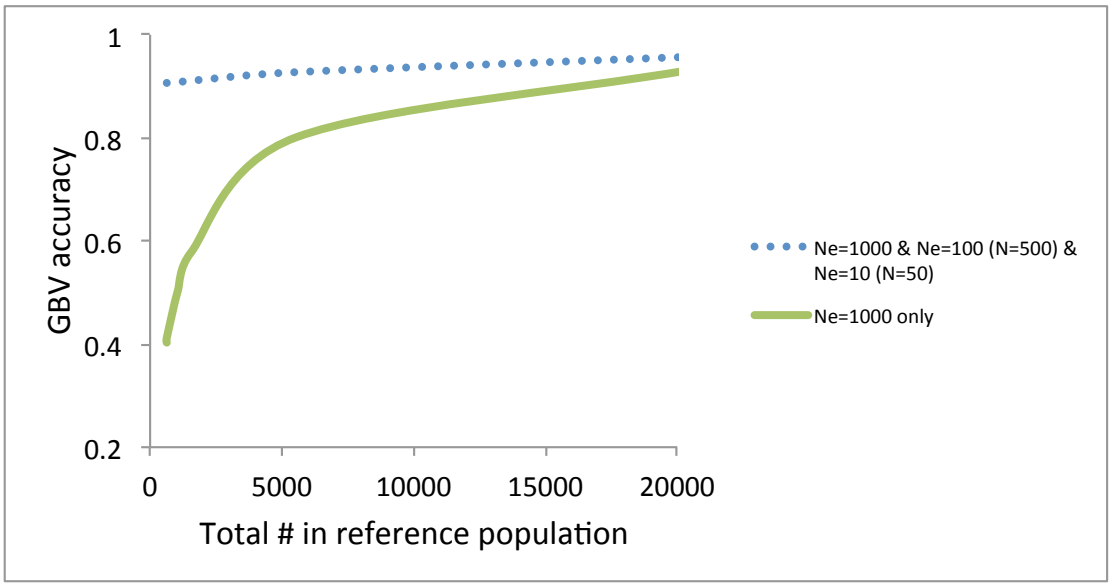


Figure 7. Accuracy of GBV when using a composite design, e.g. N_e=1000 + N_e=100 (N=500) + N_e=10 (N=50), compared to Ne=1000 only with a heritability of 0.25 **(A)** and with a heritability of 0.75 **(B).**

