PLOS | ONE

RESEARCH ARTICLE

The Impact of Genetic Relationship and Linkage Disequilibrium on Genomic Selection

Hongjun Liu[1☯], Huangkai Zhou[2☯], Yongsheng Wu[3☯], Xiao Li[4], Jing Zhao[5], Tao Zuo[6], Xuan Zhang[2], Yongzhong Zhang[1], Sisi Liu[1], Yaou Shen[1], Haijian Lin[1], Zhiming Zhang[1], Kaijian Huang[3], Thomas Lübberstedt[5], Guangtang Pan[1]*

1 Maize Research Institute of Sichuan Agricultural University, Chengdu, China, **2** Guangzhou Genedenovo Biotechnology Co., Ltd, Guangzhou, China, **3** Guangxi Maize Research Institute, Guangxi Academy of Agricultural Sciences, Nanning, China, **4** Institute of Plant Protection, Sichuan Academy of Agricultural Sciences, Chengdu, China, **5** Department of Agronomy, Iowa State University, Ames, IA, United States of America, **6** Interdepartmental genetics program, Iowa State University, Ames, IA, United States of America

☯ These authors contributed equally to this work.
* pangt@sicau.edu.cn

OPEN ACCESS

Citation: Liu H, Zhou H, Wu Y, Li X, Zhao J, Zuo T, et al. (2015) The Impact of Genetic Relationship and Linkage Disequilibrium on Genomic Selection. PLoS ONE 10(7): e0132379. doi:10.1371/journal.pone.0132379

Editor: Lewis Lukens, University of Guelph, CANADA

Received: October 23, 2014

Accepted: June 13, 2015

Published: July 6, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: Funding for this research was provided by National Natural Science Foundation of China (31201221, 31271740, 31060191); the National Hi-Tech program of China (2012AA10307); Maize Research & Development Center, CARS-02-07; and the Key Laboratory Construction Program of Guangxi (12-071-09). Guangzhou Genedenovo Biotechnology Co., Ltd provided support in the form of salaries for authors HZ and XZ, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Genomic selection is a promising research area due to its practical application in breeding. In this study, impact of realized genetic relationship and linkage disequilibrium (LD) on marker density and training population size required was investigated and their impact on practical application was further discussed. This study is based on experimental data of two populations derived from the same two founder lines (B73, Mo17). Two populations were genotyped with different marker sets at different density: IBM Syn4 and IBM Syn10. A high-density marker set in Syn10 was imputed into the Syn4 population with low marker density. Seven different prediction scenarios were carried out with a random regression best linear unbiased prediction (RR-BLUP) model. The result showed that the closer the real genetic relationship between training and validation population, the fewer markers were required to reach a good prediction accuracy. Taken the short-term cost for consideration, relationship information is more valuable than LD information. Meanwhile, the result indicated that accuracies based on high LD between QTL and markers were more stable over generations, thus LD information would provide more robust prediction capacity in practical applications.

Introduction

With rapid development of high density genotyping technologies, such as SNP (single nucleotide polymorphism) arrays and GBS (genotyping by sequencing) [1–3], markers covering genomes at high density become available for application in plant and animal breeding. Traditional marker-assisted selection (MAS) focuses on markers, which are significantly associated with traits of interest. MAS has been shown to bias breeding value estimates [4,5]. This limitation can be overcome by genomic selection (GS), a promising approach for improving quantitative traits [6], which has been widely used in animal breeding [7]. In GS schemes, marker effects are estimated in a training population. Ideally, this training population is genetically

The specific roles of these authors are articulated in the 'author contributions' section.

Competing Interests: HZ and XZ are employees of Guangzhou Genedenovo Biotechnology Co. There are no patents, products in development, or marketed products to declare. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

closely related to the breeding population. Using the training population, both marker genotypes and trait phenotypes are combined in a statistical model. Effects of all markers are simultaneously estimated without statistical threshold. Breeding values of individuals related to the training population can then be predicted based on genomic estimated breeding values (GEBVs) from genotyping data only, using the marker effects estimated from the training population.

With high density genotypic data, there are not enough degrees of freedom to fit all marker effects to a data set of limited trait observations by least squares [5]. To overcome this limitation, several statistical models have been developed, including Bayesian shrinkage regression [6,8], random regression best linear unbiased prediction RR-BLUP [6,9], kernel regression [10] and machine learning methods [11,12]. Statistical models have been examined using empirical data of cattle [13], barley, maize, wheat, and Arabidopsis [14,15]. In most cases, RR-BLUP showed a robust prediction accuracy for different genetic architectures, especially when marker density was low to medium.

Accuracy of GS depends on the characteristics of the training population and validation population, such as heritability of traits [16], extent of linkage disequilibrium (LD), genetic distance [17], and genetic relationship between two populations [18,19]. Expected accuracy depends on marker density, effective sample size, and the number of included phenotypes [16]. Marker density greatly affects the costs of genotyping and is, therefore, a key factor in large scale application of GS [20]. Marker number had very little effect on prediction accuracies within families from various plant species, if marker densities were not very low (hundreds of markers with interval of 8~25cM for most traits) [14]. High marker coverage (thousands of markers or more) was needed, when GS was applied in a more diverse population [21]. In biparental populations, prediction accuracy is mostly due to genetic relationship captured by markers that contribute most to trait expression. However, if the genetic relationship between individuals is weak in a diverse population, dense coverage of markers is required to capture most of the markers in LD with QTL. The effect of these two factors for accuracy of GS were addressed in a simulation study [22], but empirical data in plants have not yet been reported.

