Journal of Animal Science and Biotechnology

RESEARCH Open Access

Improving genomic predictions by correction of genotypes from genotyping by sequencing in livestock populations

Xiao Wang[1,2], Mogens Sandø Lund[1], Peipei Ma[1,3], Luc Janss[1], Haja N. Kadarmideen[2] and Guosheng Su[1*]

Abstract

Background: Genotyping by sequencing (GBS) is a robust method to genotype markers. Many factors can influence the genotyping quality. One is that heterozygous genotypes could be wrongly genotyped as homozygotes, dependent on the genotyping depths. In this study, a method correcting this type of genotyping error was demonstrated. The efficiency of this correction method and its effect on genomic prediction were assessed using simulated data of livestock populations.

Results: Chip array (Chip) and four depths of GBS data was simulated. After quality control (call rate ≥ 0.8 and MAF ≥ 0.01), the remaining number of Chip and GBS SNPs were both approximately 7,000, averaged over 10 replicates. GBS genotypes were corrected with the proposed method. The reliability of genomic prediction was calculated using GBS, corrected GBS (GBSc), true genotypes for the GBS loci (GBSr) and Chip data. The results showed that GBSc had higher rates of correct genotype calls and higher correlations with true genotypes than GBS. For genomic prediction, using Chip data resulted in the highest reliability. As the depth increased to 10, the prediction reliabilities using GBS and GBSc data approached those using true GBS data. The reliabilities of genomic prediction using GBSc data were 0.604, 0.672, 0.684 and 0.704 after genomic correction, with the improved values of 0.013, 0.009, 0.006 and 0.001 at depth = 2, 4, 5 and 10, respectively.

Conclusions: The current study showed that a correction method for GBS data increased the genotype accuracies and, consequently, improved genomic predictions. These results suggest that a correction of GBS genotype is necessary, especially for the GBS data with low depths.

Keywords: Genomic prediction, Genotype correction, Genotyping by sequencing, Simulation

Background

Genotyping by sequencing (GBS) can produce multiplex libraries of samples based on restriction enzyme and DNA barcoded adapters, and potentially reduce the cost of genotyping [1]. With the reduced-representation sequencing of multiplexed samples, GBS has been developed as a robust method to discover and genotype genome-wide molecular markers [2]. For some species, a commercial chip array is not available, thus GBS will be a good approach to obtain genotypes of DNA markers [3]. However, genotyping quality of GBS tends to be lower than for a chip array [4]. Since genome-wide sequence read depth varies along each sequenced genome of different individuals, genotype quality also varies accordingly [5]. Therefore, the proportion of correctly called genotypes will decrease after decreasing read depths.

Several studies have suggested that it is more powerful to sequence more individuals at the lower coverage [6]. Low-coverage sequencing could capture as much of the variation across the genome as SNP arrays and yielded a commensurate increase in statistical power, which would be a more attractive strategy for the studies of complex trait genetics [7, 8]. In Gorjanc's report [4], expanding the training set resulted in higher overall accuracy of estimated breeding value (EBV), even with reducing the quality of genotyping for lower expense, but genotyping quality may be more important for the prediction set. It

* Correspondence: guosheng.su@mbg.au.dk
[1] Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark
Full list of author information is available at the end of the article

was shown that prediction accuracy increased greatly when read depths also increased in the prediction set [4].

Due to the lower coverage, heterozygous genotypes wrongly genotyped as homozygotes are considered to be a serious problem in GBS data. For example, a read depth of one would genotype only one allele of a diploid at random, so that a true *Aa* genotype is definitely genotyped into *aa* or *AA* genotype by mistake. Previous studies have proposed the maximum-likelihood (ML) method for calling genotypes in low-coverage sequencing data [9, 10], and also developed related programs, such as ANGSD [11] and polyRAD [12]. The R package polyRAD estimated a posterior probability from the priors and likelihoods for each individual and allele using Bayes' theorem. It applied information from high-depth markers to improve genotyping accuracy of low-depth markers using population structure and linkage between loci [12], Additionally, some studies investigated relationship estimation for better relatedness matrices construction using GBS with low depth [13, 14].

In practice, it is possible to correct the wrong genotype calls of GBS data based on read depths and allele frequencies and, consequently, improve the GBS quality to some extent. Therefore, genotype error correction methods are required to complement the future use of GBS data [8]. Simulation is a highly valuable tool to assess such GBS correction methods. Thus, the objective of this study was to propose a method of genotype correction for original GBS data, and then investigate the improvement of genomic prediction (GP) using the simulated data of livestock populations. In this study, four different read depths of GBS genotypes and chip array (Chip) genotypes were simulated. Breeding values were predicted using GBS, corrected GBS (GBSc) and Chip genotypes. The accuracies of genomic predictions were compared to assess the value of GBS and the improvement of GBSc from genotype correction using different genotype data sets.

Methods

In this study, genomic and phenotypic data of ten replicates for each scenario were generated by QMSim software (version 1.10) [15]. Parameters used for generating the population structure and genome are given in Table 1 and Table 2.