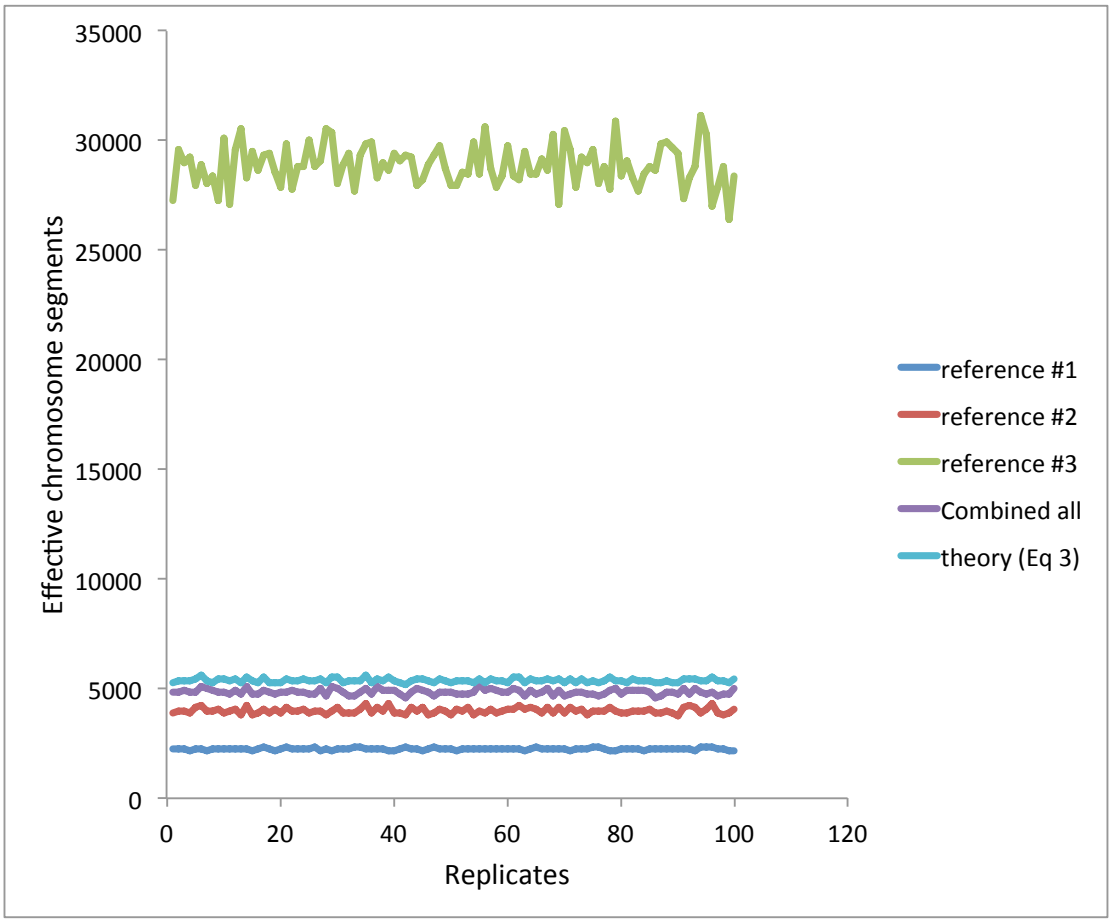


Figure 8. When using Framingham data, empirically estimated M_e based on each of the reference data sets and combined data. Empirically estimated M_e based on combined data is approximately agreed with that from theory (Eq. 4).

