Potential of low-coverage genotyping-by-sequencing and imputation for cost-effective genomic selection in bi-parental segregating populations

Published March 27, 2017

RESEARCH

Potential of Low-Coverage Genotyping-by-Sequencing and Imputation for Cost-Effective Genomic Selection in Biparental Segregating Populations

Gregor Gorjanc,* Jean-Francois Dumasy, Serap Gonen, R. Chris Gaynor, Roberto Antolin, and John M. Hickey

ABSTRACT

Genotyping-by-sequencing (GBS) is an alternative genotyping method to single-nucleotide polymorphism (SNP) arrays that has received considerable attention in the plant breeding community. In this study we use simulation to quantify the potential of low-coverage GBS and imputation for cost-effective genomic selection in biparental segregating populations. The simulations comprised a range of scenarios where SNP array or GBS data were used to train the genomic selection model, to predict breeding values, or both. The GBS data were generated with sequencing coverages (*x*) from 4*x* to 0.01*x*. The data were used either nonimputed or imputed by the AlphaImpute program. The size of the training and prediction sets was either held fixed or was increased by reducing sequencing coverage per individual. The results show that nonimputed 1*x* GBS data provided comparable prediction accuracy and bias, and for the used measurement of return on investment, outperformed the SNP array data. Imputation allowed for further reduction in sequencing coverage, to as low as 0.1*x* with 10,000 markers or 0.01*x* with 100,000 markers. The results suggest that using such data in biparental families gave up to 5.63 times higher return on investment than using the SNP array data. Reduction of sequencing coverage per individual and imputation can be leveraged to genotype larger training sets to increase prediction accuracy and larger prediction sets to increase selection intensity, which both allow for higher response to selection and higher return on investment.

G. Gorjanc, J.-F. Dumasy, S. Gonen, R.C. Gaynor, R. Antolin, and J.M. Hickey, The Roslin Institute and Royal (Dick) School of Veterinary Studies, Univ. of Edinburgh, Easter Bush Research Centre, Midlothian EH25 9RG, UK. Received 15 Aug. 2016. Accepted 30 Jan. 2017. *Corresponding author (Gregor.Gorjanc@roslin.ed.ac.uk). Assigned to Associate Editor Michael Allen.

Abbreviations: GBS, genotyping-by-sequencing; SNP, single-nucleotide polymorphism.

This study quantifies the potential of low-coverage genotyping-by-sequencing (GBS) and imputation for cost-effective genomic selection in biparental segregating populations. Genomic selection can increase the rates of genetic gain by shortening the generation interval and selecting from amongst a greater number of diverse individuals. Unfortunately, the cost of genomic selection in early segregating populations is high because the training set for the genomic selection model must be large and because genomic predictions have to be obtained for a large number of individuals (Riedelsheimer and Melchinger, 2013; Hickey et al., 2014). To enable the use of genomic selection in segregating populations, low-cost genotyping strategies need to be developed.

One such low-cost strategy is to use single-nucleotide polymorphism (SNP) genotyping arrays with high and low marker density combined with imputation. This strategy involves a combination of genotyping parents with high-density SNP arrays, genotyping progeny with low-density SNP arrays, and imputation of high-density information from the parents onto progeny (Hickey et al., 2015; Jacobson et al., 2015; Gorjanc et al., 2016). Although the imputation is imperfect, it provides enough information for effective genomic selection in a cost-effective way (Jacobson et al., 2015; Gorjanc et al., 2016).

Published in Crop Sci. 57:1–17 (2017).
doi: 10.2135/cropsci2016.08.0675

Another low-cost genotyping strategy is GBS (Altshuler et al., 2000; Davey et al., 2011). This genotyping technology enables collection of large volumes of high-density genomic data, and its flexibility has large potential for generating new cost-effective genotyping strategies. There are many variants of GBS (see a white paper provided by Illumina, 2014), including restriction-site-associated DNA sequencing (Baird et al., 2008) and the GBS method developed by Elshire et al. (2011). The latter method (Elshire et al., 2011) has received considerable interest in the plant breeding community (Poland et al., 2012b; Poland and Rife, 2012; Beissinger et al., 2013; Crossa et al., 2013; Zhang et al., 2015). Our interest in this genotyping technique, however, is in its flexibility and ability to drive the per-sample cost below US$10 (Poland and Rife, 2012). In this study, we do not refer to any specific GBS method but rather work with a generic sequencing-based process of generating genomic data, which is common to all of these methods.

Genotyping-by-sequencing allows breeders to manipulate the amount of retrieved information and its cost in three ways. First, a breeder can adjust the number of sequenced loci by choosing different restriction enzymes or through adjustments to other parts of the protocol (Elshire et al., 2011; Poland and Rife, 2012; Poland et al., 2012a). Genotyping-by-sequencing can thus generate a small or large number of markers (e.g., hundreds or millions) at different cost. Second, a breeder can influence the amount of retrieved information by manipulating sequencing depth. Increasing genomewide sequencing depth (hereinafter referred to as sequencing coverage or x) increases the average number of times a locus is sequenced, and this alters the informativeness of the resulting genomic data. Specifically, it increases the probability of correctly calling individual genotypes from such data. However, increasing sequencing coverage also increases the costs. Third, a breeder can reduce the cost by optimizing the sample preparation and sequencing process. In particular, multiplex sequencing, which involves tagging DNA fragments from multiple individuals and sequencing them jointly in one sequencing process (Craig et al., 2008), enables substantial cost reductions (Poland and Rife, 2012).

High heterozygosity in segregating populations raises a potential challenge for genomic selection with low-coverage GBS data. The success of genomic selection depends on the ability of genotypic data to capture genetic variation among the training and prediction individuals at low cost. Most studies of genomic selection with GBS data have focused on settings with inbred individuals (Poland et al., 2012b; Crossa et al., 2013; Rutkoski et al., 2013) and have shown that the accuracy of genomic prediction using low-coverage GBS data was comparable with using SNP array or diversity arrays technology (DArT) data. These results may not hold for segregating populations, because capturing genetic variation with low-coverage sequencing in such a setting is challenging. For example, sequencing a heterozygous locus once ($1x$) reveals only one allele, and the genotype from such data would be wrongly called as a reference or alternative homozygote. It is unknown if such low-coverage GBS data are useful for genomic selection in segregating populations. A simulation study in an outbred livestock population shows that low-coverage GBS data enable accurate and unbiased genomic predictions when a sufficient number of markers is available and coverage per individual is at least $1x$ (Gorjanc et al., 2015), which holds promise for segregating plant populations.

Imputation could increase the informativeness of low-coverage GBS data. As imputation is used with low-density SNP array data, it could also be used with low-coverage GBS data, provided that the imputation method takes into account the probabilistic nature of sequencing data (Li et al., 2010; Pasaniuc et al., 2012; Huang et al., 2014). The study in outbred livestock populations shows that sequencing coverage per individual should be at least $1x$ for accurate genomic predictions. This observation suggests that a low-cost strategy could be to obtain GBS data with less than $1x$ and use imputation to increase coverage up to or above $1x$. This strategy could be used to decrease the cost of sequencing the large training and prediction sets required for accurate genomic selection. It could also be used to increase the size of sets that could be generated at a predefined cost.

In this study, we used simulation to quantify the potential of low-coverage GBS and imputation for cost-effective genomic selection in biparental segregating populations.

MATERIALS AND METHODS

The simulations comprised a range of scenarios where non-imputed or imputed GBS data were used to train the genomic selection model, to predict breeding values, or both. The size of the training or prediction sets was either held fixed or was increased by reducing sequencing coverage per individual. The results were compared with those obtained from the true (SNP array) genotypic data. The simulation involved the following steps, most of which were performed with the AlphaSim program (Faux et al., 2016) available at http://www.AlphaGenes.Roslin.ed.ac.uk/AlphaSuite/AlphaSim:

1. Generate founder genomes.
2. Select causal loci and markers.
3. Generate a breeding program and breeding and phenotypic values.
4. Generate marker allele dosages using SNP array and GBS technology.
5. Impute GBS data.
6. Measure accuracy of GBS data.
7. Estimate marker associations and breeding values.
8. Measure the accuracy and bias of genomic predictions, response to selection, and return on investment.

The results were summarized over 10 replicates and presented graphically, whereas Supplemental Table S1 provides results in a tabular form. Data manipulation and summaries were performed with the R program (R Development Core Team, 2016).

Founder Genomes

Base population genome sequences were simulated for 10 chromosomes using the Markovian Coalescent Simulator (Chen et al., 2009). The chromosomes comprised 1.0×10^8 base pairs and were simulated with a per-site mutation rate of 1.0×10^{-8}, a per-site recombination rate of 1.0×10^{-8} (100 cM in total per chromosome), and an effective population size that varied over time to mimic the historical changes. The effective population size was set to 50 in the final generation of the coalescent simulation, to 100 at 10 generations ago, to 1000 at 100 generations ago, to 6000 at 1000 generations ago, to 12,000 at 10,000 generations ago, and to 32,000 at 100,000 yr ago with linear changes between. The simulated chromosomes collectively had ~1,000,000 segregating variants (biallelic SNPs).

Causal Loci and Markers

Among the segregating variants, 10,000 were chosen at random as causal loci of a quantitative trait with the restriction of an equal number from each chromosome. The effect of each causal locus was sampled from a normal distribution with a mean of zero and variance of one divided by the number of causal loci. Additionally, a set of 10,000 (10K) and 100,000 (100K) variants were chosen at random as codominant markers.

Breeding Program and Breeding and Phenotypic Values

A breeding program of a self-pollinating species with inbred parents and biparental populations (families) was simulated (Fig. 1). The program was initiated by establishing a base population of 40 inbred parents. Each parent had one haplotype per chromosome sampled from the base haplotypes, allowing for recombination between base haplotypes. The sampled haplotypes were doubled to create inbreds. The parents were then crossed at random to create 160 biparental populations with a restriction that any pair of parents could only be crossed once. Each biparental population was constructed by (i) mating at random two inbred parents to generate F_1 individuals, (ii) selfing F_1 individuals to generate 400 F_2 individuals, and (iii) selfing F_2 individuals to generate 400 F_3 individuals. True breeding value for the F_3 individuals was calculated as a sum of the effect of causal alleles an individual inherited. The phenotypic values that pertained to the F_3 individuals (collected in the $F_{3:4}$ stage) were simulated by adding a random residual to the true breeding value. Heritability was set to 0.1 by scaling the residual variance relative to the variance of the true breeding values in inbred lines.

Marker Allele Dosages

Marker data were generated with either SNP array or GBS technology. Each marker genotype was represented numerically as the number of alternative alleles (i.e., allele dosage; 0 for reference homozygote 0/0, 1 for heterozygote 0/1, and 2 for alternative homozygote 1/1). We assumed that both technologies covered the same set of loci and that neither technology introduced errors. With SNP arrays, allele dosages were set directly to 0, 1, or 2 depending on the marker genotype. With GBS, allele dosages were obtained by simulating the GBS process with a targeted sequencing coverage, x (Li et al., 2010; Pasaniuc et al., 2012). The GBS process had four steps:

1. Sequenceability of each marker locus (s_j) was sampled from a Gamma distribution with shape and rate parameter equal to 4, $s_j \sim \text{Gamma}(a = 4, b = 4)$.
2. The number of sequence reads for an individual i at a locus j ($n_{i,j}$) was sampled from a Poisson distribution with mean equal to xs_j, $n_{i,j} \sim \text{Poisson}\ (l = xs_j)$.
3. The sequence reads were distributed at random between the two alleles of an individual, $n_{i,j,1} \sim \text{Binomial}(p = 0.5, k = n_{i,j})$ and $n_{i,j,2} = n_{i,j} - n_{i,j,1}$.
4. Allele dosages ($m_{i,j}$) were calculated as a weighted average of the allele reads $m_{i,j} = (n_{i,j,1}a_{i,j,1} + n_{i,j,2}a_{i,j,2})/n_{i,j}$, where $a_{i,j,k}$ is the kth allele code (0 for the reference allele and 1 for the alternative allele).