Table 1 Simulation parameter of population structure

Step	Population structure	Value
	Number of replicates	10
	Overall heritability	0.3
	QTL heritability	0.3
	Phenotypic variance	1.0
Step1: Historical generation (HG)	Foundation population size of HG	2000
	Number of generation in phase 1	1000
	Population size in the end of phase 1	2000
	Number of generation in phase 2	200
	Population size in the end of phase 2	400
	Number of male in the last HG	200
	Number of female in the last HG	200
	Number of male from HG	40
	Number of female from HG	200
Step 2: Expanded generation (EG)	Number of generation	1
	Litter size	5
	Proportion of male progeny	50%
	Mating design	Random
	Number of male from EG	100
	Number of female from EG	500
Step 3: Recent generation	Number of generation	10
	Litter size	5
	Proportion of male progeny	50%
	Mating design	Random
	Sire replacement	80%
	Dam replacement	40%
	Selection design	EBV

Population structure

During the historical generations, the foundation population of 2,000 individuals (1,000 males and 1,000 females) was kept at a constant size across 1000 generations, and then reduced gradually to 400 individuals (200 males and 200 females) in the following 200 generations, generating linkage disequilibrium (LD) as a result of the domestication process. Among the 400 individuals in the last generation of historical population, 40 males and 200 females were randomly chosen. Each male mated randomly with five females and each female produced five progeny for population expansion. In the recent generations, 100 males and 500 females from the last generation of the expanded population were selected as the parents of the next generation. This continued for ten generations, keeping a male to mate randomly with five females and each female producing five progeny. Selection and replacement was performed based on EBV. The replacement rate was 80% for males and 40% for females. The breeding values were estimated by best linear unbiased prediction (BLUP) using an animal model [16]. In the whole process of simulation, the

Table 2 Simulation parameter of genome

Genome	Value
Number of chromosome	5
Chromosome length	100 Mb
Number of marker loci on one chromosome	1,000,000
Marker position	Evenly
Number of marker alleles in the first HG	2
Marker allele frequencies in the first HG	Random
Number of QTL loci on one chromosome	100
QTL position	Random
Number of QTL allele in the first HG	2
QTL allele frequency in the first HG	Random
QTL allele effect	From gamma distribution with shape 0.4
Maker mutation rate in the historical population	2.5×10^{-5}
QTL mutation rate in the historical population	2.5×10^{-5}

individuals of each sex were produced with equal probability based on the random union of gametes, which were sampled from both the male and female gamete pools. The overall heritability, quantitative trait locus (QTL) heritability and phenotypic variance were set as 0.3, 0.3 and 1.0, respectively. No remaining polygenic effect was simulated, so all the genetic variance was explained by QTLs. The phenotypes were created by adding random residuals to the true breeding values (TBV); TBVs were defined as the sum of individual QTL additive effects. The targeted level of LD in this study was close to the values for cattle [17, 18] and pig [19]. The decay of LD between the markers is shown in Fig. 1, which indicates that the mean r-squared of LD in the last (10th) generation of the population was 0.259 (SE = 0.004) based on markers with interval less than 50 kb (0 ~ 0.05 cM) and minor allele frequencies (MAF) > 0.01, averaged over 10 replicates.

Genome

Initial LD was created by the process of mutation-drift equilibrium in the historical generations. In this process, mutation and drift were considered as the only two evolutionary forces due to no selection, no migration and random mating. Crossovers were simulated to be randomly located across the chromosome and the number of crossovers was sampled from a Poisson distribution. A total of 5×10^6 SNP markers were evenly distributed on five chromosomes of length 100 Mb. Allele frequencies of the bi-allelic markers and QTLs were initiated through randomly sampling from a uniform distribution in the first historical generation. In total, 500 QTLs were simulated and randomly distributed on these five chromosomes. Thus, QTL positions for each replicate were different due to the random distribution. QTL allele effects were sampled from gamma distribution with the shape parameter equal to 0.4. The original QTL effect was that one allele had effect and the other allele had zero effect, and then QTL allelic effects at each locus were scaled as (Allelic effect − QTL mean of population) $\times \frac{\sqrt{\text{Defined QTL variance}}}{\sqrt{\text{Observed QTL variance}}}$, where the observed QTL variance were the sum of QTL variances in the last historical generation. By the scaling, the sum of QTL variances in the last historical generation equals to the defined QTL variance (total additive genetic variance in this study). The effect sizes of two alleles of 500 QTLs in one replicate are shown in Additional file 1. In order to establish mutation-drift equilibrium in historical generations, marker and QTL recurrent mutation rates in historical population were both set as 2.5×10^{-5}. The recurrent mutations assumed that a mutation altered an allele to another instead of creating a new allele and these transition probabilities were equal. The number of mutations for one chromosome of an individual was sampled from a Poisson distribution with the mean u ($u = 2 \times$ number of loci × mutation rate), and then each mutation was assigned to a random locus in the genome. However, recurrent mutations were generally very rare and there was no evidence that these mutations contributed significantly to the erosion of LD between SNPs [20]. In the recent populations, no mutations were generated.

Creating GBS and chip array data

De Donato et al. [21] reported that the distribution of distances between the GBS SNPs differed to that from a chip array data in cattle. For GBS data, the percentage of neighboring SNP less than 50 kb was 44.0% and of more than 150 kb was 13.8%. Following the results of De Donato et al. [21], the distribution of the distances between the neighboring SNPs in this study were set as 13, 8, 8, 12, 9, 6, 5, 16, 7 and 16% for 0.5 kb, 2.5 kb, 7.5 kb, 15 kb, 25 kb, 35 kb, 45 kb, 75 kb, 125 kb and 200 kb, respectively (Fig. 2).