Here, we used empirical data to study the contribution of genetic relationship and LD on prediction accuracy of GS under different scenarios due to different population structures, marker densities, and progeny numbers. Our objectives were, to (1) study the impact of LD on the number of markers and progenies required to obtain acceptable prediction accuracy within or across populations, and (2) investigate the effect of genetic relationship and LD on prediction accuracy of GS.

Materials and Methods

Genotypic and phenotypic data

Data from two maize (*Zea mays* L.) populations, IBM Syn10 [23] and IBM Syn4 [24], were used for this study (hereafter referred to as Syn10 and Syn4). In short, Syn10 is a B73×Mo17 doubled haploid (DH) population obtained after 10 generations of intermating, and Syn4 is a B73×Mo17 recombinant inbred line population obtained after four generations of intermating. Both Syn4 and Syn10 populations were planted in separate but adjacent experiments at the Agronomy Agricultural Engineering Research Center, Ames, Iowa, in 2006 and 2007. Performance *per se* was tested for each line. Growing degree days (GDD) were calculated in °C day from planting until the date, when at least 50% of the tassels in the plots were shedding pollen. Plant height (PH) was measured from soil surface to the flag leaf collar on five representative plants within each plot. For each experiment, both populations were repeated twice in a row column alpha lattice experimental design of seven columns and 37 rows. The two parents B73

and Mo17 were planted in a randomized fashion eight times within replications in each experiment, respectively. Each plot size consisted of 5.3 x 1.5 m^2, at a density of 69,187 plants/ha. 15 days prior to planting in May, 175 kg urea/ha were applied, as well as Metolachlor and Atrazine herbicides at a rate of 1.86 and 1.12 kg of active ingredient per ha. Neither herbicides nor insecticides were applied on the experiments after planting.

Genotype data for a set of 244 Syn4 recombinant inbred lines (RILs) are available at MaizeGDB [25], with 1339 polymorphic markers covering an approximately 6,240 cM (CentiMorgan) linkage map. A set of 194 Syn10 DH lines was genotyped using a genotyping by low coverage whole genome sequencing procedure [26]. Briefly, the sequencing reads from DH lines were aligned to the B73 reference genome. The genotype of each genomic region in 100 kb increments was determined by the proportion of reads derived from both parents, and adjacent regions with completely identical genotype in each line were integrated as bin-marker. A high-density genetic map of Syn10 consists of 6611 bin markers, with a genetic distance of 11198.5 cM (unpublished data).

Imputation, LD, and kinship measures

Missing marker genotypes of Syn4 were imputed using a function in the package *R/QTL* that uses a hidden Markov model to predict missing marker genotypes, given observed multipoint marker data [27]. Considering convenience of subsequent comparisons between Syn4 and Syn10, genotypes from Syn10 were imputed onto Syn4 by first determining the physical position of the two markers sets in the B73 reference genome (B73 RefGen_v2) [28]. Then, for each Syn10 marker, the unknown Syn4 genotype values were imputed based on the nearest flanking markers in a given Syn4 line, similar to the method used to impute missing marker values in Syn4. Finally, all 6611 bin markers from Syn10 were imputed into Syn4 and used for further analyses.

The degree of LD between markers was quantified using the parameter r^2 [29], estimated using *GOLD* software [30]. LD within populations (Syn4, Syn10) and across populations was measured. Average LD was calculated in increments of 100 kb, according to marker distances. The realized relationship (kinship) matrix of two populations was calculated using TASSEL software [31] with imputed genotype data.

BLUP of marker breeding values

As the aim of our project was to predict breeding values (BV), we used models to fit only the additive effects at each marker. The method used for BV prediction in this study was RR-BLUP [6]. The phenotypic values for a set of progenies were modeled as $y = \mu l + Xg + e$, where y is an $N_p \times l$ vector of phenotypic means of the progenies; l is an $N_p \times l$ vector of 1s (N_P was number of individuals in training populations); μ is the overall mean; X is an $N_p \times N_m$ matrix of marker genotype indicators (N_M is number of markers in model), with elements $X_{ij} = 0$ or 2, if the genotype of line i at marker j is AA, or BB, respectively; g is an $N_m \times l$ vector of marker breeding value; and e is an $N_p \times l$ vector of residuals.

Estimates of genetic variance (V_g) and residual variance (V_e) were obtained from an analysis of variance of phenotype within the two IBM populations, and a "mixed population" after combination of both IBM populations. In the following parts, we used V_g and heritability (h^2) estimates from mixed population data (Table 1). The variance of breeding values at each marker locus was assumed to be equal to ${}^{V_g}/_{N_m}$, g is assumed to be normally distributed, $g_i \sim N(0, \sigma_g^2)$.

Table 1. Genetic variance (σ_g^2) and heritability (h^2) in different populations.

Population	Number of progenies	Total genetic distance (cM)	GDD		PH	
			σ_g^2	h^2	σ_g^2	h^2
Syn10	194	11,198.50	895.96**	0.79	279.18**	0.92
Syn4	244	6,240	1,074.74**	0.81	298.81**	0.90
Mixed population	438	-	1,022.71**	0.80	296.21**	0.90

**Significantly different from zero at the 0.01 level of probability.

doi:10.1371/journal.pone.0132379.t001

Data analyses and validation

We estimated the marker effects and predicted the genomic breeding values for seven different scenarios (Table 2). We studied the effect of number of markers and also number of progenies in the training population for estimating marker effects. Therefore, we randomly selected seven marker sets with different marker-genome coverage (100, 200, 400, 800, 1600, 3200, and 6611; Table 3) and an even distribution across the genetic map. The number of lines in training populations varied from 30 to 180 with an increment of 30 lines.

