

RESEARCH

Open Access

# Identity-by-descent genomic selection using selective and sparse genotyping

Jørgen Ødegård<sup>1\*</sup> and Theo HE Meuwissen<sup>2</sup>

## Abstract

**Background:** Genomic selection methods require dense and widespread genotyping data, posing a particular challenge if both sexes are subject to intense selection (e.g., aquaculture species). This study focuses on alternative low-cost genomic selection methods (IBD-GS) that use selective genotyping with sparse marker panels to estimate identity-by-descent relationships through linkage analysis. Our aim was to evaluate the potential of these methods in selection programs for continuous traits measured on sibs of selection candidates in a typical aquaculture breeding population.

**Methods:** Phenotypic and genomic data were generated by stochastic simulation, assuming low to moderate heritabilities (0.10 to 0.30) for a Gaussian trait measured on sibs of the selection candidates in a typical aquaculture breeding population that consisted of 100 families (100 training animals and 20 selection candidates per family). Low-density marker genotype data (~ 40 markers per Morgan) were used to trace genomic identity-by-descent relationships. Genotyping was restricted to selection candidates from 30 phenotypically top-ranking families and varying fractions of their phenotypically extreme training sibs. All phenotypes were included in the genetic analyses. Classical pedigree-based and IBD-GS models were compared based on realized genetic gain over one generation of selection.

**Results:** Genetic gain increased substantially (13 to 32%) with IBD-GS compared to classical selection and was greatest with higher heritability. Most of the extra gain from IBD-GS was obtained already by genotyping the 5% phenotypically most extreme sibs within the pre-selected families. Additional genotyping further increased genetic gains, but these were small when going from genotyping 20% of the extremes to all phenotyped sibs. The success of IBD-GS with sparse and selective genotyping can be explained by the fact that within-family haplotype blocks are accurately traced even with low-marker densities and that most of the within-family variance for normally distributed traits is captured by a small proportion of the phenotypically extreme sibs.

**Conclusions:** IBD-GS was substantially more effective than classical selection, even when based on very few markers and combined with selective genotyping of small fractions of the population. The study shows that low-cost GS programs can be successful by combining sparse and selective genotyping with pedigree and linkage information.

## Background

Recently, genomic selection (GS) [1] has led to a paradigm shift in quantitative genetic analyses and selective breeding programs. GS was originally developed on the basis that all selection candidates and training animals are individually and densely genotyped. However, implementation of GS has been hampered in many species by the cost of genotyping large numbers of animals. For

species like dairy cattle, for which traditional selection schemes focus mainly on the male candidates and necessarily involve progeny testing, GS can often be implemented more cost-effectively by genotyping a limited number of male selection candidates and reducing the need for progeny testing. In species with a higher female fecundity, selective breeding focuses on both sexes. Aquaculture breeding programs represent the most extreme of these, in which both males and females have an extremely high fecundity. Moreover, since some of the most important aquaculture species are single spawners

\* Correspondence: jorgen.odegard@nmbu.no

<sup>1</sup>AquaGen AS, P.O. Box 1240, Sluppen, NO-7462 Trondheim, Norway  
Full list of author information is available at the end of the article

under certain conditions (e.g., Atlantic and Pacific salmon), progeny testing is uncommon and organized breeding programs are often based on sib-testing (for invasive traits). Such breeding populations are usually large, with many training animals and numerous selection candidates of both sexes. Genotyping costs are thus the main obstacle for the successful large-scale implementation of GS in aquaculture species. Such costs can be reduced by: (1) genotyping fewer samples (either fewer animals or pooled samples) and (2) genotyping fewer marker loci per sample (e.g., genotyping by sequencing regions of the genome). However, the challenge is to find a balance between genotyping costs and accuracy of selection. Pooled genotyping will be investigated in a future study and will not be further discussed here. The current study focuses on the use of combined selective and sparse genotyping in GS, which can greatly reduce genotyping costs, but the potential advantage of GS over classical pedigree-based selection in these conditions remains uncertain.

The methods suggested by Meuwissen et al. [1] imply that identity-by-state (IBS) information from all available marker loci is used, i.e., a given marker allele is assumed to have the same statistical effect on phenotype across different familial backgrounds. Thus, all marker alleles that are IBS are assumed to have the same effect, which can be explained by the existence of linkage disequilibrium (LD) between the marker and one or more quantitative trait loci (QTL). However, Luan et al. [2] showed small differences in accuracies of genetic evaluations of dairy bulls when using genomic identity-by-descent (IBD) relationships within the known pedigree (traced by markers) instead of the IBS relationships suggested by Meuwissen et al. [1]. In the following, these two methods of GS will be termed IBD-GS and IBS-GS, respectively. Assuming base animals are unrelated, IBD-GS uses only information from known, and thus close relatives, which typically share long chromosomal segments IBD. In comparison, the (LD-based) haplotype blocks found across a population are much shorter [3], both for domesticated breeds and, even more so, for large wild populations. Tracing large IBD-blocks within a known pedigree does not require dense markers, which makes IBD-GS less sensitive to marker density, compared to the more widely used IBS-GS.

Although IBS-GS does not use the pedigree explicitly (and thus IBD information), it does use the available population structure, i.e., IBD relationships necessarily imply sharing of marker alleles (but not vice versa). Hence, in a domesticated population under artificial selection, differences in IBS relationships between individuals are largely explained by close relationships, rather than by relationships due to common ancestors prior to the known base population. For IBS-GS, it has also been

shown that relationships with reference individuals have a much higher effect on accuracy of resulting genomic breeding value predictions than LD *per se* [4]. Furthermore, IBS-GS in pedigreed populations can be used to predict genomic breeding values even in the absence of LD, which shows that markers capture close relationships among genotyped animals, and thereby affect accuracy of predictions [5]. IBD-GS is expected to use this information more accurately but at the cost of not using LD that may exist between markers and QTL.