Sequencing errors were not simulated, but low read depths could result in incorrect genotype calls with each allele sampled from Binomial distribution $p \sim B(n, \frac{1}{2})$. Thus, the probability of a heterozygous genotype being correctly called was $1-2\left(\frac{1}{2}\right)^n$ (e.g., 0.00 when n = 1, 0.50 when n = 2, and 0.25 when n = 3). In other words, the probability of a true heterozygous genotype being wrongly called as one of the observed homozygous genotype was $\left(\frac{1}{2}\right)^n$. In the simulation, the number of reads (n) per locus was drawn from a Poisson distribution $n \sim P(x)$, where x was the average

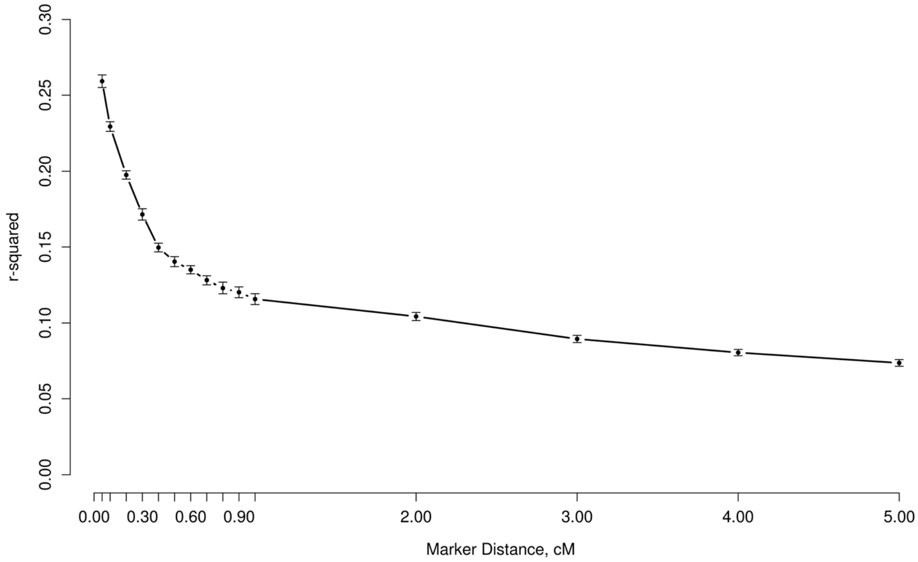

Fig. 1 Decay of LD (r-squared) between markers averaged over 10 replicates. Lines combined with solid circle are average r-squared values in the last (10th) generations of recent population based on MAF > 0.01 and bars indicate SE

depth (x = 2, 4, 5, 10). Actually, the distribution of reads might not reflect the observed distributions accurately due to many factors influencing the variability. Nevertheless, it should not affect the comparisons of genomic prediction results, even though the simulated data has much less variability than observed in practice [22].

We created the GBS genotypes at a heterozygous locus by sampling a random number (r) from a uniform distribution $r{\sim}U(0,1)$. If $r \le \left(\frac{1}{2}\right)^n$, the heterozygous genotype was replaced by aa. If $\left(\frac{1}{2}\right)^n < r < 2\left(\frac{1}{2}\right)^n$, the heterozygous genotype was replaced by AA, else the heterozygous genotype was correctly assigned as Aa. Afterwards, the GBS genotypes aa, Aa and AA were recorded as 0, 1 and 2, respectively.

Quality control criteria of call rate ≥ 0.8 and MAF ≥ 0.01 for SNPs were used to edit the GBS data. After quality control, missing genotypes (zero reads) were set as the mean genotype value for the same loci before genomic prediction. In addition, chip array (Chip) data with no sequencing errors or missing genotypes was also simulated for comparison. The SNP markers of Chip were evenly distributed on five chromosomes with a distance of 50 kb between two adjacent markers.

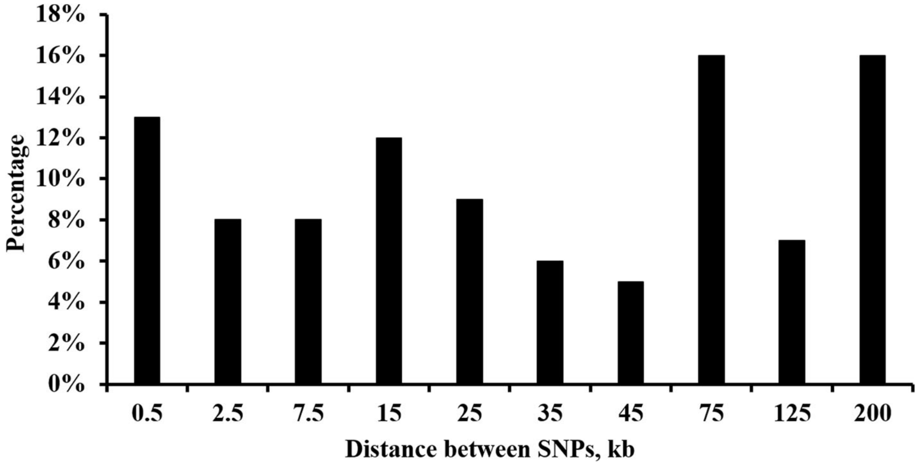

Fig. 2 Distribution of distances between the neighboring SNPs

Genotype correction for GBS data

The genotype correction in this study is to adjust GBS genotypes according to Bayes' conditional probability $P(G|\,GBS)$, where G is the true genotype (unknown) and GBS is the GBS genotype (known) which is subject to genotyping errors. The expected genotype dosage after genotype correction can be rounded to GBSc genotype types (most probable GBSc genotype, i.e., dosage < 0.5: *aa*; dosage > 1.5: *AA*; else: *Aa*). Right and false correction of GBSc were displayed when comparing GBSc genotypes with GBS and true (GBSr) genotypes (Table 3). Homozygous genotypes of GBSc could be corrected into heterozygous genotypes or keep the same homozygous genotypes. If the homozygosity was corrected into heterozygosity and the true genotype of GBSr was also heterozygous, this kind of genotype correction was right genotype correction. However, such genotype correction could be false genotype correction when the true genotype of GBSr was homozygous.