To obtain training populations within a population, lines were randomly selected from the original population. Then the remaining lines were classified as validation population by default. In scenarios 5–7, to obtain training populations from both IBM populations, an equal number of lines were randomly selected from Syn4 and Syn10, respectively. For each scenario (Table 2), sampling of training and validation set was repeated 100 times. The correlation between observed and predicted phenotypes (r_{MP}) was estimated for each repeat. The accuracy of genomic selection was expressed as $r_{MG} = {}^{r_{MP}}/_{h}$ [32,33], where h refers to the square root of heritability. Reported prediction accuracies are the mean r_{MG} values across 100 repeats. Least significant differences ($P = 0.05$, adjusted by the *Bonferroni* procedure for the effect of multiple comparisons) for r_{MG} were calculated for each scenario, with the combinations of N_P and N_M as independent variables. All data analyses were done in R [34].

Results

LD and kinship in two populations

To understand the pattern of LD decay, estimates of pair-wise LD were averaged in increments of 100 kb distance between markers (Fig 1). The average physical distance between adjacent

Table 2. Scenarios for GS tests.

Scenario	Training population	Validation population	Prediction Type
1	Syn10	Syn10	Within population[b]
2	Syn4	Syn4	Within population
3	Syn10	Syn4	Between populations[c]
4	Syn4	Syn10	Between populations
5	Mixed population[a]	Mixed population	Across populations[d]
6	Mixed population	Syn10	Across populations
7	Mixed population	Syn4	Across populations

[a] Mixed population was a combination of the Syn4 and Syn10 populations

[b] Lines for training and validation came from the same population

[c] Lines for training and validation came from different populations

[d] Lines for training came from both populations.

doi:10.1371/journal.pone.0132379.t002

Table 3. Average distance and LD between adjacent markers in different marker sets.

Marker set (N_M)	Physical Distance (Mb) [a]	Genetic Distance [b] in Syn4 (cM)	Genetic Distance in Syn10 (cM)	LD [c] in Syn10 (r^2)	LD in Mixed Population (r^2)	LD in Syn4 (r^2)
6,611	0.31	0.94	1.69	0.78	0.81	0.87
3,200	0.64	1.95	3.50	0.66	0.71	0.78
1,600	1.28	3.90	7.00	0.50	0.55	0.63
800	2.57	7.80	14.00	0.32	0.37	0.44
400	5.13	15.60	28.00	0.19	0.22	0.27
200	10.26	31.20	55.99	0.09	0.11	0.15
100	20.53	62.40	111.99	0.05	0.06	0.08

[a] Average physical distance (in Mb) between adjacent markers.

[b] Average genetic distance (in cM) between adjacent markers.

[c] Linkage disequilibrium as estimated by the mean pairwise r^2 values between adjacent markers.

doi:10.1371/journal.pone.0132379.t003

markers varied in different marker sets (Table 1), and average LD between adjacent markers (ALAM) in different marker sets was estimated according to the average distance. Results were listed in Table 3. The ALAM of full marker sets (N_M = 6611) was 0.78, 0.87, and 0.81 in Syn10, Syn4, and mixed population (Table 3), respectively. A high degree of LD was observed even for extended distances between markers. For example, for the marker set of N_M = 800 for Syn10, the average distance between adjacent markers was 2.57 Mb (which equates to 14 cM), but the corresponding LD was 0.32. For marker set of N_M = 400 for Syn4, the average distance between adjacent markers was 5.13 Mb (equal to 15.6 cM), but the corresponding LD was 0.27. As

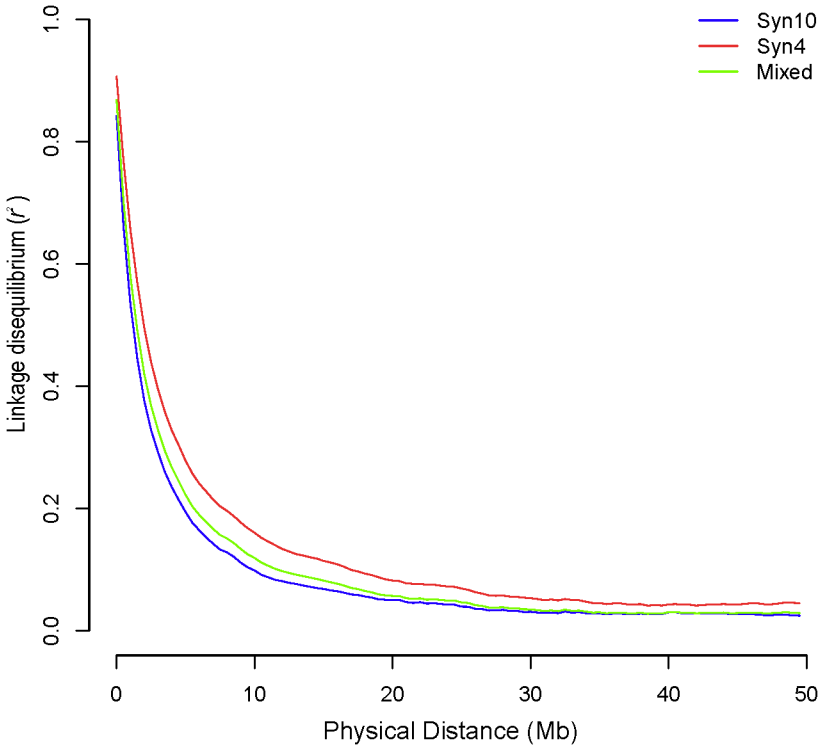

Fig 1. LD decay measured in different populations.

doi:10.1371/journal.pone.0132379.g001

expected, the LD in Syn4 was higher than in Syn10 and in the mixed population. The average realized genetic relationship (kinship) between any progeny pair was 0.63 and 0.44 for Syn4 and Syn10, respectively.