The GBS process gave continuous allele dosages with values ranging from 0 to 2. These values depend on the true genotype at a locus of an individual, number of sequence reads, and randomness of the process. If no sequence reads occurred for an individual at a particular locus and imputation was not used, the allele dosage was set equal to two times the allele frequency computed from the observed GBS allele dosages within a biparental family. When imputation was used, the fourth step was skipped.

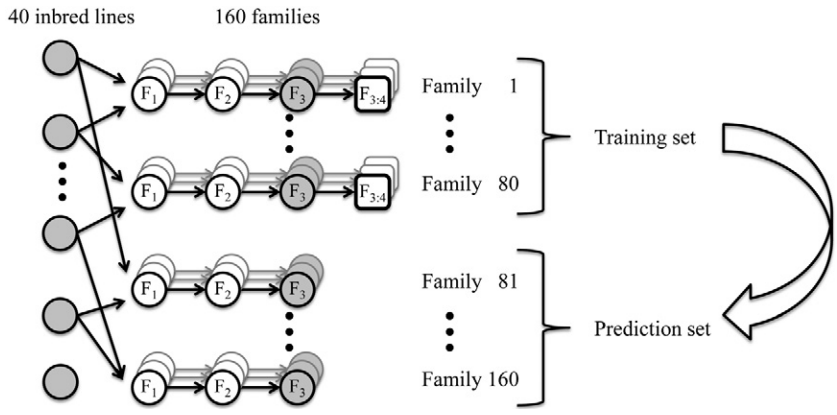

Fig. 1. Breeding program design with 40 inbred lines used to generate 160 families, of which 80 families comprised a genomic selection training set and 80 families comprised a prediction set. Circles represent individuals (shaded had genotypic data) and squares represent phenotypic data.

Imputation

Imputation increases the informativeness of low-coverage GBS data. The low-coverage GBS data had many data points with no or few sequence reads. To increase the quantity and quality of this data, we used the modified Hidden Markov Model of Li et al. (2010) as implemented in the AlphaImpute program version 1.3 (Hickey et al., 2012b; Antolin et al., 2017), available at http://www.AlphaGenes.Roslin.ed.ac.uk/AlphaSuite/AlphaImpute. Each family and chromosome was imputed independently. The imputation procedure involved:

1. Reading sequence reads
2. Constructing a set of template haplotypes
3. Estimating Hidden Markov Model parameters that describe mapping of the observed sequence reads onto the template haplotypes
4. Estimating (imputing) genotype probabilities given the observed sequence reads ($G_{i,j,0/0}$, $G_{i,j,0/1}$, and $G_{i,j,1/1}$)
5. Estimating allele dosages ($m_{i,j} = G_{i,j,0/1} + 2G_{i,j,1/1}$)

The input for imputation was the number of reference and alternative allele reads at each locus for the two parents and 400 F_3 individuals. The allele reads must be ordered according to marker positions on a chromosome. It was assumed that parents had high-quality GBS data with an equivalent of 20*x*. Preliminary tests showed that accurate imputation can be achieved in such a setting by constructing 100 template haplotypes and running the imputation procedure for 100 rounds in total with five rounds of burn-in. The average running time of imputing the 400 individuals was about 14 min with 1000 markers per chromosome and 99 min with 10,000 markers per chromosome. The template haplotypes were not available and had to be constructed from the nonimputed GBS data of the two crossed parents and their progeny in two steps. First, genotype probabilities were computed for each individual given the obtained numbers of reference and alternative alleles for the individual and observed allele frequency in the family. Second, the obtained genotype probabilities were used to sample template haplotype alleles one locus at a time.

Accuracy of GBS Data

The accuracy of GBS data was measured using the Pearson correlation between the GBS allele dosages and the true (SNP array) allele dosages. This was performed for both the nonimputed and imputed GBS data in two steps (Hickey et al., 2012a; Calus et al., 2014). First, marker allele dosages within each family were standardized (subtracted the mean and divided by standard deviation). If a marker was fixed within a family, we omitted dividing by standard deviation of zero to avoid numerical instabilities. Second, correlations were computed using the standardized marker allele dosages on a per-individual basis and averaged over individuals of a family and then over families.

Genomic Prediction

Genomic prediction of breeding values for nonphenotyped but genotyped individuals in a prediction set was based on estimates of associations between marker allele dosages and phenotypes in a training set. The construction of the training and prediction sets for each scenario is described in the subsection Scenarios. In general, the principle involved the construction of a training set by selecting a random set of genotyped and phenotyped families and the prediction of breeding values for nonphenotyped individuals in another set of families (i.e., across-family prediction). Marker associations were estimated using the ridge-regression model (Hoerl and Kennard, 1976; Whittaker et al., 1997; Meuwissen et al., 2001) as implemented in the program AlphaBayes, available at http://www.AlphaGenes.Roslin.ed.ac.uk/AlphaSuite/AlphaBayes. The model parameters were estimated using a Monte Carlo Markov Chain method, with one chain of 10,000 iterations with the first 1000 discarded as burn-in. Posterior means were used as estimates of marker associations.

Prediction Accuracy

The accuracy of genomic prediction was measured with the Pearson correlation between predicted and true breeding values. We measured accuracy in two ways, jointly across families and within each family, to account for the effect of family structure on genomic prediction (Windhausen et al., 2012). In the Results, we refer to this as the scope of prediction. The within-family correlation measures the accuracy of predicting the within-family variation commonly referred to as Mendelian sampling variation. The across-family correlation measures accuracy of predicting the within-family and between-family variation. The aim of genomic prediction is to capture variation due to both components, but it is harder and more important to capture the within-family variation because it is the variation that genomic selection primarily targets. Within-family variation is harder to predict because there is less information to accurately estimate unique variation within a family. We therefore focus largely on the within-family accuracy in the Results and Discussion.

Prediction Bias

The bias of genomic prediction was measured as the regression of the true breeding values on the estimated breeding values. This metric shows systematic underestimation or overestimation of estimated breeding values, which is a potential problem when genomic-based estimates are compared or combined with unbiased phenotype-based estimates. The desired value for this metric is one; values above one indicate underestimation, and values below one indicate overestimation. As with accuracy, we have computed bias within and across families to respectively measure the bias of estimated Mendelian sampling terms, and the joint bias of estimated parent average and Mendelian sampling terms.

Response to Selection

The response to selection was measured only for selection within a family for the same reasons as described for accuracy (see previous paragraph). It was calculated by subtracting the mean true breeding value of a family from the mean true breeding value of the top 10 individuals. Ranking of individuals was based on genomic predictions of breeding values.

Return on Investment

Return on investment was measured by dividing the response to selection within a family by the accrued genotyping costs to

achieve that response to selection. We expressed it relative to a chosen baseline so that all the evaluated scenarios could be compared. There are important assumptions behind this calculation, which we address in the Discussion. The chosen baseline was scenario 1 with SNP array data (see subsection Scenarios). We considered only the costs of genotypic data, as we assumed that a breeding program would already have phenotypic data available. For simplicity, any other costs were ignored. We divided the cost of genotyping the training set by 80 because we performed predictions in 80 families and all of them used the same training set. We believe this is a conservative choice as a breeding program could spread this cost over many more families generated over several cycles of genomic selection. The cost of genotyping the prediction set was considered for each family separately because the response to selection was measured for each family separately.

The cost of genotyping is determined by many factors. We have assumed only the following costs and factors: SNP array with 10K markers costs \$30, SNP array with 100K markers costs \$70, GBS library costs \$5, 10 sites in the genome need to be sequenced to obtain one reliable marker, sequence reads are 100 bases long, and 1x sequencing of 1 Gb costs \$100. The cost of GBS was calculated as the cost of preparing the GBS library plus the cost of sequencing a part of genome at a targeted sequencing coverage. For example, the cost of 10K GBS markers with coverage of 4x was \$9, of which library was \$5 and sequencing \$4 (= 10,000 targeted markers × 10 sequenced sites for a marker × 100 bases per sequence read × 4 coverage × $100/10^9$). With these assumptions, the costs of 10K GBS markers ranged between \$9.00 at 4x and \$5.01 at 0.01x (Table 1), whereas the costs of 100K GBS markers ranged between \$45.00 at 4x and \$5.10 at 0.01x (Table 1). We provide a spreadsheet in the supplement that details the calculations (Supplemental Table S1), which can be used to modify our assumptions.

Scenarios

We analyzed the simulated data in four sets of scenarios in which low-coverage GBS and imputation were used in either training, prediction, or both (Table 2, Supplemental Table S1). In each of the scenarios, we tested different sequencing coverages with the aim to either reduce the total cost of genotyping a fixed number of individuals or to increase the number of individuals genotyped at a fixed cost. When applied in training, the aim of this strategy was to increase prediction accuracy for a given investment. When applied in prediction, the aim of this strategy was to increase selection intensity for a given investment. Ultimately, both approaches should increase response to selection and return on investment. All scenarios involved the two sets of marker densities (10K and 100K) genotyped with SNP array or GBS technology. Genotyping-by-sequencing data used were either nonimputed or imputed.

Table 1. Assumed costs (US$) of 10,000 (10K) and 100,000 (100K) marker data with the single-nucleotide polymorphism (SNP) array or genotyping-by-sequencing (GBS) technology.

Technology	10K	100K
SNP array	30.00	70.00
GBS (coverage)		
4.00	9.00	45.00
2.00	7.00	25.00
1.00	6.00	15.00
0.50	5.50	10.00
0.25	5.25	7.50
0.10	5.10	6.00
0.05	5.05	5.50
0.01	5.01	5.10

The first scenario quantified the prediction accuracy, prediction bias, and return on investment when GBS data with different sequencing coverage were used in the training and prediction sets of fixed size (Table 2, Supplemental Table S1). The following sequencing coverages (x) were evaluated: 0.01x, 0.05x, 0.10x, 0.25x, 0.50x, 1.00x, 2.00x, and 4.00x. The training set consisted of 2000 genotyped and phenotyped individuals from a random set of 80 families, with each contributing a random set of 25 individuals. The prediction set consisted of 32,000 genotyped individuals from the remaining 80 families, with each family contributing 400 individuals.

The second scenario quantified the response to selection and return on investment when the prediction set (the number of selection candidates) was enlarged by reducing sequencing coverage per individual (Table 2, Supplemental Table S1). The total sequencing coverage used in the prediction set was always the same (25x) but was distributed amongst an increasing number of individuals in a family (25, 50, 100, 200, or 400). Specifically, the following five strategies were analyzed (denoted as y individuals genotyped with a sequencing coverage of x, y@x): 25@1x, 50@0.50x, 100@0.25x, 200@0.125x, and 400@0.0625x. These prediction sets had nonimputed or imputed GBS data. Predictions with SNP array data were also performed to assess the upper limit. The training set was the same as in the first scenario (2000 individuals) and was genotyped with SNP arrays to enable quantification of the effect of increasing the prediction set without confounding the results with the quality of genotypic data in training.