If GBS_{aa} (genotype is labeled in the subscript for the nomenclature) is observed, there are two possible true genotypes (G_{aa} and G_{Aa}), and the probabilities are

$$P(G_{aa}|GBS_{aa}) = \frac{P(G_{aa})\,P(GBS_{aa}|G_{aa})}{P(GBS_{aa})},$$

$$P(G_{Aa}|GBS_{aa}) = 1 – P(G_{aa}|GBS_{aa}).$$

Similarly, If GBS_{AA} is observed,

$$P(G_{AA}|GBS_{AA}) = \frac{P(G_{AA})\,P(GBS_{AA}|G_{AA})}{P(GBS_{AA})},$$

$$P(G_{Aa}|GBS_{AA}) = 1 – P(G_{AA}|GBS_{AA}).$$

If GBS_{Aa} is observed, G_{Aa} is the only possible true genotype,

$$P(G_{Aa}|GBS_{Aa}) = 1.$$

Let assume $p = P(A)$ and $q = P(a)$ for the true genotype, the relevant probabilities can be written as:

$$\begin{aligned} P(GBS_{aa}) &= P(G_{aa}) + P(GBS_{aa}|G_{Aa}) \\ &= q^2 + 2pq\left({}^1\!/_2\right)^n, \end{aligned}$$

$$\begin{aligned} P(G_{aa}|GBS_{aa}) &= P(G_{aa}) * P(GBS_{aa}|G_{aa})/P(GBS_{aa}) \\ &= q^2/P(GBS_{aa}), \end{aligned}$$

Table 3 Genotype changes of corrected GBS (GBSc) in the lower panel of Fig. 3

Genotype change in the lower panel of Fig. 3	GBS	GBSc	GBSr
Right correction of GBSc (GBS ≠ GBSc = GBSr)	aa/AA	Aa	Aa
False correction of GBSc (GBSc ≠ GBS = GBSr)	aa/AA	Aa	aa/AA

$$\begin{aligned} P(G_{Aa}|GBS_{aa}) &= P(G_{Aa}) * P(GBS_{aa}|G_{Aa})/P(GBS_{aa}) \\ &= 2pq\left({}^1\!/_2\right)^n/P(GBS_{aa}), \end{aligned}$$

$$\begin{aligned} P(GBS_{AA}) &= P(G_{AA}) + P(GBS_{AA}|G_{Aa}) \\ &= p^2 + 2pq\left({}^1\!/_2\right)^n, \end{aligned}$$

$$\begin{aligned} P(G_{AA}|GBS_{AA}) &= P(G_{AA}) \\ &\quad * P(GBS_{AA}|G_{AA})/P(GBS_{AA}) \\ &= p^2/P(GBS_{AA}), \end{aligned}$$

$$\begin{aligned} P(G_{Aa}|GBS_{AA}) &= P(G_{Aa}) * P(GBS_{AA}|G_{Aa})/P(GBS_{AA}) \\ &= 2pq\left({}^1\!/_2\right)^n/P(GBS_{AA}). \end{aligned}$$

Let 0, 1, and 2 denote *aa*, *Aa* and *AA* genotype, respectively. The original GBS genotypes are scored as $GBS_{aa} = 0$, $GBS_{Aa} = 1$ and $GBS_{AA} = 2$. The correction of GBS used in this study is to correct GBS homozygous genotypes to be expected genotype dosage. Thus,

$$GBSc_{aa} = 2pq\left({}^1\!/_2\right)^n/\left(q^2 + 2pq\left({}^1\!/_2\right)^n\right),$$

$$GBSc_{AA} = \left(2p^2 + 2pq\left({}^1\!/_2\right)^n\right)/\left(p^2 + 2pq\left({}^1\!/_2\right)^n\right),$$

$$GBSc_{Aa} = 1.$$

Allele frequency can be calculated from the data including all reads when assuming Hardy-Weinberg equilibrium. It can be estimated from the known GBS data in this way:

$$\begin{aligned} P(GBS_{AA}) – P(GBS_{aa}) &= \left(p^2 + 2pq\left({}^1\!/_2\right)^n\right) \\ &\quad – \left(q^2 + 2pq\left({}^1\!/_2\right)^n\right) = p^2 – q^2 = 2p – 1, \end{aligned}$$

$$p = \frac{P(GBS_{AA}) – P(GBS_{aa}) + 1}{2}$$

Statistical analysis

Based on the GBS, GBSc, GBSr and Chip data, genomic breeding values (GEBV) were predicted by a SNP-BLUP model using the BayZ package (http://www.bayz.biz/). The model was

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Mg} + \mathbf{e},$$

where $\mathbf{y}$ was the vector of phenotypic values, $\mathbf{1}$ is the vector of ones, μ is the overall mean, $\mathbf{g}$ is the vector of random unknown marker effects to be estimated, $\mathbf{M}$ is the coefficient matrix of genotypes which links $\mathbf{g}$ to $\mathbf{y}$, and $\mathbf{e}$ is the vector of random residuals. It was assumed that $\mathbf{g} \sim N\left(0, \mathbf{I}\sigma_g^2\right)$.

Validation

In the 6th to 9th generations of the recent population, 7,500 individuals were used as a training set, in which all individuals were genotyped and phenotyped. The test set comprised of 2,500 genotyped individuals from the 10th generation. The reliabilities of genomic predictions using marker data from GBS, GBSc, GBSr and Chip were compared. The reliabilities of genomic predictions were calculated as squared correlations between the predicted and true breeding values for individuals in the test data set.