Trend of prediction accuracies across scenarios

A nonlinear increase in prediction accuracies with increasing size of N_P and N_M was observed for both traits within both populations (Fig 2, S1 Table). Generally, the highest r_{MG} values were obtained for the highest N_P, and an increase in N_M generally led to increased accuracies. Increase in accuracies was smooth and did not reach an obvious plateau, while increasing N_P using most fixed N_M of all scenarios (S1 Fig). However, the effect of increasing N_M was different under fixed N_P in different scenarios.

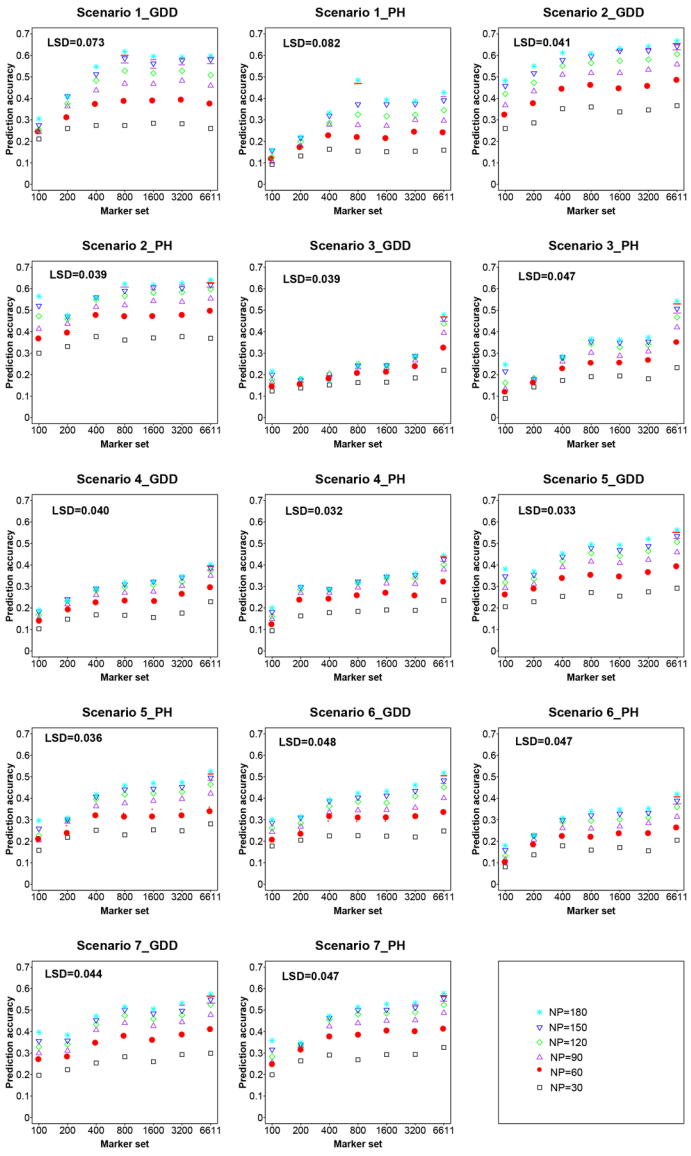

Fig 2. Prediction accuracies depending on marker number (N_M) and training population size (N_P) for different scenarios and traits.

doi:10.1371/journal.pone.0132379.g002

Trend of prediction accuracy within populations

Within populations (Scenario 1 and Scenario 2), accuracy increases became negligible, reached a plateau, or even decreased, when a certain level of N_M was reached. In Scenario 1, when using N_P = 180 for predicting GDD, there was no significant difference in r_{MG} for N_M from 400–6611. The prediction accuracy for GDD was higher than that for PH (S1 Table). The highest r_{MG} was obtained (0.49) when N_P = 180 and N_M = 800, although it was not significantly different from r_{MG} for the highest N_P and N_M. In Scenario 2, when marker coverage reached N_M = 1600 (for GDD) or 800 (for PH), r_{MG} did not significantly increase, if marker density was continually increased. However, when N_P = 180, r_{MG} for the lowest marker density N_M = 100 could not be ignored. Using N_P = 180, there were no major differences between r_{MG} of N_M = 100 and N_M = 6611(0.48 *vs*. 0.67 for GDD, 0.57 *vs*. 0.64 for PH). However, in Scenario 2 of PH, r_{MG} of N_M = 100 was significantly higher compared to N_M = 200, when N_P = 150 (0.52 *vs*. 0.47) or 180 (0.57 *vs*. 0.48) (S1 Table).

Trend of prediction accuracies between populations

Both in Scenario 3 and Scenario 4, under the same N_P, r_{MG} of N_M = 6611 was significantly higher than for any other marker set, even compared to the next lower marker density N_M = 3200, whereas there was no significant difference between most other adjacent marker densities. However, the average difference between N_M = 6611 and N_M = 3200 under different N_P was more pronounced in Scenario 3 compared to Scenario 4 (0.13 *vs*. 0.05 for GDD, 0.12 *vs*. 0.07 for PH). In Scenario 3, for the highest N_M and N_P, r_{MG} for GDD was 0.48, and r_{MG} for PH was 0.55. In Scenario 4, for the highest N_M and N_P, r_{MG} for GDD was 0.40, and r_{MG} for PH was 0.44. In Scenario 4, r_{MG} was higher than that for Scenario 3 under most different N_M, but r_{MG} of the highest marker sets N_M = 6611 was lower than that for Scenario 3.