The third scenario quantified the prediction accuracy and return on investment when the training set was enlarged by reducing sequencing coverage per individual (Table 2, Supplemental Table S1). The total sequencing coverage used in training was always the same (2000x) but was distributed amongst an increasing number of individuals (25, 50, 100, 200, or 400) from each of the 80 training families. Specifically, the following five strategies were analyzed (denoted as y individuals genotyped with a sequencing coverage of x, y@x): 2000@1x, 4000@0.50x, 8000@0.25x, 16,000@0.125x, and 32,000@0.0625x. These training sets had nonimputed or imputed GBS data. Predictions with SNP array data were also performed to assess the upper limit. The prediction set was the same as in the first scenario, with 400 individuals per family genotyped with SNP arrays to enable quantification of the effect of increasing the training set without confounding the results with the quality of genotypic data in prediction.

The fourth scenario quantified the response to selection and return on investment when both the training and prediction sets were enlarged by reducing sequencing coverage per individual (Table 2, Supplemental Table S1). This scenario is a combination of all the strategies from the second and third scenarios (i.e., increasing both the size of the training or prediction set jointly). Both sets had nonimputed or imputed GBS data. Predictions with SNP array data were also performed to assess the upper limit.

Table 2. Summary of scenarios.

Scenario	Description	Training	Prediction	Metric
1	Fixed training and prediction sets with different sequencing coverage	SNP† array data or GBS‡ data (4*x* to 0.01*x*), fixed set size (2000)	SNP array data or GBS data (4*x* to 0.01*x*), fixed set size (400)	Prediction accuracy and bias and return on investment
2	Enlarged prediction set by reduced sequencing coverage per individual	SNP array data, fixed set size (2000)	GBS data (25@1*x*, 50@0.5*x*, 100@0.25*x*, 200@0.125*x*, 400@0.0625*x*)	Response to selection and return on investment
3	Enlarged training set by reduced sequencing coverage per individual	GBS data (2000@1*x*, 4000@0.5*x*, 8000@0.25*x*, 16,000@0.125*x*, 32,000@0.0625*x*)	SNP array data, fixed set size (400)	Prediction accuracy and return on investment
4	Enlarged training and prediction sets by reduced sequencing coverage per individual	GBS data (2000@1*x*, 4000@0.5*x*, 8000@0.25*x*, 16,000@0.125*x*, 32,000@0.0625*x*)	GBS data (25@1*x*, 50@0.5*x*, 100@0.25*x*, 200@0.125*x*, 400@0.0625*x*)	Response to selection and return on investment

† SNP, single-nucleotide polymorphism.

‡ GBS, genotyping-by-sequencing.

RESULTS

Low-coverage GBS data and imputation delivered the same level of prediction accuracy and bias as SNP array data and higher return on investment. Reducing individual sequencing coverage enabled increasing the size of the training set to increase prediction accuracy. It also enabled increasing the size of the prediction set to increase the response to selection. These two approaches can be combined to maximize the return on investment in a genomic selection program.

Accuracy of GBS Data

Accuracy of nonimputed GBS data was reduced substantially with reduced sequencing coverage, whereas accuracy of imputed GBS data was less so. This is shown in Fig. 2, which plots the accuracy of nonimputed and imputed GBS data within a family against the sequencing coverage at two marker densities. The accuracy of nonimputed GBS data (10K or 100K markers) was 0.91 at 4*x*, 0.68 at 1*x*, and reduced almost linearly to only 0.08 at 0.01*x*. Imputation increased accuracy at low coverages, particularly when 100K markers were used. High accuracy of imputed GBS data (≥0.97) was obtained when 10K markers were used and coverage was at least 0.50*x*, or when 100K markers were used and coverage was at least 0.05*x*.

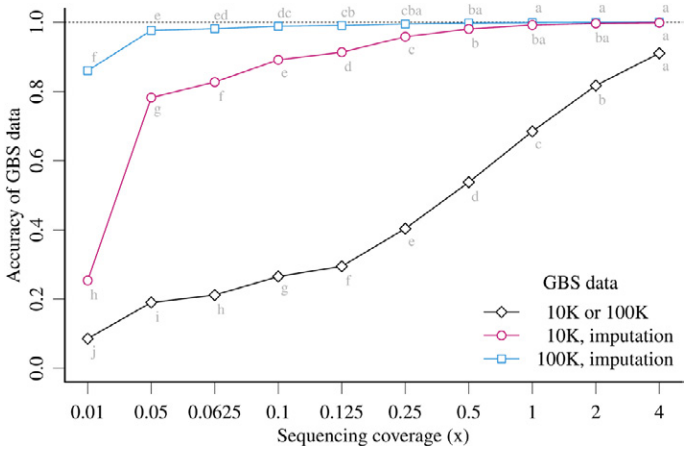

Fig. 2. Accuracy of nonimputed and imputed genotyping-by-sequencing (GBS) data within a family against the sequencing coverage at two marker densities (letters denote significant difference between different coverages within the type of GBS data at $p \leq 0.01$ according to the Tukey's multiple comparison test).

Fixed Size Training and Prediction Sets with Different Sequencing Coverage (Scenario 1)

The prediction accuracy using nonimputed GBS data with medium to low sequence coverage was equivalent to using SNP array data, whereas imputation extended this equivalence to very low sequence coverage. This is shown in Fig. 3, which plots the prediction accuracy within a family and across families against sequencing coverage at two marker densities using nonimputed or imputed GBS data or SNP array data. The prediction accuracy within a family was ~0.50 when using SNP array data, irrespective of marker density. The same level of accuracy (>95%) was achieved when using nonimputed GBS data with sequencing coverage of at least 2*x* with 10K markers and 0.5*x* with 100K markers. At lower coverages, the accuracy dropped; at 0.01*x*, the accuracy was 0.11 with 10K markers and 0.22 with 100K markers. Imputation recovered the loss of accuracy at low coverages. Specifically, it moved the turning point in the loss of accuracy from 2*x* to 0.1*x* when 10K markers were used and from 0.5*x* to 0.01*x* when 100K markers were used. The prediction accuracy across families was higher than within a family and was less influenced by the decreasing sequencing coverage. The small loss in accuracy across families was also recovered by imputation.

Prediction bias within a family was large with low-coverage GBS. This is shown in Fig. 4, which plots the prediction bias within a family and across families against sequencing coverage at two marker densities using nonimputed or imputed GBS data or SNP array data. Predictions were slightly overestimated with SNP array data (bias was 0.91 with both marker densities) and progressively underestimated with nonimputed low-coverage GBS data. At 0.01*x*, the bias was 15.59 with 10K markers and 65.76 with 100K markers. Imputation removed underestimation at all

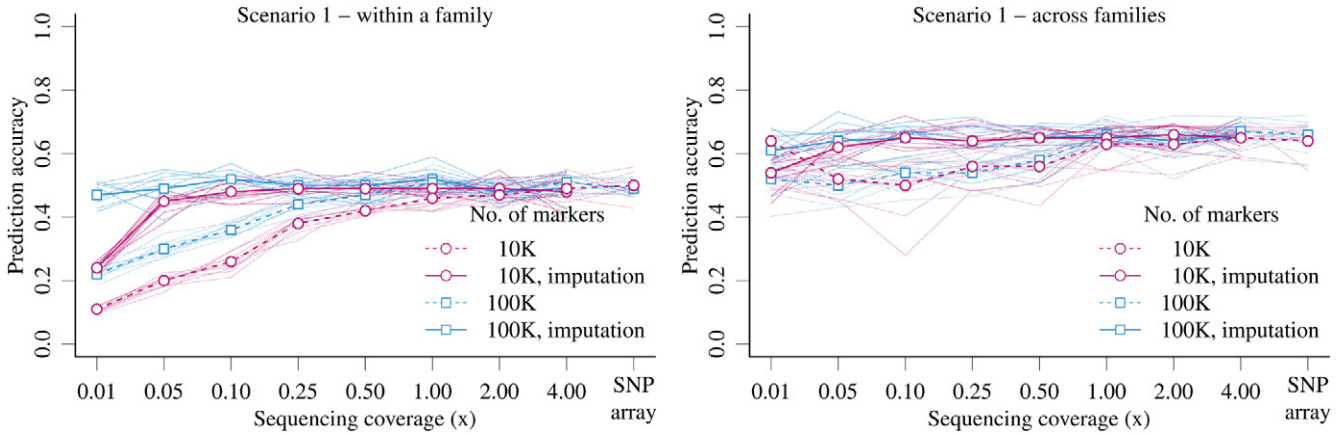

Fig. 3. Prediction accuracy within a family (left) and across families (right) against sequencing coverage at two marker densities. A training set had 2000 individuals and a prediction set had 400 individuals (individual replicates are represented with thin lines and average with a thick line). SNP, single-nucleotide polymorphism.

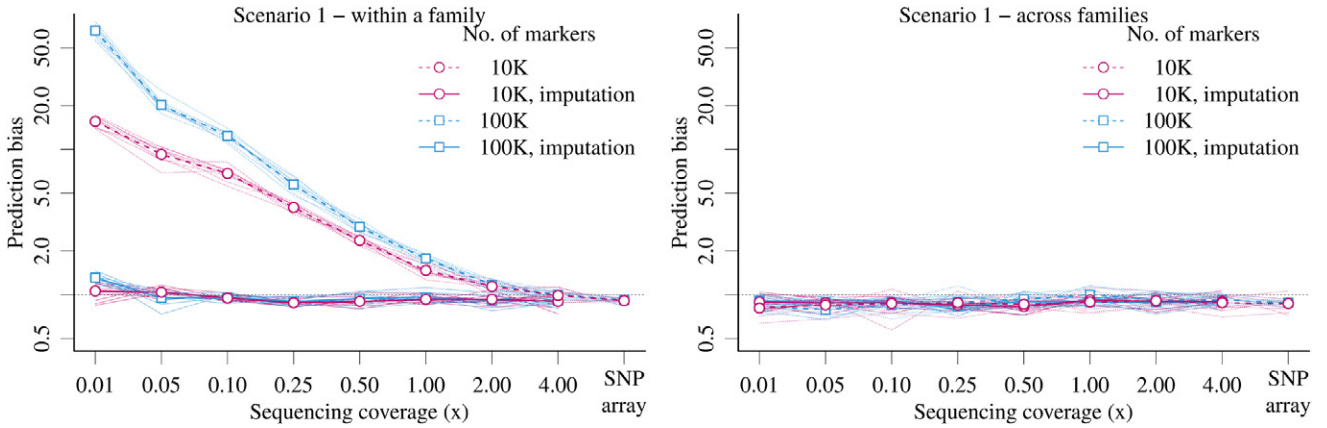

Fig. 4. Prediction bias (on a log scale) within a family (left) and across families (right) against sequencing coverage at two marker densities. A training set had 2000 individuals and a prediction set had 400 individuals (individual replicates are represented with thin lines and average with a thick line). SNP, single-nucleotide polymorphism.

of tested sequencing coverages, except when 100K markers had the lowest sequence coverage (0.01*x*, the bias was 1.31). Contrary to predictions within a family, there was no underestimation in predictions across families.