Results

Distributions of read depth (*n*) at four levels of mean read depth (*x*)

Additional file 2 showed the realized frequency distributions of reads at four mean depths (x = 2, 4, 5, 10), which were highly consistent with the theoretical frequencies of Poisson distribution. The percentages of read depth ≤ 5 at mean read depth = 2, 4, 5 and 10 were 98.3%, 78.5%, 61.6% and 6.71%, respectively. The percentages of missing genotypes at mean read depth = 2, 4, 5 and 10 were 13.5%, 1.83%, 0.673% and 0.00464%, respectively. Standard deviation (SD) of ten replicates were all less than 4.74×10^{-5}.

Incorrect genotype calls

It was expected that a heterozygous genotype may be wrongly assigned to a homozygous genotype with probability of $2(\frac{1}{2})^n$. The upper panel of Additional file 3 shows the proportions of incorrect genotype calls over true genotypes and the SDs of ten replicates were all less than 6.32×10^{-3}. Although the average number of realized reads for different loci was nearly the same, there was much variation in the proportion of incorrect genotypes observed for different loci (lower panel of Additional file 3). Having a large proportion of incorrect genotypes in the loci close to QTL regions could affect genomic prediction.

Improvement of accuracies of GBS genotype by genotype correction

The accuracies of GBS genotypes were measured as correlations between the true genotype and the GBS genotype, as well correct rates of GBS genotype calls. As shown in the upper panel of Fig. 3, genotype correction improved the accuracies of GBS genotype. The correlations between reported genotype and true genotype were highest for the GBSc genotype dosage and lowest for the original GBS, while GBSc genotype type (most probable genotype) was in between. The differences among these three genotype data were larger for lower depth, and not for depth = 10. Similarly, correct rates of GBSc genotype type were higher than those of original GBS for depth = 2, 4, and 5, but not for depth = 10. GBSc genotype type occupied a larger proportion of right genotype correction than false genotype correction (lower panel of Fig. 3). As expected, larger improvement was observed in the genotype data with lower depth.

Reliabilities of genomic prediction

After quality control (call rate ≥ 0.8 for loci and MAF ≥ 0.01), the number of GBS SNPs obtained for the four depths was approximately 7,000 averaged over 10 replicates. As shown in the Additional file 2, the missing genotypes at mean depth = 2, 4, 5 and 10 were less than 13.5%, so call rate criteria of 80% for loci had nearly no effect on genotypes editing. However, the missing genotypes at mean depth = 1 were high, up to approximately 30%; therefore, a large number of loci did not meet this criterion, and this depth was discarded. In addition, approximately 7,000 SNPs remained for the Chip data using the same quality control criteria (call rate ≥ 0.8 for loci and MAF ≥ 0.01).

The reliabilities (r^2) of genomic prediction averaged over 10 replicates at four depths are shown in Fig. 4. Prediction reliability using Chip data (0.710) was slightly higher than that using true genotype for the GBS loci (GBSr) (0.706). As depth increased from 2 to 10, the prediction reliabilities using GBS and GBSc data gradually approached the reliabilities using GBSr data. The worst prediction reliability was from depth = 2 due to having the most missing and incorrect genotypes (Additional file 2 and Additional file 3). Genotype correction improved genomic prediction to different degrees, consistent with the accuracies of corrected genotypes (Fig. 3). Thus, the reliabilities of genomic prediction using GBSc data were higher than those using GBS data at all four depths (Fig. 4). The standard error (SE) of prediction reliabilities in 10 replicates was approximately 0.025 for scenarios of different depths and types of the marker data.

Without quality control (call rate ≥ 0.8 for loci and MAF ≥ 0.01) for editing genotype data, approximately 8,000 GBS SNPs were obtained at the four depths, averaged over 10 replicates. Table 4 reveals that reliabilities of genomic prediction using GBS and GBSc data before editing genotypes were better than those using the data after editing genotypes. In addition, GBSc led to higher reliability than GBS no matter that correction was performed before or after genotype editing.

Discussion

Potential application of GBS in breeding programs

As a highly multiplexed technology for constructing reduced representation libraries, GBS generates large numbers of SNPs for use in genetic analyses [2]. Unlike