Trend of prediction accuracies for the combined IBM populations

For mixed population (Scenarios 5, 6, and 7), the increased accuracies did not reach a plateau, when increasing N_M. Similarly, with the trend of prediction accuracies between populations (Scenarios 3 and 4), r_{MG} of N_M = 6611 was higher than for any other marker sets, even for the adjacent marker density N_M = 3200 under the same N_P. However, not all comparisons showed significant differences. The average difference of r_{MG} between N_M = 6611 and any other marker set was higher than differences between any of the marker set pairs. For example, in Scenario 5, the average difference of r_{MG} between N_M = 6611 and 3200 was 0.04 or 0.03, respectively, compared to 0.02 or 0.004 between N_M = 3200 and 1600 for GDD or PH. In these three scenarios, the highest r_{MG} was found for a combination of the highest N_P and N_M. The value of highest r_{MG} of Scenario 5 was between those for Scenario 6 and Scenario 7 (Scenarios 5, 6, 7 were 0.56, 0.52, 0.57 for GDD, and 0.53, 0.42, 0.58 for PH).

Discussion

The accuracy of RR-BLUP is due to two main factors, genetic relationship between training and validation individual, and LD between markers and QTL [19,35]. Traditional pedigree-BLUP (Henderson 1975) uses pedigree relationship to determine expected genetic relationship, while RR-BLUP also captures genetic relationships by markers as genomic relationship. Genomic relationships based on genotype data has the capacity to explain actual genetic relationships, deviating from pedigree-based relationships as a consequence of Mendelian sampling [36,37]. Goddard (2009) [38] showed that RR-BLUP is statistically equivalent to a BLUP model, without explicit marker effects and markers only being used to estimate the relationship

between lines. Through a simulation study, Habier *et al.* (2007) [19] showed that markers in LE (linkage equilibrium) with QTL could capture genetic relationships, and therefore, affect accuracy of prediction. However, this study also showed that LD between markers and QTL contributed a substantial proportion of the accuracy of RR-BLUP, and accuracy due to LD was more persistent than accuracy due to genetic relationship over generations. Based on the study, we will discuss the effect of genetic relationship within the Syn10 or Syn4 population, respectively, and then explore the potential contribution of LD for prediction accuracy between populations.

Effect of genetic relationship on prediction accuracy

GS with r_{MG} of only 0.5 can still result in a two fold higher gain per year compared to traditional MAS in a low-investment wheat breeding program (h^2 = 0.13) and a three fold increase in a high-investment maize breeding program (h^2 = 0.11) [39]. Lorenzana and Bernardo [14] proposed that 100 markers for bi-parental populations and 200–800 markers for random-mated maize populations are sufficient for genome-wide prediction of genotypic values. Using a panel of 788 individuals obtained from a half-diallel cross between four dent inbreeds, [40] showed that marker density is not a major limitation for the accuracy of GS. In scenarios within the two IBM populations (Scenarios 1 and 2), different combinations of N_P and N_M were validated to seek for the lowest marker density to obtain a r_{MG} comparable to 0.5. In Scenario 2, there was no r_{MG} in any marker set significantly lower than 0.5, when N_P = 180. One abnormal decrease of r_{MG} for PH was observed, when the marker density increased from 100 (ALAM = 0.08) to 200 (ALAM = 0.15) (the same phenomenon was observed in Scenario 3). The marker sets of N_M = 100 and N_M = 200 were randomly and independently selected from the whole marker set. Potentially, more markers of N_M = 100 were in LD with QTL of PH than for N_M = 200 by chance. In Scenario 1 of N_P = 180 and N_M = 100, r_{MG} of GDD and PH were only 0.304 and 0.159, respectively (compared to 0.48 and 0.57 in Scenario 2), and r_{MG} exceeded or was near 0.5 until marker density reached 400 (ALAM = 0.19) or 800 (ALAM = 0.32) for GDD or PH, respectively. Thus, more markers were needed in Syn10 to obtain the same accuracy as in Syn4.

The expected (coancestry) relationship within and between pairs of Syn4 and 10 lines does not differ as they derived from the same parents. However, the actual relationships deviated as a consequence of Mendelian sampling. Thus the realized genetic relationship, which is more accurate than expected genetic relationship, was calculated based on markers genotype. The average realized genetic relationship in Syn4 was higher than that in Syn10 (0.63 *vs*. 0.44). Possibly the higher number of generations of intermating in Syn10 reduces average genetic relationship between individuals compared to Syn4.