Using low-coverage GBS data and imputation increased the return on investment in comparison with using SNP array data. This is shown in Fig. 5, which plots the return on investment for selection within a family against sequencing coverage at two marker densities using nonimputed or imputed GBS data or SNP array data. The baseline for comparison was a strategy where 10K SNP array data was used. Using GBS data instead of SNP array data gave higher return on investment. The optimal sequencing coverage depended on marker density and use of imputation. When we used 10K markers and no imputation, the return on investment with GBS data was highest at 0.5*x*, 4.66 times that of the baseline scenario. Imputation increased the return on investment by recovering information at lower coverages; it was highest at 0.1*x*, 5.63 times that of the baseline scenario. When we used 100K markers, the return on investment was generally lower than with 10K markers, except at very low coverages and when imputation was used. The highest return on investment with 100K markers and no imputation was achieved at 0.1*x*, 3.66 times that of the baseline scenario. Imputation increased it to the same level as achieved with 10K markers, but only when coverage was below 0.1*x*.

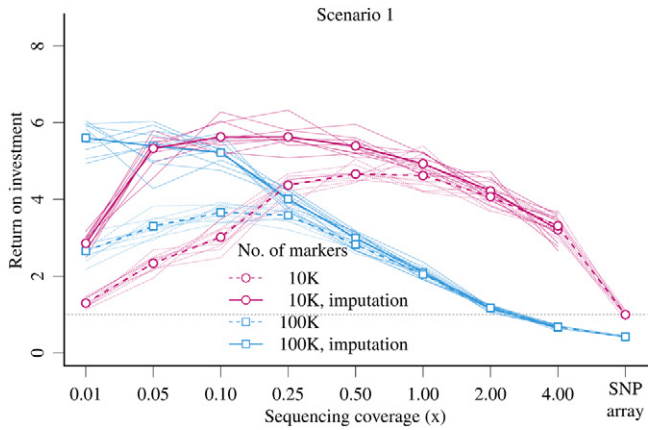

Fig. 5. Return on investment for selection within a family against sequencing coverage at two marker densities. A training set had 2000 individuals and a prediction set had 400 individuals (individual replicates are represented with thin lines and average with a thick line). SNP, single-nucleotide polymorphism.

Enlarged Prediction Set with Reduced Sequencing Coverage (Scenario 2)

Enlarging a prediction set by reducing sequencing coverage per individual more than doubled the response to selection. This is shown in Fig. 6, which plots the response to selection within a family at two marker densities against an increasing prediction set that either had SNP array data or GBS data with decreasing sequencing coverage per individual (the total coverage was always 25*x*). The GBS data were either nonimputed or imputed. As expected, an increased response to selection was observed when enlarging a prediction set that had SNP array data (from 0.22 to >0.50). An increased response to selection was also observed when GBS data with decreasing coverage were used. However, the response to selection stopped increasing when the coverage was reduced below 0.25*x* with 10K markers or 0.125*x* with 100K markers, because such GBS data failed to capture sufficient amounts of genetic variance among selection candidates. Imputation recovered this loss for all the tested combinations so that the imputed GBS data enabled the same level of response to selection as the SNP array data.

Return on investment with an enlarged prediction set was highest with intermediate numbers of selection candidates that had GBS data with intermediate sequencing coverages. This is shown in Fig. 7, which plots the return on investment for selection within a family at two marker densities against an increasing prediction set that had either SNP array data or GBS data with decreasing sequencing coverage per individual (the total coverage was always 25*x*). The GBS data used were either nonimputed or imputed. The baseline for comparison was a strategy where 10K SNP array data were used in a prediction set of 400 selection candidates. When 10K markers were used, the highest return on investment was achieved by selecting among 50 selection candidates with 0.5*x* GBS data (7.19 and 7.67 times that of the baseline scenario without and with imputation, respectively) or 100 selection candidates with 0.25*x* GBS data (6.80 and 7.50 times that of the baseline scenario without and with imputation, respectively). Increasing the number of selection candidates beyond 100 started to decrease return on investment, even when imputation was used. When 100K markers were used, the same pattern was observed, but the returns on investment were almost half those obtained with 10K markers.

Enlarged Training Set with Reduced Sequencing Coverage (Scenario 3)

Enlarging the training set by reducing sequencing coverage per individual and using imputation nearly doubled prediction accuracy. This is shown in Fig. 8, which plots the prediction accuracy within a family at two marker densities against an increasing training set that had either SNP array or GBS data with decreasing sequencing coverage per individual (the total coverage was always 2000*x*). The GBS data used was either nonimputed or imputed. When the SNP array data were used, enlarging the training set from 2000 to 32,000 individuals increased prediction accuracy from ~0.50 to >0.80. Accuracy also increased when GBS data with reduced sequencing coverage were used, but the increase was smaller as accuracy started to level off when training individuals were sequenced at coverage of <0.5*x*, irrespective of marker density. Imputation recovered this loss entirely for all the tested combinations.

Return on investment with an enlarged training set was highest when a large training set had 10K markers sequenced at low coverage and imputed. This is shown in Fig. 9, which plots the return on investment for selection within a family at two marker densities against an increasing training set that either had GBS data with decreasing sequencing coverage per individual (the total coverage was always 2000*x*) or SNP array data. The GBS data used were either nonimputed or imputed. The baseline for comparison was a strategy where 10K SNP array data were used in the training set of 2000 individuals. Enlarging the training set genotyped with a

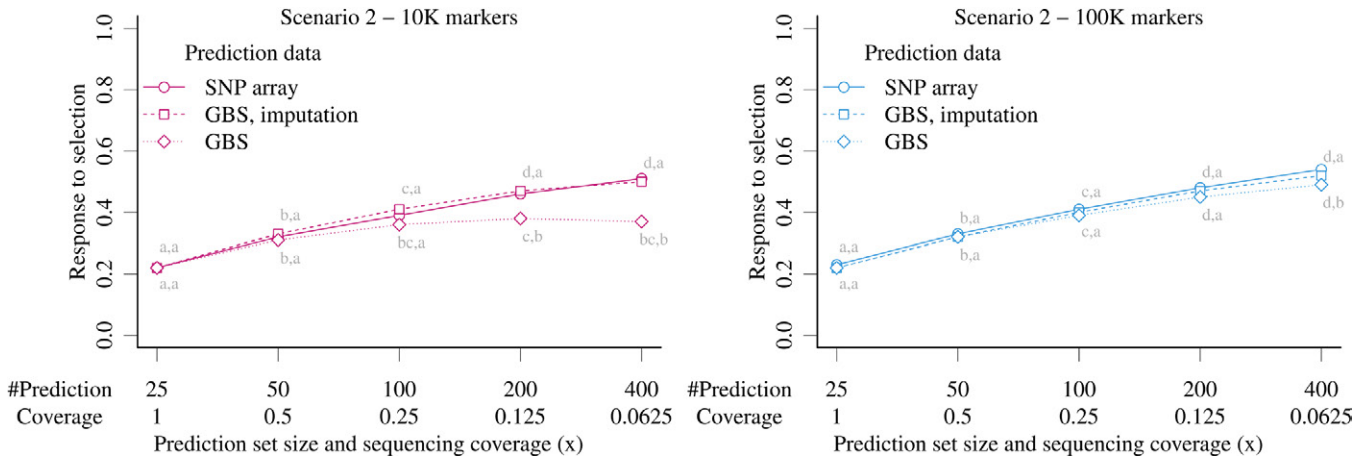

Fig. 6. Response to selection within a family with 10,000 (10K, left) or 100,000 (100K, right) markers against increasing prediction set with single-nucleotide polymorphism (SNP) array data or genotyping-by-sequencing (GBS) data at different sequencing coverage. A training set had 2000 individuals with SNP array data (letters denote significant difference against an enlarged training set [the first letter] and the SNP array data [the second letter] at $p \leq 0.01$ according to the Tukey's multiple comparison test).

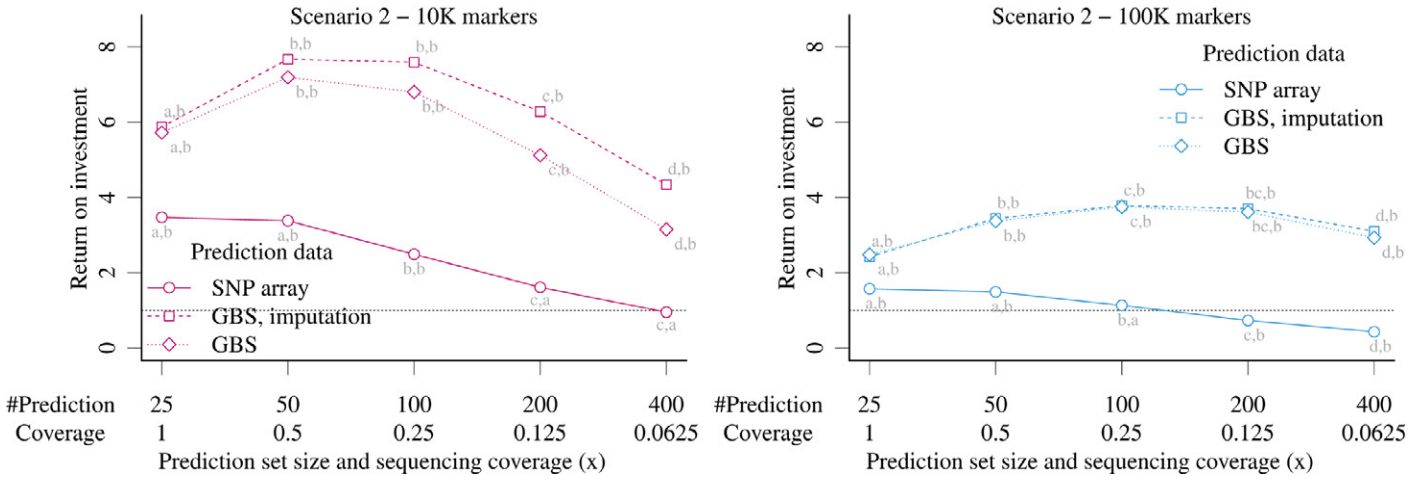

Fig. 7. Return on investment for selection within a family with 10,000 (10K, left) or 100,000 (100K, right) markers against increasing prediction set with single-nucleotide polymorphism (SNP) array data or genotyping-by-sequencing (GBS) data at different sequencing coverage. A training set had 2000 individuals with SNP array data (letters denote significant difference against an enlarged training set [the first letter] and the baseline SNP array data scenario [the second letter] at $p \leq 0.01$ according to the Tukey's multiple comparison test).