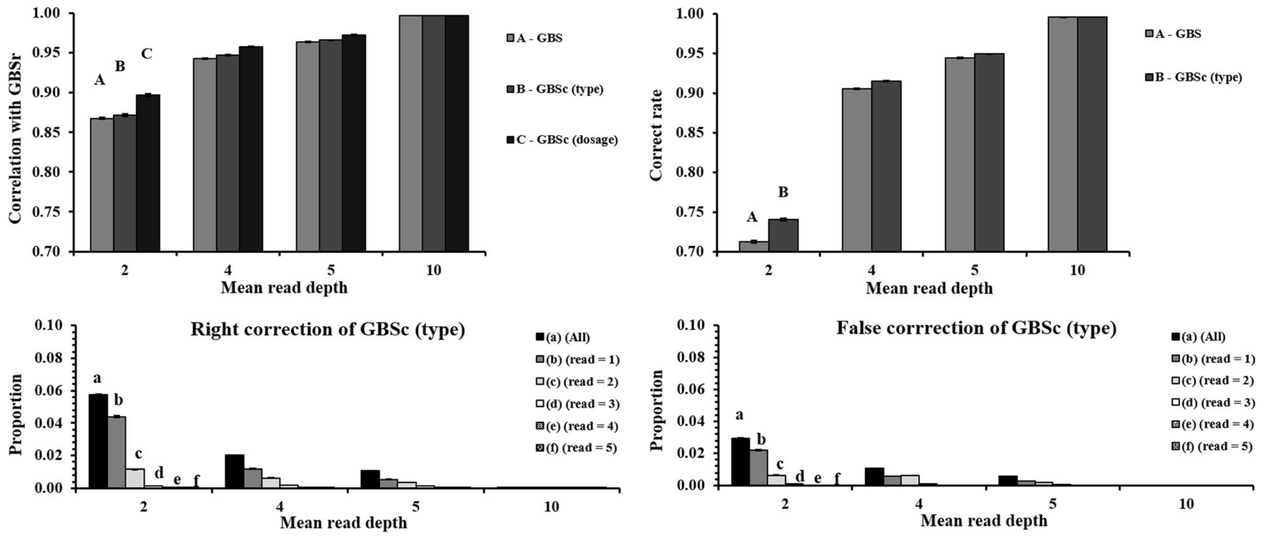

Fig. 3 Correlations and correct rates for original GBS (GBS), corrected GBS genotype type (GBSc type), corrected GBS genotype dosage (GBSc dosage) with true genotype in GBS loci (GBSr) data (the upper panel), as well right and false genotype correction of GBSc (type) data (the lower panel) at four mean depths, averaged over 10 replicates

genotyping approach of chip arrays, GBS allows de novo marker discovery, even when there is no reference genome [3]. The GBS technology includes digestion by a restriction enzyme, ligation of barcode adapter, amplification by PCR and sequencing of amplified DNA [3]. Repetitive regions of genome can be avoided and lower copy regions can be targeted efficiently to simplify the alignment problems by using appropriate restriction enzymes [2, 23]. Single-well digestion and barcode adapter ligation results in reduced sample handling, less PCR amplification and no size fractionation limitation [24]. Currently, GBS has been widely used in many breeding programs. Poland et al. [25] presented a GBS approach for barley and wheat lacking of a reference genome sequence and found GBS to have broad applications in plant breeding programs. The applications of genomic selection in aquaculture species has been underpinned by GBS techniques, which are available for a handful of aquaculture species [26]. Additionally, GBS has great potential application in domestic species whose reference sequences are either being developed or have been fully sequenced, as it enables acceptable marker density for genomic selection in cattle at one third of the cost of the current genotyping technologies [21].

In this study, the highest reliability of genomic prediction was from using the chip array (Chip) data. Obviously, Chip SNPs were evenly distributed along the genome, so at least one SNP was in strong LD with QTLs. However, large distances between some

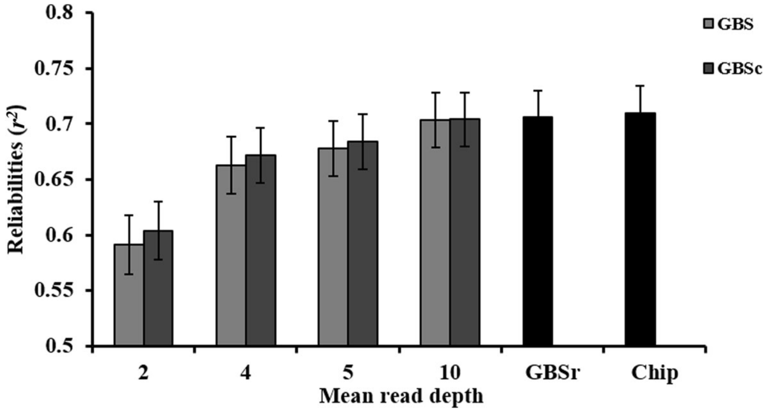

Fig. 4 Reliabilities (r^2) of genomic prediction using original GBS (GBS), corrected GBS (GBSc dosage), true genotype in GBS loci (GBSr) and chip array (Chip) data, at four depths, averaged over 10 replicates. Bars indicate SE

Table 4 Reliabilities of genomic prediction using original GBS (GBS) and corrected GBS (GBSc) data before and after editing genotypes, at four mean depths, averaged over 10 replicates

Reliability (SE)	GBS (after editing genotypes)	GBS (no editing genotypes)	GBSc (after editing genotypes)	GBSc (no editing genotypes)
Depth = 2	0.591 (0.026)	0.598 (0.023)	0.603 (0.026)	0.610 (0.024)
Depth = 4	0.662 (0.025)	0.663 (0.024)	0.671 (0.025)	0.672 (0.024)
Depth = 5	0.678 (0.025)	0.683 (0.023)	0.684 (0.025)	0.687 (0.023)
Depth = 10	0.703 (0.024)	0.704 (0.024)	0.704 (0.024)	0.704 (0.024)

neighboring GBS SNPs weakened the LD between SNP and QTL. De Donato et al. [21] reported that the BovineSNP50 chip array had a large proportion of intervals from 20 kb to 100 kb and only 3% had an interval larger than 150 kb, while GBS data showed that about 14% of SNP intervals were more than 150 kb. Even if LD exists at long distances (longer than 1 cM in some regions and some species), LD will decay as the distance between marker and QTL increases [27]. This study restricted the numbers of Chip SNPs and GBS SNPs to be same. If the number of GBS SNPs increases, GBS data could perform as well as Chip data in genomic prediction. In general, GBS can produce enough information even at relatively low coverage, but wrong calls of genotype are the main disadvantage that requires further improvement.