For N_M = 400 in Scenario 1, the average distance between adjacent markers was 5.1 Mb (28.0 cM) and LD was 0.08 (r^2 = 0.19). For N_M = 100 in Scenario 2, the distance and LD between adjacent markers was 20.5 Mb (62.4 cM) and LD was 0.19 (r^2 = 0.08). Accuracy due to realized genetic relationships can be regarded as lower bounds, if accuracy due to LD is small [22]. Therefore, prediction accuracy in this study was mainly due to realized genetic relationship as the contribution of LD should be limited using low density markers. In previous studies, the accuracy of GS was non-zero, even when the LD between marker and QTL was zero, as the GS model also captures realized genetic relationships [19,35]. However, r_{MG} of Scenario 2 was higher than that of Scenario 1 at lower marker density. We inferred that more benefits were obtained for related individuals from Syn4 due to higher realized genetic relationships compared to Syn10. This also explains, why the accuracy of Scenario 7 was higher than that of Scenario6. Although both scenarios used mixed populations as training population, the validation

populations were from Syn4 and Syn10 for Scenario 7 and Scenario 6, respectively. The difference between the highest r_{MG} of Scenario 1 and Scenario 2 was 0.05 and 0.18 for GDD and PH, respectively. Similarly, the difference between the highest r_{MG} of Scenario 6 and Scenario 7 was 0.06 and 0.16 for GDD and PH, respectively. We concluded that both differences were derived from different contributions of realized genetic relationships from Syn4 and Syn10 populations.

In addition to heritability and training population size, the effective number of quantitative trait loci (QTL) was also a factor influencing prediction accuracy [41–43]. Thus, the QTL numbers of two traits were investigated. The result showed that the average number of QTL of GDD for Syn10 and Syn4 was 17.5 and 17 respectively, and the number of QTL of PH for Syn10 and Syn4 was 17 and 16, respectively (unpublished data). The result excluded the effect of number of QTL for prediction accuracy between different Scenarios in this study.

In Scenario 3 and Scenario 4, no r_{MG} value (correlation between the true genotypic values and the predicted genotypic values based on markers) reached 0.4 when $N_M = 800$. It seems that the impact of genetic relationship is limited compared to Scenarios within populations. LD should be a good complement in this situation.

Effect of LD on prediction accuracy

Zhong *et al.* (2009) [18] discussed the effect of LD in RR-BLUP models in another simulation study. According to the study, the dichotomy between contributions "due to LD" *vs.* those "due to genetic relationship" is useful for considering the strengths of different methods, but the two contributions are confounded in practice. Indeed, in most empirical studies, the effect of LD and genetic relationship exist simultaneously. However, the accuracy due to LD may be a lower boundary for the accuracy of an individual that is unrelated to the training population [41]. In our study, due to being derived from the same founders B73 and Mo17, any pair of individuals across two populations was expected to share 50% of the genome. However, two populations have experienced generations of independent intermating. There is no close pedigree relationship between Syn4 and Syn10 lines (other than sharing the same parents), and the genetic relationship between both populations is complicated. This might explain, why Scenarios 3 or 4 did not result in a comparable accuracy at low marker density compared to Scenarios 1 or 2. As Scenarios 3 and 4 remove at least part of these effects due to a close direct genetic relationship, prediction accuracies depends mostly on LD between markers and QTL. It is thus excellent material for studying the effect of LD on prediction accuracy.

In Scenario 3, the training population Syn10 has a lower LD than Syn4. A higher marker density needed to maintain LD between markers and underlying QTL in Syn10 compared to Syn4. Low LD in the training population also means that recombination reduces the size of parental haplotypes, and that the effects of segments can be evaluated more accurately, given sufficient marker coverage. At low marker density in Scenario 3, the prediction accuracy increased slowly with increasing size of the training population. Even at $N_M = 3200$ (ALAM was 0.66), r_{MG} of both traits was lower than 0.4. When the marker density was near saturation ($N_M = 6611$,with ALAM of 0.78), r_{MG} increased substantially, and reached about 0.5 with $N_P = 180$. In Scenario 4, the situation was opposite, as the training population Syn4 has a higher LD than Syn10. Low density coverage with markers was sufficient to capture LD, but higher colinearity between adjacent markers (and underlying QTL) hinders evaluation of genome segment effects accurately. Prediction accuracy was higher for Scenario 4 compared to Scenario 3 at low marker coverage, as low density markers also capture part of LD in Scenario 4. With increasing marker density, the r_{MG} with $N_M = 6611$ and $M_P = 180$ was almost 0.1l lower in Scenario 4

than the corresponding r_{MG} in Scenario 3, as the effect of segments were estimated more accurately in Scenario 3.

In case of a weak genetic relationship between training and breeding population, LD contributed most of the prediction accuracy, thus marker density (LD between adjacent markers) becomes the most important factor. Calus and Veerkamp [44] suggested that an ALAM>0.15 was sufficient for a highly heritable trait. Our study indicated that the contribution of LD for prediction accuracy was limited, when the LD between adjacent markers was only 0.15 (about N_M = 200 for Syn4 in our study). Increasing N_P has almost no effect, until the marker density was high enough to capture LD and to make predictions between 'distant related' populations. We further infer that the prediction accuracy under low marker density is due to other factors, such as genetic relationships.

The effect of LD and pedigree relationship on GS in maize breeding

A large number of maize DH lines are produced every year. It seems impossible to evaluate all inbred lines for testcross performance in repeated field trials with limited breeding resources. The combination of DH technology and GS revolutionizes maize breeding [45]. Using GS, DH lines can be evaluated based on markers only, thus the cost of genotyping will become the key factor to determine the intensity of application of GS. If, for example, the cost of genotyping on a large scale is 10 cents per data point and the cost of growing a maize yield experimental plot is U.S. $20, genotyping for 200 markers would cost less ($20) than conducting yield trials at three locations ($60) [20]. Costs of genotyping can be expected to drop further, while costs of labor will likely increase (increasing field trial costs) in future.

In our study, a marker density ensuring a LD of 0.1 between adjacent markers was sufficient for acceptable prediction accuracies in bi-parental populations. If training and prediction populations are not closely related, the prediction accuracy will only depend on LD. In this situation, the number of markers required will be larger, a high LD with r^2>0.8 is needed for acceptable r_{MG} values. Undoubtedly, this will increase the cost of GS. Therefore, the genetic relationship and LD decay of the population will be of great importance in GS practice.