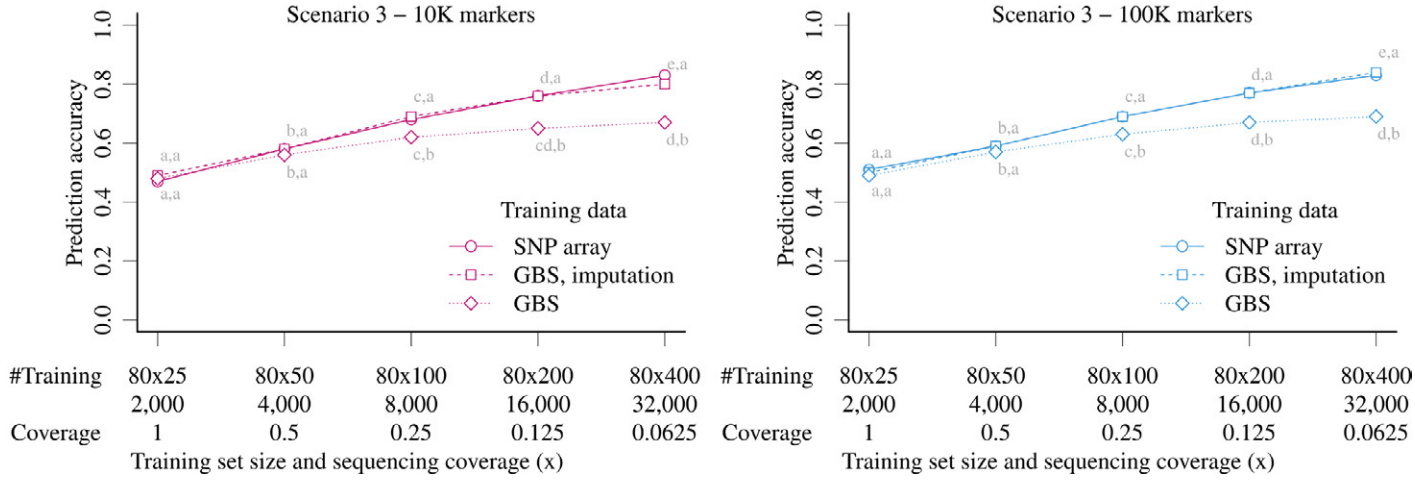

Fig. 8. Prediction accuracy within a family with 10,000 (10K, left) or 100,000 (100K, right) markers against increasing training set with single-nucleotide polymorphism (SNP) array data or genotyping-by-sequencing (GBS) data at different sequencing coverage. A prediction set had 400 individuals with SNP array data (letters denote significant difference against an enlarged training set [the first letter] and the SNP array data [the second letter] at $p \leq 0.01$ according to the Tukey's multiple comparison test).

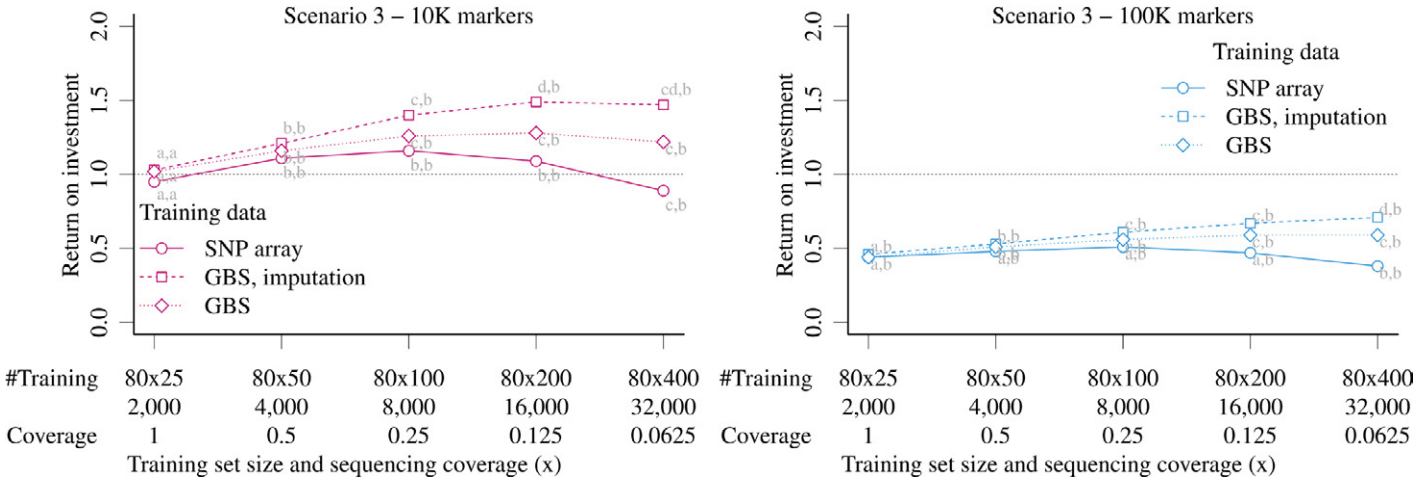

Fig. 9. Return on investment for selection within a family with 10,000 (10K, left) or 100,000 (100K, right) markers against increasing training set with single-nucleotide polymorphism (SNP) array data or genotyping-by-sequencing (GBS) data at different sequencing coverage. A prediction set had 400 individuals with SNP array data (letters denote significant difference against an enlarged training set [the first letter] and the baseline SNP array data scenario [the second letter] at $p \leq 0.01$ according to the Tukey's multiple comparison test).

10K marker SNP array increased return on investment, but only marginally—it was highest (1.16 times that of the baseline scenario) with a training set size of 8000 individuals, whereas increasing the training set further reduced return on investment. Using 10K GBS markers gave a higher return on investment—up to 1.28 times that of the baseline scenario when 16,000 training individuals had 0.125*x* GBS data that were not imputed and up to 1.49 times that of the baseline scenario when 16,000 training individuals had 0.125*x* GBS data that were imputed. When 100K markers were used, the return on investment was substantially lower, and while enlarging the training set increased returns, they were still considerably smaller than with 10K markers; the highest return was 0.71 times that of the baseline scenario when a training set had 32,000 individuals with 0.0625*x* GBS data that were imputed.

Enlarged Training and Prediction Sets with Reduced Sequencing Coverage (Scenario 4)

Enlarging both the training and prediction sets by reducing sequencing coverage per individual, and using imputation quadrupled the response to selection. This is shown in Fig. 10, which plots the response to selection within a family at two marker densities against increasing training and prediction sets that had either SNP array data or GBS data with decreasing sequencing coverage per individual (the total coverage was always 2000*x* in training and 25*x* in prediction). The GBS data used were either nonimputed or imputed. Increasing the size of the training and prediction sets increased the response to selection, as expected. Using SNP array data with a training set of 2000 individuals and a prediction set of 25 selection candidates gave a response to selection of 0.21 with 10K and 0.23 with 100K markers. Increasing the training set to 32,000 individuals and the prediction set to 400 selection candidates increased the response to selection to 0.89 with 10K and 0.88 with 100K markers, which is a fourfold increase. Nonimputed GBS data achieved half of this potential with 10K markers and three quarters of this potential with 100K markers. Imputed GBS data achieved 90% of this potential with 10K markers and reached the potential with 100K markers.

Return on investment with enlarged training and prediction sets varied substantially and was strongly dependent on the size of sets and marker density. This is shown in Fig. 11, which plots the return on investment for selection within a family at two marker densities against increasing training and prediction sets that had either SNP

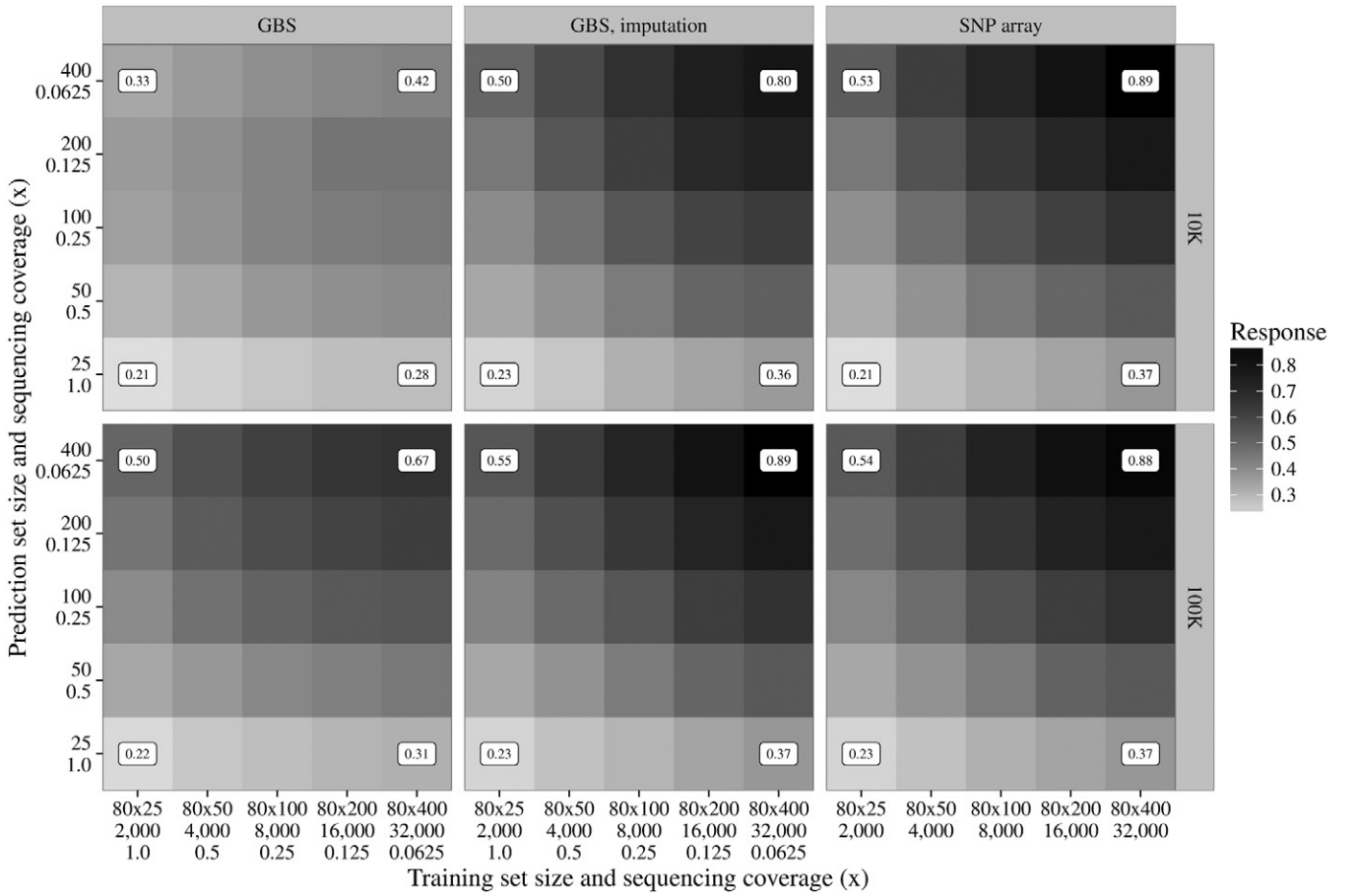

Fig. 10. Response to selection within a family with 10,000 (10K, top row) or 100,000 (100K, bottom row) markers against increasing training and prediction sets with single-nucleotide polymorphism (SNP) array data or genotyping-by-sequencing (GBS) data at different sequencing coverage.