Genotype dosage and most probable genotype

This study demonstrated the method of correction for GBS genotype to improve the accuracy of GBS data, following the previous studies [9–14]. This correction method resulted in GBSc genotype type and GBSc genotype dosage. The GBSc genotype type was derived from rounding the GBSc genotype dosage. The inferred GBSc genotype dosage of genotype *aa* was $\text{dosage}(GBSc_{aa}) = 1 - \frac{q^2}{q^2 + 2pq\left(\frac{1}{2}\right)^n} = \frac{1}{\frac{q}{2p\left(\frac{1}{2}\right)^n} + 1}$, which ranges between 0 and 1. For example, dosage($GBSc_{aa}$) = 1 – q if read depth = 1, which is always larger than 0.5 when q is less than 0.5. When reads depths become higher, dosage($GBSc_{aa}$) was closer to zero (Fig. 5) and the inferred GBSc genotype dosage and inferred GBSc genotype type were more consistent. The results also indicate that the rounded genotypes derived from dosage were more accurate than original GBS genotype, but some of them might be far from the true genotype, dependent on the number of reads and allele frequency.

Correction of GBS genotype improved genomic prediction

The proportions of right genotype correction were more than the proportions of false genotype correction in this study, so the correct rates of GBSc increased 0.028, 0.010, 0.005 and 0.001 from original GBS at depth = 2, 4, 5 and 10, respectively (Fig. 3). Therefore, genotype correction increased the accuracy of GBS genotypes. GBSc genotype dosage had higher correlation than GBSc genotype type, so GBSc genotype dosage was more accurate (Fig. 3). In fact, all meaningful information about uncertainty is lost by choosing the largest probability [28], such as the GBSc genotype type rounded from GBSc genotype dosage in this study. For genotype imputation, a general approach is also the use of posterior probabilities. Imputed genotypes are predictions instead of actual observations of genotyping, so incorporating

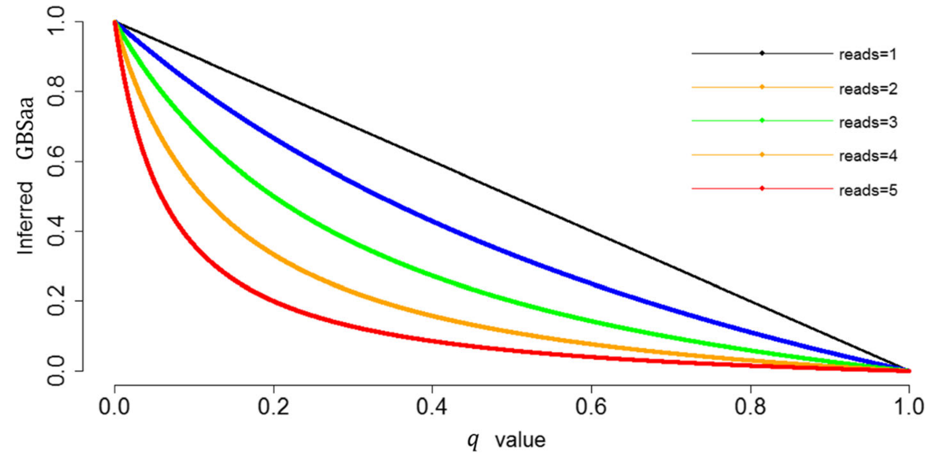

Fig. 5 Inferred genotype dosage of GBS_{aa} versus q value for different read depths (n)

the uncertainty of these predictions could avoid spurious results in some cases [29].

We only used a SNP-BLUP model for predicting breeding value in this study. The result indicated that genomic prediction using corrected GBS data were more accurate than using the original GBS data (Fig. 4). It is expected that the gain in reliability of genomic prediction from genotype correction will also present when using other genomic prediction models such as Bayesian variable selection models, because genotype correction increases the accuracies of genotype assignment (Fig. 3).

In this study, retaining more SNPs resulted in higher prediction reliabilities (Table 4), which meant that a decreasing or less stringent thresholds of call rate and MAF led to an increase in prediction reliability [30]. Our study also revealed that genotype correction improved genomic prediction more than removing the MAF threshold. Cooke et al. [31] expected to reduce genotype errors in GBS data by estimating allelic dropout. Their simulation studies using their GBStools package improved genotyping accuracy more than hard filters [31]. Meanwhile, Furuta et al. [32] used their post SNP-calling error correction to eliminate most errors of raw GBS data, but some remained. Their studies also found that simple imputation methods can reinforce the usefulness of GBS data tremendously, even if up to 75% of missing data for each marker existed in the raw GBS data [32]. However, methods rely on population types and reference genomes, so they may not always be applied.

Conclusions

The current study demonstrated a method for the correction of GBS genotypes. The results showed that the correction increased the accuracy of GBS genotype and increased the accuracy of genomic prediction. Therefore, a correction method for GBS genotype is necessary, especially for GBS data with low depth.

Additional files

Additional file 1: Effect sizes of 500 QTLs in one replicate. (TIF 257 kb)

Additional file 2: Read depths at four mean read depths in one replicate. (TIF 171 kb)

Additional file 3: Proportions of wrongly called genotypes at four depths averaged over the whole genome and over 10 replicates (the upper panel), as well proportions of wrongly called genotypes (the lower panel) along 50 loci in one replicate. (TIF 187 kb)

Abbreviation

BLUP: Best linear unbiased prediction; Chip: Chip array; cM: Centimorgan; EBV: Estimated breeding value; EG: Expanded generation; GBS: Genotyping by sequencing; GBSc: Corrected GBS; GBSr: True genotype in the GBS loci; GEBV: Genomic estimated breeding value; GP: Genomic prediction; HG: Historical generation; LD: Linkage disequilibrium; MAF: Minor allele frequency; Mb: Megabase pair; ML: Maximum likelihood; QTL: Quantitative trait loci; SD: Standard deviation; SE: Standard error; SNP: Single nucleotide polymorphism; TBV: True breeding value

Acknowledgements

Not applicable.