GS was tested by cross-validation for the Dent heterotic pool, as used on one side of the Flint × Dent pattern in central Europe. Prediction accuracies were near 0.8 for seven biomass- and bioenergy-related traits at a marker density of 5000 (LD between adjacent markers was about 0.1) [21]. The materials from the Dent heterotic pool were highly inbred. Genetic relationships were maintained between “unrelated” lines and could be tracked by genotyping data. In GS application practice, we could genotype and phenotype different populations presenting widely diverse genetic backgrounds to construct a training database. When a new population needs to be evaluated by GS, we can select the best training set from a training database according to the genetic relationship between prediction population and potential training set, to reduce the marker density requirement for prediction.

LD is an important factor that needs to be considered, when there is only a weak genetic relationship for prediction. In Scenarios 3 and 4, the impact of LD in the training population was observed. High LD in the training population not only means a lower marker density requirement to cover the genome, but also means a higher colinearity between linked markers, hindering evaluation of small genome segment effects accurately. This will reduce the robustness of prediction models for other populations. Therefore, if the prediction model is used temporarily, a training population with high LD is suggested for lower marker density. But if the prediction model is very important and will be used frequently, a low LD training population is recommended for extensive and persistent fitness over generations. A saturated marker density is suggested for such training populations to obtain an accurate and robust prediction model.

The marker density for the prediction population will be determined according to LD in the prediction population, and the process of imputation (such as for Syn4 in our study) will help to ensure consistent marker sets between training and prediction populations to obtain accurate prediction.

In our study, Syn4 and Syn10 were derived from the same parents and under high levels of identity by descent, however they didn't predict each other well. While the reason related to genetic relationship and LD was discussed above, there were two more potential factors that shouldn't be ignored. One was the genotyping error in syn10 and syn4 populations. We planned to randomly extract parts of the lines from two populations and validate their genotype by restriction-site associated DNA tags (RAD) sequencing. The other potential factor was that two populations grew in adjacent locations, which might induce extra environmental error. In the two populations, limited genetic diversity and clear genetic background help us simplify the analysis and draw a preliminary conclusion. However, the shortcoming of our study is also obvious. Diverse materials were used in practical breeding program, and several issues should be solved for application of genomics selection. For instance, lower identity by descent, faster LD decay, and three or more genotypes in one locus due to high genetic diversity. Next step, we will gradually extend our study to more diverse materials, so as to further study the impact of genetic relationship and LD in practice.

Supporting Information

S1 Table. Prediction accuracy (r_{MG}) obtained from the different combinations of N_P and N_M.
(XLSX)

S1 Fig. Prediction accuracy with increasing N_P using fixed N_M.
(TIFF)

Acknowledgments

Funding for this research were provided by National Natural Science Foundation of China (31201221, 31271740, 31060191); the National Hi-Tech program of China (2012AA10307); Maize Research & Development Center, CARS-02-07; and the Key Laboratory Construction Program of Guangxi (12-071-09).

Author Contributions

Conceived and designed the experiments: H. Liu GP KH TL ZZ. Performed the experiments: H. Liu HZ YW JZ TZ XZ YS YZ SL. Analyzed the data: H. Liu HZ YW JZ TZ XZ YS YZ SL. Contributed reagents/materials/analysis tools: XL TL. Wrote the paper: H. Liu XL H. Lin TL GP.

References

1. Ganal MW, Durstewitz G, Polley A, Berard A, Buckler ES, Charcosset A, et al. A large maize (Zea mays L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PloS one. 2011; 6(12):e28334. doi: 10.1371/journal.pone.0028334 PMID: 22174790
2. Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, et al. High-throughput genotyping by whole-genome resequencing. Genome research. 2009; 19(6):1068–76. doi: 10.1101/gr.089516.108 PMID: 19420380
3. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one. 2011; 6(5):e19379. doi: 10.1371/journal.pone.0019379 PMID: 21573248