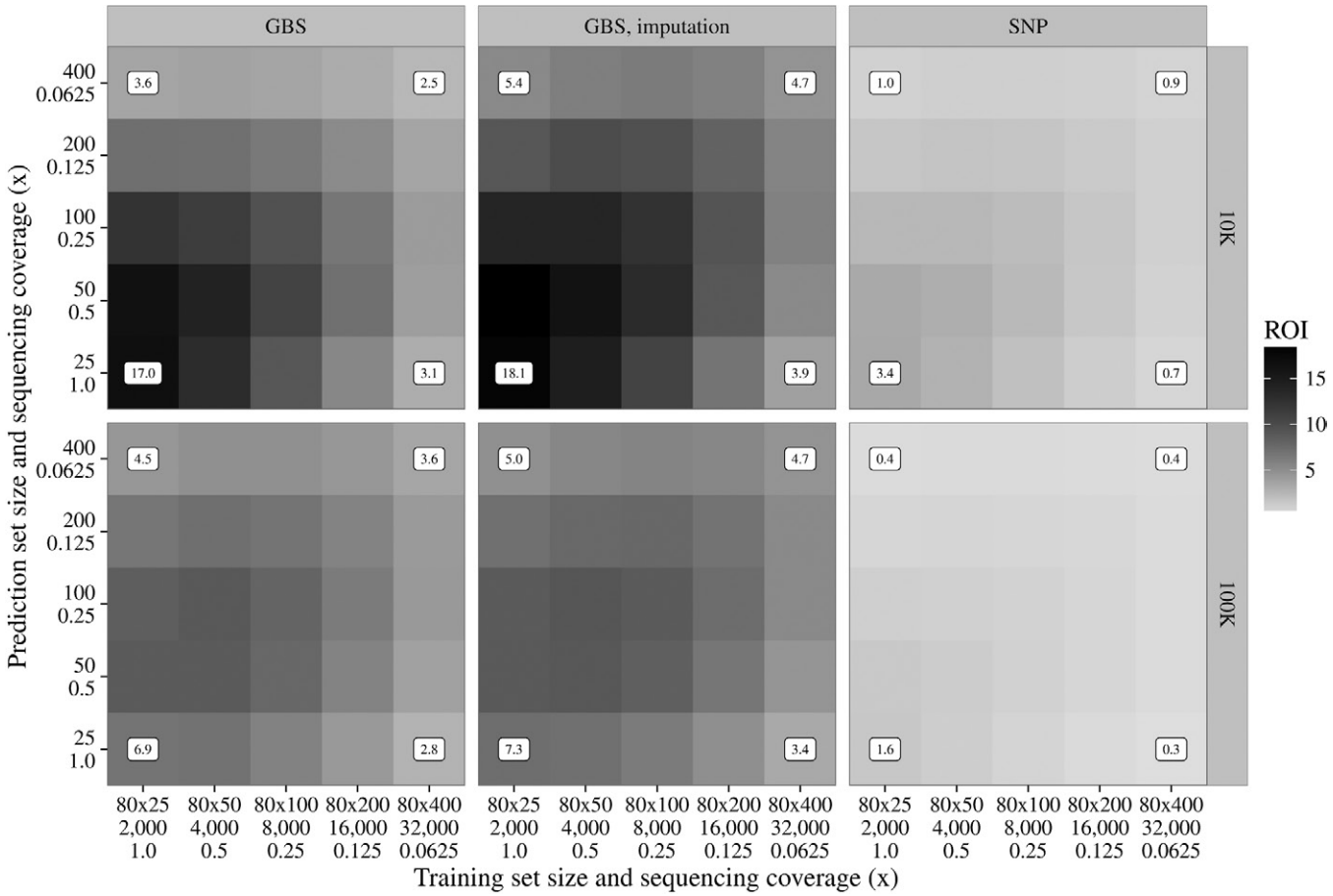

Fig. 11. Return on investment for selection within a family (ROI) with 10,000 (10K, top row) or 100,000 (100K, bottom row) markers against increasing training and prediction sets with single-nucleotide polymorphism (SNP) array data or genotyping-by-sequencing (GBS) data at different sequencing coverage.

array data or GBS data with decreasing sequencing coverage per individual (the total coverage was always 2000*x* in training and 25*x* in prediction). The GBS data used were either nonimputed or imputed. The baseline for comparison was a strategy where 10K SNP array data were used in a training set of 2000 individuals and a prediction set of 400 selection candidates. Averaged over all of the tested strategies, the return on investment was highest using imputed GBS data with 10K markers (9.3 times that of the baseline scenario), followed by using nonimputed GBS data with 10K markers (7.3 times that of the baseline scenario), using imputed GBS data with 100K markers (6.4 times that of the baseline scenario), and using non-imputed GBS data with 100K markers (5.8 times that of the baseline scenario). Using SNP array data was less cost effective than using GBS data. The highest return on investment (18.3 times that of the baseline scenario) was obtained with a strategy of building a training set with 2000 individuals that had imputed 1*x* GBS data and a prediction set with 50 selection candidates that had imputed 0.5*x* GBS data. The next two best strategies were to either halve the prediction set to 25 selection candidates or double the training set to 4000 individuals. In general, when we used either nonimputed or imputed GBS data, it was more cost effective to increase the prediction set than to increase the training set.

DISCUSSION

The results show that low-coverage GBS and imputation enable accurate, unbiased, and cost-effective genomic selection in biparental segregating populations. These results highlight three main topics for discussion, specifically (i) accuracy of low-coverage GBS data for genomic selection, (ii) cost-effective genomic selection with low-coverage GBS and imputation, and (iii) the assumptions made by the study.

Accuracy of Low-Coverage GBS Data for Genomic Selection

Homozygosity of sequenced individuals affects the required sequencing coverage and cost to obtain accurate GBS data. Genotyping-by-sequencing interrogates an individuals' genomewide genotype by generating sequence reads at targeted genome sites. A single sequence read provides genomic information about one chromosome of the targeted site and therefore about a single allele of one individual. The quality of inferred allele dosages and genotypes from such data depends largely

on the number of sequence reads per locus, the sequencing coverage (x). When the aim is to obtain very precise information about genotype at many loci, deep sequencing is required with coverage of 20 to 30x or more. Such deep sequencing is too expensive for routine breeding program activities that involve large numbers of individuals. However, if there is little variation between the alleles of an individual, as is the case with inbred individuals, sequencing coverage can be greatly reduced. In absence of sequencing error, sequencing each targeted site once would suffice and would keep cost low. This is one of the reasons why low-coverage GBS data is attractive to plant breeders (Poland et al., 2012b; Crossa et al., 2013), in addition to other benefits that GBS technology has over SNP arrays (e.g., capturing private genetic variation; Elshire et al., 2011; Heslot et al., 2013).

Low-coverage GBS data allows for accurate genomic selection in biparental segregating populations. The potential low cost of low-coverage GBS data (Poland et al., 2012b) and positive initial reports of accurate genomic predictions with such data in inbred individuals (Poland et al., 2012b; Crossa et al., 2013) inspired us to quantify the potential of low-coverage GBS data for genomic selection in segregating populations. We hypothesized that there is an intermediate coverage that potentially reduces prediction accuracy but increases return on investment through lower costs. Such a strategy would enable plant breeders to implement genomic selection more aggressively, such as in early-segregating populations, where the potential for genomic selection is the greatest (Bernardo and Yu, 2007; Bassi et al., 2016). The results show that GBS data with coverage of about 1x delivers comparable prediction accuracy and bias to SNP array data and higher return on investment. This is in line with our previous simulation studies in an outbred livestock population (Gorjanc et al., 2015).

It is surprising that such low coverage (~1x) is sufficient, because low-coverage GBS data have two major drawbacks in comparison with SNP array data in segregating populations. First, due to the random sequencing process, some sites are sequenced more than once and some sites are not sequenced at all, which means that low-coverage GBS data has many missing data points. When we did not use imputation, we filled these missing data points with the naïve imputation of two times the allele frequency in a family. Second, sequencing coverage of 1x provides only one sequence read per marker on average, which means that, on average, only one allele is sequenced and consequently all heterozygous loci are assumed homozygous. This is potentially a severe drawback for analysis of segregating populations.

Five complementary factors explain why low-coverage GBS data provides sufficient information for accurate genomic selection in biparental segregating populations. First, 1x GBS data provides a lot of information for genomic selection. The correlation between 1x GBS allele dosages and the true allele dosages within a family was 0.68 without imputation and 0.99 with imputation. These values show that the low-coverage GBS data can be expected to capture association signal between markers and phenotypes within a family, albeit the data are noisier than the true state. Second, there are only about a quarter to a third of markers segregating in a type of family used in this study, which means that the rest of markers will have correct GBS data even at low coverage. Third, large linkage blocks that segregate in families with inbred parents preserve association signals and counterbalance some noise in the GBS data. Fourth, genomic selection models usually fit only additive effects, and providing GBS data in the form of realized allele dosages to these models captures all the available information and uncertainty without the need to first call the genotypes correctly. Fifth, the genomic selection models tend to be heavily overparameterized, and the joint information from tens of thousands or hundreds of thousands of markers can counterbalance some noise in the GBS data. When sequencing coverage is reduced too much, however, the signal-to-noise ratio is reduced and prediction accuracy decreases.

Imputation can recover loss of information in low-coverage GBS data. When we used imputation, we were able to recover most of the missing or erroneous information in GBS data with extremely low-coverage data (even as low as 0.10x with 10K genomewide markers and 0.01x with 100K genomewide markers) and recover prediction accuracy and bias almost to the level of the 1x GBS data or the SNP array data. Although this might seem another surprising result, it is expected because biparental populations derived from inbred parents are ideal for imputation (Swarts et al., 2014; Hickey et al., 2015; Jacobson et al., 2015). Our results show that, in such a setting, ~100 sequence reads at marker sites of a chromosome (100 = 0.10x × 10K genomewide markers/10 chromosomes or 0.01x × 100K genomewide markers/10 chromosomes) provide enough information for accurate imputation of marker data and consequently accurate, as well as cost-effective, genomic selection. These results complement previous studies that optimized sequencing coverage for genomic analyses of complex traits (Li et al., 2011; Pasaniuc et al., 2012). However, these values need to be interpreted from the perspective of haplotypes and not individuals. With 100 sequence reads per chromosome in a family with 400 individuals, we effectively generate 40,000 allele reads for the two parental haplotypes that segregate in a biparental family. All of these allele reads are used to construct the template haplotypes, to estimate hidden Markov model parameters, and finally to impute the missing alleles or whole genotypes of each individual. The optimal number of reads that need to be generated

both per individual and per haplotype such that imputation accuracy is high will be subject to future research.

Low sequencing coverage has a larger effect on the prediction accuracy and bias of the Mendelian sampling term than of the parent average term. We observed consistently lower accuracy and greater bias with nonimputed low-coverage GBS data when predicting within a biparental family than when predicting across biparental families. The main cause of this is that individuals within a family differ genetically due to segregation of parental genomes, whereas individuals across families differ genetically due to differences among parental genomes and segregation. Accurately predicting the variation due to segregation equals explaining the Mendelian sampling term of a breeding value, whereas accurately predicting the variation due to parental genomes and segregation equals explaining both the parent average term and the Mendelian sampling term of a breeding value (VanRaden and Wiggans, 1991; Mrode, 2005). In our study, we assumed that the parents had 20*x* GBS data, which means that we had very accurate information about differences among parental genomes. This resulted in almost constant accuracy of genomic prediction across families over a range or sequencing coverages, whereas this was not the case for prediction within a family. For example, with 10K nonimputed GBS markers with coverage of 1*x* and 0.05*x*, the prediction accuracy across families was 0.63 and 0.52, respectively, while prediction accuracy within a family was 0.46 and 0.20, respectively.

These results indicate that care should be taken when interpreting prediction accuracies with nonimputed low-coverage GBS data across families, because such data seem to capture mostly variation between (sub)population structures (e.g., families) and not within. This is in line with the observations in population genomic studies (Buerkle and Gompert, 2013). Capturing the within-family variation is one of the prime reasons to perform genomic selection, particularly in early-segregating populations. This is because breeders usually have a good indication about the breeding value of parents and can accurately assess the parent average of their progeny without collecting data on the progeny. On the other hand, breeders have no information about the Mendelian sampling variation amongst progeny in a family, which can be captured using genomic prediction. In addition, sustainable long-term genetic gain depends mostly on the ability to capture the within-family (Mendelian sampling) variation (Woolliams et al., 1999).