Availability of data and material

The datasets used in the current study are available from the corresponding author. Guosheng Su: guosheng.su@mbg.au.dk

Funding

This work was supported by the Genomic Selection in Plants and Animals (GenSAP) research project financed by the Danish Council of Strategic Research (Aarhus, Denmark); The first author acknowledges the scholarship provided by the China Scholarship Council (CSC).

Authors' contributions

XW, MSL and GS conceived and designed the experiments. GS and XW demonstrated the genotype correction method. XW analyzed the data. PM and LJ helped the analysis. XW and GS wrote the manuscript. MSL and HK improved the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark. [2]Department of Bio and Health Informatics and Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark. [3]School of Agriculture and Biology, Shanghai Jiaotong University, Shanghai, China.

Received: 15 August 2018 Accepted: 4 January 2019
Published online: 24 January 2019

References

1. Poland JA, Rife TW. Genotyping-by-sequencing for plant breeding and genetics. Plant Genome J. 2012;5:92–102.
2. 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:e19379.
3. He J, Zhao X, Laroche A, Lu Z, Liu H, Li Z. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front Plant Sci. 2014;5:484.
4. Gorjanc G, Cleveland MA, Houston RD, Hickey JM. Potential of genotyping-by-sequencing for genomic selection in livestock populations. Genet Sel Evol. 2015;47:12.
5. Beissinger TM, Hirsch CN, Sekhon RS, Foerster JM, Johnson JM, Muttoni G, et al. Marker density and read depth for genotyping populations using genotyping-by-sequencing. Genetics. 2013;193:1073–81.
6. Alex Buerkle C, Gompert Z. Population genomics based on low coverage sequencing: how low should we go? Mol Ecol. 2013;22:3028–35.
7. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. Low-coverage sequencing: implications for design of complex trait association studies. Genome Res. 2011;21:940–51.
8. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet. 2012;44:631–5.
9. Maruki T, Lynch M. Genotype calling from population-genomic sequencing data. G3-genes Genom. Genet. 2017;7:1393–404.
10. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–93.
11. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. BMC Bioinformatics. 2014;15:356.

12. Clark LV., Lipka AE, Sacks EJ polyRAD: Genotype calling with uncertainty from sequencing data in polyploids and diploids bioRxiv 2018; https://doi.org/10.1101/380899.
13. Dodds KG, McEwan JC, Brauning R, Anderson RM, van Stijn TC, Kristjánsson T, et al. Construction of relatedness matrices using genotyping-by-sequencing data. BMC Genomics. 2015;16:1047.
14. Cericola F, Lenk I, Fè D, Byrne S, Jensen CS, Pedersen MG, et al. Optimized use of low-depth genotyping-by-sequencing for genomic prediction among multi-parental family pools and single plants in perennial ryegrass (Lolium perenne L.). Front Plant Sci. 2018;9:369.
15. Sargolzaei M, Schenkel FS. QMSim: a large-scale genome simulator for livestock. Bioinformatics. 2009;25:680–1.
16. Henderson CR. Best linear unbiased estimation and prediction under a selection model. Biometrics. 1975;31:423–47.
17. Makina SO, Taylor JF, van Marle-Köster E, Muchadeyi FC, Makgahlela ML, MacNeil MD, et al. Extent of linkage disequilibrium and effective population size in four south African Sanga cattle breeds. Front Genet. 2015;6:337.
18. McKay SD, Schnabel RD, Murdoch BM, Matukumalli LK, Aerts J, Coppieters W, et al. Whole genome linkage disequilibrium maps in cattle. BMC Genet. 2007;8:74.
19. Du FX, Clutter AC, Lohuis MM. Characterizing linkage disequilibrium in pig populations. Int J Biol Sci. 2007;3:166–78.
20. Ardlie KG, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. 2002;3:299–309.
21. De Donato M, Peters SO, Mitchell SE, Hussain T, Imumorin IG. Genotyping-by-sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. PLoS One. 2013;8:e62137.
22. Hess AS, Hess MK, Dodds KG, Mcewan JC, Clarke SM, Rowe SJ. A method to simulate low-depth genotyping-by-sequencing data for testing genomic analyses. Proc 11th World Congr Genet Appl to Livest Prod. 2018:385.
23. Gore M, Bradbury P, Hogers R, Kirst M, Verstege E, Van Oeveren J, et al. Evaluation of target preparation methods for single-feature polymorphism detection in large complex plant genomes. Crop Sci. 2007;47:135–48.
24. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. 2011;12:499–510.
25. Poland JA, Brown PJ, Sorrells ME, Jannink JL. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One. 2012;7:e32253.
26. Robledo D, Palaiokostas C, Bargelloni L, Martínez P, Houston R. Applications of genotyping by sequencing in aquaculture breeding and genetics. Rev Aquac. 2018;10:670–82.
27. Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet. 2009;10:381–91.
28. Palmer C, Pe'er I. Bias characterization in probabilistic genotype data and improved signal detection with multiple imputation. PLoS Genet. 2016;12:e1006091.
29. Ellinghaus D, Schreiber S, Franke A, Nothnagel M. Current software for genotype imputation. Hum Genomics. 2009;3:371–80.
30. Edriss V, Guldbrandtsen B, Lund MS, Su G. Effect of marker-data editing on the accuracy of genomic prediction. J Anim Breed Genet. 2013;130:128–35.
31. Cooke TF, Yee MC, Muzzio M, Sockell A, Bell R, Cornejo OE, et al. GBStools: a statistical method for estimating allelic dropout in reduced representation sequencing data. PLoS Genet. 2016;12:e1005631.
32. Furuta T, Ashikari M, Jena KK, Doi K, Reuscher S. Adapting genotyping-by-sequencing for Rice F2 populations. G3-genes Genom. Genet. 2017;7:881–93.