4. Melchinger AE, Utz HF, Schön CC. Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. Genetics. 1998; 149(1):383–403. PMID: 9584111
5. Lande R, Thompson R. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics. 1990; 124(3):743–56. PMID: 1968875
6. Meuwissen THE, Hayes B, Goddard M. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157(4):1819–29. PMID: 11290733
7. VanRaden P, Van Tassell C, Wiggans G, Sonstegard T, Schnabel R, Taylor J, et al. Invited Review: Reliability of genomic predictions for North American Holstein bulls. Journal of dairy science. 2009; 92 (1):16–24. doi: 10.3168/jds.2008-1514 PMID: 19109259
8. Xu Y, Crouch JH. Marker-assisted selection in plant breeding: from publications to practice. Crop Science. 2008; 48(2):391–407.
9. Whittaker JC, Thompson R, Denham MC. Marker-assisted selection using ridge regression. Genetical research. 2000; 75(2):249–52. PMID: 10816982
10. Bennewitz J, Solberg T, Meuwissen T. Genomic breeding value estimation using nonparametric additive regression models. Genet Sel Evol. 2009; 41:20. doi: 10.1186/1297-9686-41-20 PMID: 19284696
11. Wei Z, Wang K, Qu H-Q, Zhang H, Bradfield J, Kim C, et al. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS genetics. 2009; 5(10):e1000678. doi: 10.1371/journal.pgen.1000678 PMID: 19816555
12. Long N, Gianola D, Rosa G, Weigel K, Avendano S. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. Journal of animal breeding and genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie. 2007; 124(6):377–89. PMID: 18076475
13. Luan T, Woolliams JA, Lien S, Kent M, Svendsen M, Meuwissen TH. The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. Genetics. 2009; 183(3):1119–26. doi: 10.1534/genetics.109.107391 PMID: 19704013
14. Lorenzana RE, Bernardo R. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theoretical and applied genetics. 2009; 120(1):151–61. doi: 10.1007/s00122-009-1166-3 PMID: 19841887
15. Crossa J, de Los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, et al. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics. 2010; 186 (2):713–24. doi: 10.1534/genetics.110.118521 PMID: 20813882
16. Hayes B, Bowman P, Chamberlain A, Goddard M. Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of dairy science. 2009; 92(2):433–43. doi: 10.3168/jds.2008-1646 PMID: 19164653
17. Solberg T, Sonesson A, Woolliams J. Genomic selection using different marker types and densities. Journal of animal science. 2008; 86(10):2447–54. doi: 10.2527/jas.2007-0010 PMID: 18407980
18. Zhong S, Dekkers JC, Fernando RL, Jannink J- L. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics. 2009; 182(1):355–64. doi: 10.1534/genetics.108.098277 PMID: 19299342
19. Habier D, Fernando R, Dekkers J. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007; 177(4):2389–97. PMID: 18073436
20. Bernardo R, Yu J. Prospects for genomewide selection for quantitative traits in maize. Crop Science. 2007; 47(3):1082–90.
21. Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R, et al. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nature genetics. 2012; 44(2):217–20 doi: 10.1038/ng.1033 PMID: 22246502
22. Habier D, Tetens J, Seefried F-R, Lichtner P, Thaller G. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics, selection, evolution: GSE. 2010; 42 (1):5.
23. Hussain T, Tausend P, Graham G, Ho J. Registration of IBM2 SYN10 doubled haploid mapping population of maize. Journal of Plant Registrations. 2007; 1(1):81–81.
24. Lee M, Sharopova N, Beavis WD, Grant D, Katt M, Blair D, et al. Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. Plant molecular biology. 2002; 48(5–6):453–61. PMID: 11999829.
25. Lawrence CJ, Harper LC, Schaeffer ML, Sen TZ, Seigfried TE, Campbell DA. MaizeGDB: the maize model organism database for basic, translational, and applied research. International journal of plant genomics. 2008; 2008.

26. Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, et al. A map of rice genome variation reveals the origin of cultivated rice. Nature. 2012; 490(7421):497–501. doi: 10.1038/nature11532 PMID: 23034647

27. Broman KW, Wu H, Sen , Churchill GA. R/qtl: QTL mapping in experimental crosses. Bioinformatics. 2003; 19(7):889–90. PMID: 12724300

28. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009; 326(5956):1112–5. doi: 10.1126/science.1178534 PMID: 19965430

29. Hill W, Robertson A. Linkage disequilibrium in finite populations. Theoretical and Applied Genetics. 1968; 38(6):226–31. doi: 10.1007/BF01245622 PMID: 24442307

30. Abecasis GR, Cookson W. GOLD—graphical overview of linkage disequilibrium. Bioinformatics. 2000; 16(2):182–3. PMID: 10842743

31. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007; 23(19):2633–5. PMID: 17586829

32. Dekkers J. Prediction of response to marker-assisted and genomic selection using selection index theory. Journal of animal breeding and genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie. 2007; 124(6):331–41. PMID: 18076470

33. Lee SH, van der Werf JH, Hayes BJ, Goddard ME, Visscher PM. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLoS genetics. 2008; 4(10):e1000231. doi: 10.1371/journal.pgen.1000231 PMID: 18949033

34. Team RDC R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012; http://wwwR-projectorg.

35. Habier D, Fernando RL, Garrick DJ. Genomic-BLUP Decoded: A Look into the Black Box of Genomic Prediction. Genetics. 2013.

36. Hill W, Weir B. Variation in actual relationship as a consequence of Mendelian sampling and linkage. 2011.

37. Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the realized relationship matrix. Genetics Research. 2009; 91(01):47–60.

38. Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009; 136(2):245–57. doi: 10.1007/s10709-008-9308-0 PMID: 18704696

39. Heffner E, Lorenz A, Jannink J-L, Sorrells M. Plant breeding with genomic selection: gain per unit time and cost. Crop science 2010; 50(5):1681–90.

40. Zhao Y, Gowda M, Liu W, Würschum T, Maurer HP, Longin FH, et al. Accuracy of genomic selection in European maize elite breeding populations. Theoretical and Applied Genetics. 2012; 124(4):769–76. doi: 10.1007/s00122-011-1745-y PMID: 22075809

41. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PloS one. 2008; 3(10):e3395. doi: 10.1371/journal.pone.0003395 PMID: 18852893

42. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 2010; 185(3):1021–31. doi: 10.1534/genetics.110.116855 PMID: 20407128

43. Combs E, Bernardo R. Accuracy of Genomewide Selection for Different Traits with Constant Population Size, Heritability, and Number of Markers. The Plant Genome. 2013; 6(1):1–7.

44. Calus M, Veerkamp R. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. Journal of animal breeding and genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie. 2007; 124(6):362–8. PMID: 18076473

45. Morrell PL, Buckler ES, Ross-Ibarra J. Crop genomics: advances and applications. Nature reviews Genetics. 2011; 13(2):85–96. Epub 2011/12/31. doi: 10.1038/nrg3097 PMID: 22207165.