Cost-effective Genomic Selection with Low-Coverage GBS and Imputation

Low-coverage GBS and imputation enable substantial reduction of genotyping costs in genomic selection programs. One of the barriers for adoption of genomic selection is the high cost of genomewide genotyping large numbers of individuals. Our results show that by using low-coverage GBS breeders can reduce genotyping costs substantially, particularly in combination with imputation. We first evaluated the benefit of such data in a fixed-size genomic selection program. Using 10K nonimputed GBS markers sequenced at 0.5*x* gave return on investment that was 4.66 times higher than using the SNP array data. This improvement was brought about by a large reduction in costs (82%) and only a small reduction in prediction accuracy (16%). Imputation further increased the return on investment to 5.60 times that with the SNP array data by reducing sequencing coverage even more to 0.10x with 10K markers and 0.01*x* with 100K markers.

There are two rate-limiting factors in increasing the return on investment through reducing the sequencing coverage in a fixed-size genomic selection program. First, when the sequence coverage is reduced too much, the accuracy of resulting GBS data decreases and the consequent loss in prediction accuracy is not counterbalanced by the decreased cost of GBS data. The rate of this decrease depends on whether imputation is used or not. When imputation is used, the coverage can be reduced to ~100 sequence reads at marker sites of a chromosome. Second, the cost of GBS data has a fixed component, of which preparation of the sequencing library is the key part. When sequencing coverage is reduced to a low level, the fixed-cost component dominates, and any further reduction in sequencing coverage does not reduce the total cost substantially. For example, in the scenarios that gave the highest return on investment (10K markers with 0.1*x* or 100K markers with 0.01*x*), library accounted for 98% of the genotyping cost.

Low-coverage GBS and imputation enable the assembling of larger and more cost-effective genomic selection programs. Effective genomic selection programs have large training sets that enable high prediction accuracy and large prediction sets that enable high selection intensity. When these two principles are combined, a genomic selection program can deliver high response to selection. However, it is expensive to genotype a large number of individuals, and a balance needs to be found that maximizes return on investment. The results show that low-coverage GBS and imputation can be leveraged to increase the training set, the prediction set, or both in a cost-effective manner. The results are qualitatively similar to our previous simulation study that evaluated the prospect of cost-effective genomic selection through imputation of SNP array data (Gorjanc et al., 2016). The main difference in this study was that we were able to impute GBS data with higher accuracy, which allowed us to reduce sequencing coverage considerably, genotype larger training and prediction sets, and achieve higher returns on investment.

Expanding the training set gave smaller changes in return on investment than expanding the prediction set. This result is due to three interacting factors. First, we evaluated return on investment for selection within a family, and the costs accrued for this were genotyping the selection candidates and genotyping the training individuals. Since we used the training set for predictions in several families, we divided the cost of genotyping the training set by the number of families predicted. Therefore, any cost reductions in assembling a larger training set by lowering sequencing coverage per individual were diluted. Second, the initial size of the training set (2000 individuals) provided a good baseline with prediction accuracy (within a family) of about 0.5. Although expanding the training set size did increase prediction accuracy (up to ~0.8 when the training set comprised 32,000 individuals), the additional cost of genotyping many more individuals limited the increase in return on investment. Third, the design of the simulation allowed us only to assess cost effectiveness of genomic selection in closely related individuals, meaning the predicted families had one parent in common with some of the training families and "background" relationships with the other training families. Having a large training set in such a case is not as important as when more distantly related individuals are predicted (Clark et al., 2012; Pszczola et al., 2012; Hickey et al., 2014). The ability to build large training sets at low cost is, however, very important for enabling high prediction accuracy across several cycles of genomic selection (Michel et al., 2016; Pszczola and Calus, 2016).

Expanding the prediction sets (increasing selection intensity) increases response to selection, but this benefit needs to be balanced against larger genotyping costs. The results showed that expanding the prediction sets with SNP array data was not cost effective, while it was with both nonimputed and imputed GBS data, although the scope for increase was also limited with GBS data. The reason for this is that there is limited genetic variation within biparental families. This diversity can be sampled well with a limited number of candidates, and further increases will lead to diminishing return in genetic gain and even more so in return on investment. In this study, the breakpoint was 50 individuals per family. We expect that such a breakpoint would be higher in multiparental families.

Assumptions of the Study

The results are based on four important assumptions. These are (i) known marker order, (ii) the way we generated GBS data, (iii) the type of populations analyzed, and (iv) the assumed costs and return on investment calculation.

We have assumed that marker order was known. All accurate imputation methods rely heavily on at least approximate marker order to infer haplotypes and impute them in either low-density genotyped or low-coverage sequenced individuals. When marker order is not known, it is not possible to perform these two core tasks of imputation. In such a case, alternative imputation methods can be used (Rutkoski et al., 2013), but these methods are much less effective to deliver a cost-effective genotyping strategy than the type of method used in this study. Although ordering markers on a genome is not trivial, the recent developments in sequencing technologies hold promise in obtaining more marker maps for plant genomes (Michael and VanBuren, 2015; VanBuren et al., 2015; Chaney et al., 2016; Staňková et al., 2016). An implied assumption behind a known marker order is also that sequence reads can be uniquely aligned to originating chromosomes. Nonunique alignment is an issue in polyploid crops. Possible solutions include longer reads and pair-end reads.

We have generated the GBS data by sampling alleles from targeted loci of a genome that were also assumed to be present on the SNP array. Although the same set of loci might not be targeted with the two technologies, we did that to enable comparison of the two approaches of generating the genomewide data, the almost-exact SNP array technology versus probabilistic GBS technology. In generating both types of data, we ignored errors because sequencing error rates and error rates of array genotyping tend to be small ($\leq$1%) and were not expected to change the results. However, we have not ignored the uncertainty in sequence data regarding the true genotype of an individual when coverage is low. We have taken this uncertainty into account by working with continuous allele dosages instead of attempting to call marker genotypes prior to data analysis. We have computed the allele dosage from the sequence reads when we used the nonimputed GBS data and from the inferred genotype probabilities given the sequence reads in a family when we used the imputed GBS data. We also emphasize that, while the sequencing and array errors do generate some amount of erroneous data, the amount of this noise is negligible to the amount of variability in quantitative traits. Finally, while the simulation of GBS process followed the approach of other whole-genome sequencing simulation studies (Li et al., 2010; Pasaniuc et al., 2012), it is possible that some artifacts of sequencing process were not captured (Escalona et al., 2016). However, given that protocols are continually being improved (Islam et al., 2015; Ali et al., 2015; Schröder et al., 2016; Fu et al., 2016) we believe that our results adequately indicate the potential of low-coverage GBS data for genomic selection.

The recommendations for extremely low sequencing coverage in this study depend, to a large extent, on the type of populations analyzed. The biparental families in this study were derived from inbred parents, which enables very accurate imputation (Swarts et al., 2014; Hickey et al., 2015; Jacobson et al., 2015). This allowed us to reduce the sequencing coverage per individual considerably to

reduce the cost of a genomic selection program of fixed size or to generate data for a genomic selection program of larger size at a given cost. Other types of populations (e.g., top-crosses and other multiparental crosses) and sizes of population might need higher sequencing coverages to obtain the same benefits. Although this will be addressed in future research, we would like to point out that our previous experience with imputation of SNP array data shows comparable accuracy with biparental crosses, backcrosses, and top-crosses (Hickey et al., 2015). Finally, we would like to emphasize that, even when imputation is not used, the 1*x* GBS data seems to be adequate for both the type of populations studied here and for general outbred populations, provided that marker density is sufficient (Gorjanc et al., 2015).

We have made several assumptions about the costs and the calculation of return on investment. We have assumed only the cost of genotypic data and not the cost of phenotypic data because the main aim was to evaluate the potential of GBS versus SNP array data, and we also assumed that a breeding program would have historical phenotypic data and biological samples. We also assumed that the parents have accurate genotypic data and did not factor in this cost, because the number of parents tends to be small and they are likely to be genotyped in previous cycles (perhaps with low-coverage GBS prior to selection and with topped-up coverage after selection). We also assumed that the cost of genotyping the training set is spread over predicting several families, but only for one cycle of predictions, which is a very conservative approach. The assumed cost of the two SNP arrays was based on inquiries from different providers under the assumption that a large breeding program could benefit from the economy of scale. The exact cost of GBS data at different sequencing coverages is difficult to obtain from providers. We have approximated the cost of GBS data by considering the cost of library preparation from Rowan et al. (2015), assuming that we will have to sequence 10 sites in the genome to get one polymorphic marker, and using the sequencing cost based on inquiries from different providers. The provided spreadsheet in the Supplemental Material can be used to modify all of these assumptions and recalculate returns on investment. We have performed a sensitivity analysis of all the described factors and found out that, while the return on investment values change, the overall patterns largely do not.

Calculation of the return on investment in this study is approximate, because the value of genetic gain depends on the seed market in a nonlinear way, which is difficult to quantify. In absence of such information, we have assumed that the value of one unit of genetic gain is much larger than the costs and that the relationship between the value of return and genetic gain is linear, and we have expressed the ratio of genetic gain over cost for tested scenarios versus the baseline scenario. Under these assumptions, the reported values converge to the ratio of return on investments, where return of investment is calculated as (benefit − cost)/cost. The reported values should therefore be used as guidance in comparing the scenarios, and the spreadsheet provided in the Supplemental Material can be used to modify the assumptions that were made in the present study.

Finally, there are other mitigation factors that can affect comparison of GBS and SNP arrays. For example, GBS-like approaches require higher quality of DNA than SNP arrays. Also, working with sequence data requires additional expertise beyond that for working with the SNP array data, although this hurdle is getting smaller every day due to the availability of efficient and user-friendly data pipelines (Glaubitz et al., 2014) and imputation programs such as the AlphaImpute, which was used in this study.

CONCLUSIONS

We have evaluated the potential of low-coverage GBS and imputation for genomic selection in biparental segregating populations. The results show that nonimputed 1*x* GBS data provides comparable prediction accuracy and bias and higher return on investment than the SNP array data, by our calculations. With imputation the sequencing coverage can be further reduced, even as low as 0.1*x* with 10K markers or 0.01*x* with 100K markers. Reduction of sequencing coverage and imputation can be leveraged to genotype larger training sets to increase prediction accuracy and larger prediction sets to increase selection intensity, which both allow for higher response to selection and higher return on investment.

Supplemental Material Available

Supplemental material for this article is available online.

Acknowledgments

The authors acknowledge the financial support from the BBSRC ISPG to The Roslin InstituteBB/J004235/1, from Genus PLC and from grant numbers BB/M009254/1, BB/L020726/1, BB/N004736/1, BB/N004728/1, BB/L020467/1, and BB/N006178/1, and from Medical Research Council (MRC) grant number MR/M000370/1. The authors thank Dr. Andrew Derrington (Scotland, UK) for assistance in refining the manuscript. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/).