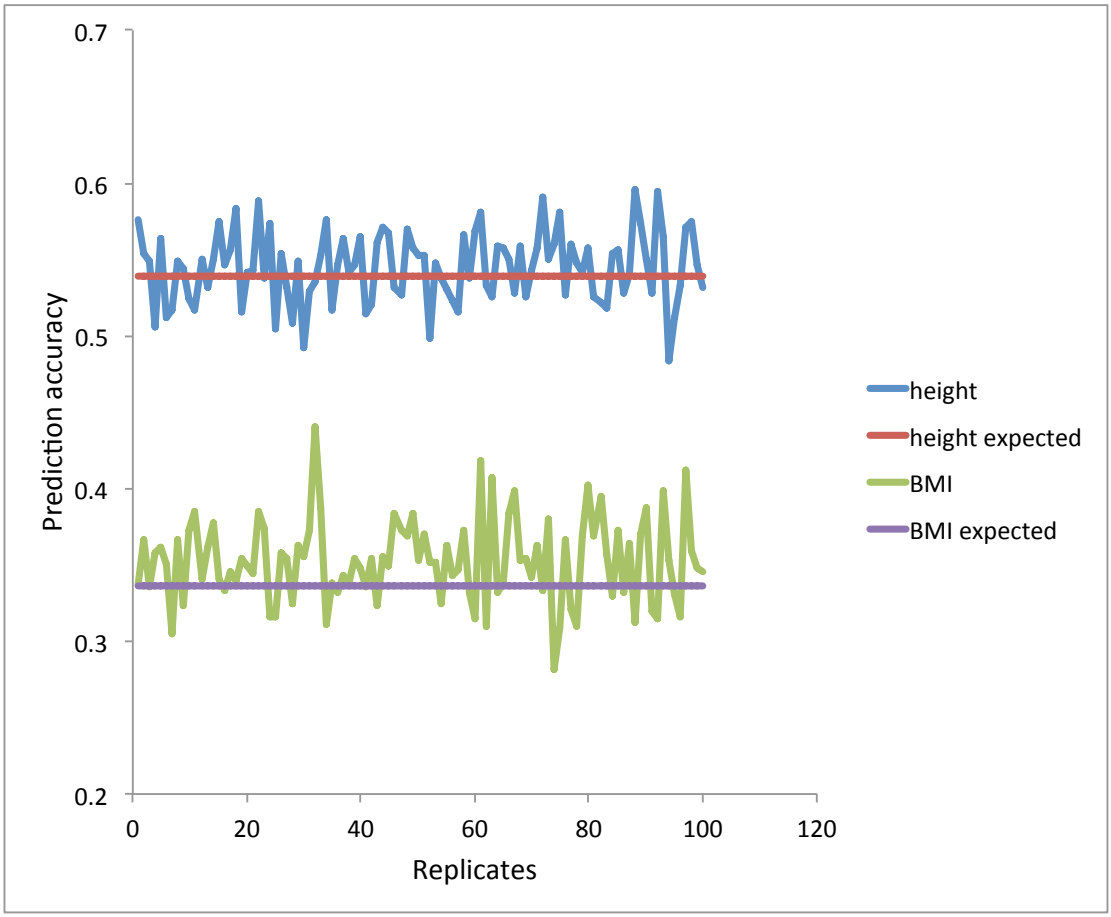


Figure 9. When using Framingham data, observed prediction accuracy and expected prediction accuracy with given M_e and N (from Eq. (2)) are agreed well. The reported heritability, h^2=0.8 [29-31] for height and h^2=0.46[32,33] for BMI, were used.

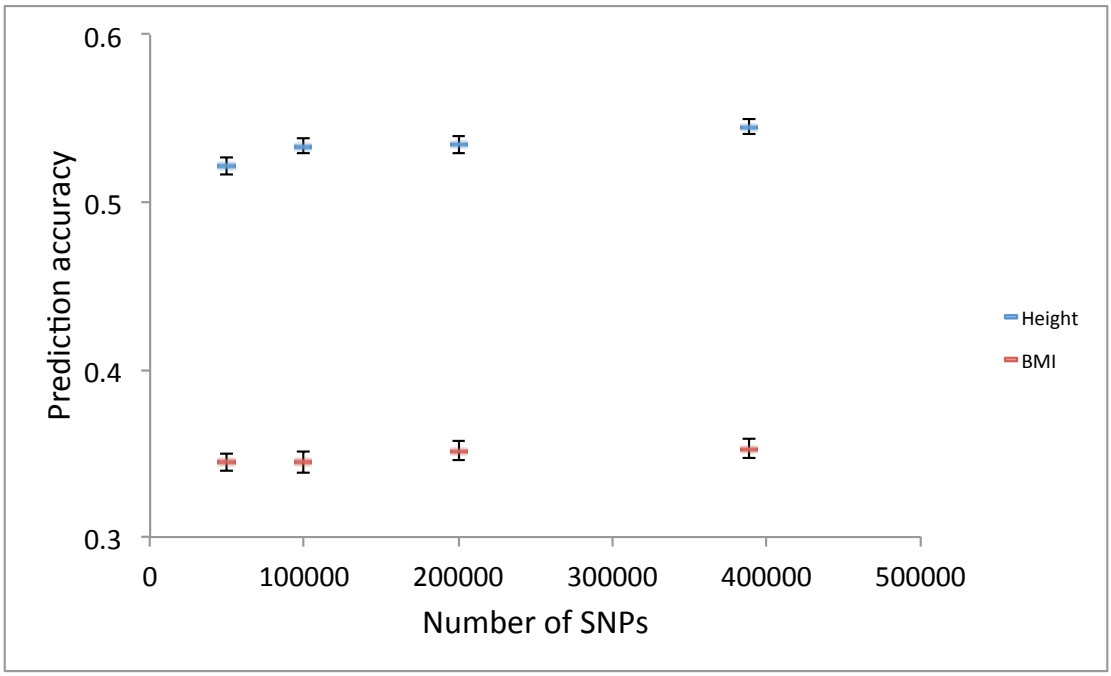


Figure 10. When using Framingham data, the prediction accuracy is not much decreased even with 50,000 SNPs that were randomly selected from 389,265 SNPs.