References

Ali, O.A., S.M. O'Rourke, S.J. Amish, M.H. Meek, G. Luikart, C. Jeffres, and M.R. Miller. 2015. RAD Capture (Rapture): Flexible and efficient sequence-based genotyping. Genetics 202:389–400. doi:10.1534/genetics.115.183665

Altshuler, D., V.J. Pollara, C.R. Cowles, W.J. Van Etten, J. Baldwin, L. Linton, and E.S. Lander. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407:513–516. doi:10.1038/35035083

Antolin, R., C. Nettelblad, G. Gorjanc, D. Money, and J.M. Hickey. 2017. A hybrid method for the imputation of genomic data in livestock populations. Genet., Sel., Evol. 49:30. doi:10.1186/s12711-017-0300-y

Baird, N.A., P.D. Etter, T.S. Atwood, M.C. Currey, A.L. Shiver, Z.A. Lewis et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3(10):e3376. doi:10.1371/journal.pone.0003376

Bassi, F.M., A.R. Bentley, G. Charmet, R. Ortiz, and J. Crossa. 2016. Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). Plant Sci. 242:23–36. doi:10.1016/j.plantsci.2015.08.021

Beissinger, T.M., C.N. Hirsch, R.S. Sekhon, J.M. Foerster, J.M. Johnson, G. Muttoni et al. 2013. Marker density and read depth for genotyping populations using genotyping-by-sequencing. Genetics 193:1073–1081. doi:10.1534/genetics.112.147710

Bernardo, R., and J. Yu. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 47:1082–1090. doi:10.2135/cropsci2006.11.0690

Buerkle, C.A., and Z. Gompert. 2013. Population genomics based on low coverage sequencing: How low should we go? Mol. Ecol. 22:3028–3035. doi:10.1111/mec.12105

Calus, M.P.L., A.C. Bouwman, J.M. Hickey, R.F. Veerkamp, and H.A. Mulder. 2014. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: A review of livestock applications. Animal 8:1743–1753. doi:10.1017/S1751731114001803

Chaney, L., A.R. Sharp, C.R. Evans, and J.A. Udall. 2016. Genome mapping in plant comparative genomics. Trends Plant Sci. 21:770–780. doi:10.1016/j.tplants.2016.05.004

Chen, G.K., P. Marjoram, and J.D. Wall. 2009. Fast and flexible simulation of DNA sequence data. Genome Res. 19:136–142. doi:10.1101/gr.083634.108

Clark, S.A., J.M. Hickey, H.D. Daetwyler, and J.H. van der Werf. 2012. 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(1):4. doi:10.1186/1297-9686-44-4

Craig, D.W., J.V. Pearson, S. Szelinger, A. Sekar, M. Redman, J.J. Corneveaux et al. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. Nat. Methods 5:887–893. doi:10.1038/nmeth.1251

Crossa, J., Y. Beyene, S. Kassa, P. Pérez, J.M. Hickey, C. Chen et al. 2013. Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3 (Bethesda) 3:1903–1926. doi:10.1534/g3.113.008227

Davey, J.W., P.A. Hohenlohe, P.D. Etter, J.Q. Boone, J.M. Catchen, and M.L. Blaxter. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12:499–510. doi:10.1038/nrg3012

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6(5):e19379. doi:10.1371/journal.pone.0019379

Escalona, M., S. Rocha, and D. Posada. 2016. A comparison of tools for the simulation of genomic next-generation sequencing data. Nat. Rev. Genet. 17:459–469. doi:10.1038/nrg.2016.57

Faux, A.-M., G. Gorjanc, R.C. Gaynor, S.M. Edwards, D. Wilson, S.J. Hearne et al. 2016. AlphaSim: Software for breeding program simulation. Plant Genome 9(3):1–14. doi:10.3835/plantgenome2016.02.0013

Fu, Y.-B., G.W. Peterson, and Y. Dong. 2016. Increasing genome sampling and improving SNP genotyping for genotyping-by-sequencing with new combinations of restriction enzymes. G3 (Bethesda) 6:845–856. doi:10.1534/g3.115.025775

Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, and E.S. Buckler. 2014. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS One 9(2):e90346. doi:10.1371/journal.pone.0090346

Gorjanc, G., M. Battagin, J.-F. Dumasy, R. Antolin, R.C. Gaynor, and J.M. Hickey. 2016. Prospects for cost-effective genomic selection via accurate within-family imputation. Crop Sci. 57:216–228. doi:10.2135/cropsci2016.06.0526

Gorjanc, G., M.A. Cleveland, R.D. Houston, and J.M. Hickey. 2015. Potential of genotyping-by-sequencing for genomic selection in livestock populations. Genet. Sel. Evol. 47(1):12. doi:10.1186/s12711-015-0102-z

Heslot, N., J. Rutkoski, J. Poland, J.-L. Jannink, and M.E. Sorrells. 2013. Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. PLoS One 8(9):e74612. doi:10.1371/journal.pone.0074612

Hickey, J.M., J. Crossa, R. Babu, and G. de los Campos. 2012a. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52:654–663. doi:10.2135/cropsci2011.07.0358

Hickey, J.M., S. Dreisigacker, J. Crossa, S. Hearne, R. Babu, B.M. Prasanna et al. 2014. Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Sci. 54:1476–1488. doi:10.2135/cropsci2013.03.0195

Hickey, J.M., G. Gorjanc, R.K. Varshney, and C. Nettelbladd. 2015. Imputation of single nucleotide polymorphism genotypes in biparental, backcross, and topcross populations with a hidden Markov model. Crop Sci. 55:1934–1946. doi:10.2135/cropsci2014.09.0648

Hickey, J.M., B.P. Kinghorn, B. Tier, J.H. van der Werf, and M.A. Cleveland. 2012b. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet. Sel. Evol. 44(9):11.

Hoerl, A.E., and R.W. Kennard. 1976. Ridge regression iterative estimation of the biasing parameter. Commun. Stat. Theory Methods 5:77–88. doi:10.1080/03610927608827333

Huang, B.E., C. Raghavan, R. Mauleon, K.W. Broman, and H. Leung. 2014. Efficient imputation of missing markers in low-coverage genotyping-by-sequencing data from multiparental crosses. Genetics 197:401–404. doi:10.1534/genetics.113.158014

Illumina. 2014. Sequence-based genotyping brings agrigenomics to a crossroads. Illumina. http://www.illumina.com/content/dam/illumina-marketing/documents/products/appspotlights/app_spotlight_ngg_ag.pdf (accessed 6 July 2016).

Islam, M.S., G.N. Thyssen, J.N. Jenkins, and D.D. Fang. 2015. Detection, validation, and application of genotyping-by-sequencing based single nucleotide polymorphisms in upland cotton. Plant Genome 8(1):1–10. doi:10.3835/plantgenome2014.07.0034

Jacobson, A., L. Lian, S. Zhong, and R. Bernardo. 2015. Marker imputation before genomewide selection in biparental maize populations. Plant Genome 8(2):1–9. doi:10.3835/plantgenome2014.10.0078

Li, Y., C. Sidore, H.M. Kang, M. Boehnke, and G.R. Abecasis. 2011. Low-coverage sequencing: Implications for design of complex trait association studies. Genome Res. 21:940–951. doi:10.1101/gr.117259.110

Li, Y., C.J. Willer, J. Ding, P. Scheet, and G.R. Abecasis. 2010. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 34:816–834. doi:10.1002/gepi.20533

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Michael, T.P., and R. VanBuren. 2015. Progress, challenges and the future of crop genomes. Curr. Opin. Plant Biol. 24:71–81. doi:10.1016/j.pbi.2015.02.002

Michel, S., C. Ametz, H. Gungor, D. Epure, H. Grausgruber, F. Löschenberger, and H. Buerstmayr. 2016. Genomic selection across multiple breeding cycles in applied bread wheat breeding. Theor. Appl. Genet. 129:1179–1189. doi:10.1007/s00122-016-2694-2

Mrode, R.A., editor. 2005. Linear models for the prediction of animal breeding values. 2nd ed. CABI, Wallingford, UK, Cambridge, MA. doi:10.1079/9780851990002.0000

Pasaniuc, B., N. Rohland, P.J. McLaren, K. Garimella, N. Zaitlen, H. Li et al. 2012. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat. Genet. 44:631–635. doi:10.1038/ng.2283

Poland, J.A., P.J. Brown, M.E. Sorrells, and J.-L. Jannink. 2012a. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7(2):e32253. doi:10.1371/journal.pone.0032253

Poland, J.A., J. Endelman, J. Rutkoski, S. Wu, Y. Manes, S. Dreisigacker et al. 2012b. Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome 5:103–113. doi:10.3835/plantgenome2012.06.0006

Poland, J.A., and T.W. Rife. 2012. Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5:92–102. doi:10.3835/plantgenome2012.05.0005

Pszczola, M., and M.P.L. Calus. 2016. Updating the reference population to achieve constant genomic prediction reliability across generations. Animal 10:1018–1024. doi:10.1017/S1751731115002785

Pszczola, M., T. Strabel, H.A. Mulder, and M.P.L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95:389–400. doi:10.3168/jds.2011-4338

R Development Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Riedelsheimer, C., and A.E. Melchinger. 2013. Optimizing the allocation of resources for genomic selection in one breeding cycle. Theor. Appl. Genet. 126:2835–2848. doi:10.1007/s00122-013-2175-9

Rowan, B.A., V. Patel, D. Weigel, and K. Schneeberger. 2015. Rapid and inexpensive whole-genome genotyping-by-sequencing for crossover localization and fine-scale genetic mapping. G3 (Bethesda) 5:385–398. doi:10.1534/g3.114.016501

Rutkoski, J.E., J. Poland, J.-L. Jannink, and M.E. Sorrells. 2013. Imputation of unordered markers and the impact on genomic selection accuracy. G3 (Bethesda) 3:427–439. doi:10.1534/g3.112.005363

Schröder, S., S. Mamidi, R. Lee, M.R. McKain, P.E. McClean, and J.M. Osorno. 2016. Optimization of genotyping by sequencing (GBS) data in common bean (*Phaseolus vulgaris* L.). Mol. Breed. 36:6. doi:10.1007/s11032-015-0431-1

Staňková, H., A.R. Hastie, S. Chan, J. Vrána, Z. Tulpová, M. Kubaláková et al. 2016. BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. Plant Biotechnol. J. 14:1523–1531. doi:10.1111/pbi.12513

Swarts, K., H. Li, J.A. Romero Navarro, D. An, M.C. Romay, S. Hearne et al. 2014. Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Genome 7(3):1–12. doi:10.3835/plantgenome2014.05.0023

VanBuren, R., D. Bryant, P.P. Edger, H. Tang, D. Burgess, D. Challabathula et al. 2015. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. Nature 527:508–511. doi:10.1038/nature15714

VanRaden, P.M., and G.R. Wiggans. 1991. Derivation, calculation, and use of national animal model information. J. Dairy Sci. 74:2737–2746. doi:10.3168/jds.S0022-0302(91)78453-1

Whittaker, J.C., C.S. Haley, and R. Thompson. 1997. Optimal weighting of information in marker-assisted selection. Genet. Res. 69:137–144. doi:10.1017/S0016672397002711

Windhausen, V.S., G.N. Atlin, J.M. Hickey, J. Crossa, J.-L. Jannink, M.E. Sorrells et al. 2012. Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 (Bethesda) 2:1427–1436. doi:10.1534/g3.112.003699

Woolliams, J.A., P. Bijma, and B. Villanueva. 1999. Expected genetic contributions and their impact on gene flow and genetic gain. Genetics 153:1009–1020.

Zhang, X., P. Pérez-Rodríguez, K. Semagn, Y. Beyene, R. Babu, M.A. López-Cruz et al. 2015. Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. Heredity 114:291–299. doi:10.1038/hdy.2014.99