Use of the IBD-GS method has the potential to reduce marker density, while maintaining the efficiency of GS and taking the entire heritable variance into account. Yet, the number of animals to be genotyped can lead to a significant cost. Thus, in this study, our aim was to investigate whether IBD-GS can be effectively combined with selective genotyping of selection candidates and training animals. From this perspective, the advantage of IBD-GS over IBS-GS is that the accuracy of predicted breeding values depends on the number of genotyped and phenotyped known relatives, rather than on genotyping of the whole population. Hence, the IBD-GS method may provide accurate genomic EBV (estimated breeding values) for specifically targeted families by preferentially genotyping high-ranking families (based on phenotypic/pedigree-based pre-selection). Furthermore, the genotyping of training animals could focus on the individuals that are most informative for prediction of Mendelian deviations from the mid-parent means, which are the phenotypically most extreme sibs (in both directions); the intermediate sibs mainly provide information about mid-parent means (which are easily estimated with classical methods), and are thus less informative for prediction of Mendelian deviations from the parental means.

The aim of the study was to quantify whether a low-cost IBD-GS scheme could provide a significant increase in genetic gain for traits evaluated on sibs of selection candidates, compared to classical selection schemes, as applied in aquaculture breeding populations. The low-cost approach combines sparse marker panels (or genotyping by sequencing) with selective genotyping of a subsample of the selection candidates and the phenotypically most extreme training sibs of these candidates.

## Methods

### Simulation

The QMSim software [6] was used to simulate all datasets. All scenarios were replicated 50 times.

### Genome

The genome was assumed to consist of 20 chromosomes, each 100 cM long, with 1200 potential marker loci and 80 potential QTL. Mutation rate for both markers and QTL

was set to  $3.0 \times 10^{-5}$ . To simplify the simulation and statistical analysis, only recurrent allele mutations were allowed, and therefore only bi-allelic loci (similar to single nucleotide polymorphisms or SNPs) existed. Loci (both marker and QTL) with minor allele frequencies (MAF) below 5% (in the founder population of the pedigree) were deleted. The QTL effects were sampled from a gamma distribution with shape parameter 0.4.

### Population

A base population with effective population size  $N_e = 500$  was simulated for 5000 generations in order to achieve mutation-drift balance. At generation 5000, ~4000 marker loci (~200 per chromosome) and ~240 to 280 QTL segregated with a MAF above 5%. From generation 5000, 100 males and 100 females were randomly selected and each male was randomly mated with a single female, resulting in 100 families, each consisting of 120 full-sibs (12 000 individuals per generation). The same population structure was used to simulate the three subsequent generations (5001 to 5003). Phenotypes and genotypes (for a fraction of the animals) were stored only for generations 5001 to 5003, which were used in the subsequent genetic analysis. No selection was applied within these generations.

### Data structure

Breeding values were defined as the sum of all QTL effects for each individual. The QTL effects were rescaled such that the total additive genetic variance (variance of breeding values summed over all QTL) was equal to 0.30 for all replicates (in generation 5001). Standardized random residuals for phenotypes were sampled independently from a standard normal distribution and subsequently scaled by the appropriate environmental standard deviation. Phenotypes were defined as the sum of the breeding values and the scaled random residuals. Two scenarios were simulated, one with heritability  $h^2 = 0.10$  and one with heritability  $h^2 = 0.30$ . Datasets with different heritabilities were generated from the same simulation datasets by changing the residual standard deviation used for scaling of the standardized residuals. Hence, breeding values and genetic variance were identical in the parallel replicates across scenarios with different heritabilities.

Of the 120 sibs in each family, 100 sibs were randomly selected as training animals (phenotyped and were not available for selection) and the remaining 20 were chosen as selection candidates (non-phenotyped validation animals). Descriptive statistics of the data sets are given in Table 1.

### Genotyping

Genotyping was performed by sampling every 5<sup>th</sup> marker (the first and last marker of each chromosome were

**Table 1 Descriptive statistics of the simulation schemes**

Number of chromosomes	20
Length per chromosome (Morgan)	1.0
Base population	
Number of generations	5000
Mutation rate, markers	$3.0 \times 10^{-5}$
Mutation rate, QTL	$3.0 \times 10^{-5}$
Effective population size	500
Evaluated population	
Number of generations	3
Genotyped markers per chromosome*	~40
Number of segregating QTL per chromosome*	~240-280
Genetic variance*	0.3
Residual variance	2.7 or 0.7
Heritability*	0.1 or 0.3
Training animals per generation	10,000
Selection candidates per generation	2000
Sires per generation	100
Dams per generation	100
Families per generation	100
Selection candidates per family	20
Training animals per family	100

\*From base population generation 5000 (base population in the statistical analyses).

always retained) from the complete dataset, i.e., ~40 markers/Morgan (M). These markers were then used to trace genomic relationships within the known pedigree, as described in the next section (considering parents from generation 5000 as base animals). To mimic a situation with selective genotyping, marker genotypes were stored only from parents (generations 5001 and 5002) and individuals of the phenotypically best (based on the average phenotype of the 100 training sibs) 30 of 100 families in the last generation (generation 5003). It was considered unlikely that candidates from the remaining 70 families would be selected (even with marker data, which was confirmed by later results). Furthermore, within the best 30 families, marker genotypes were stored for all selection candidates (20 per family) and for varying fractions of the phenotyped training sibs (the top and bottom 5 to 20%, or all sibs). Thus, the total number of genotyped individuals ranged from 900 to 3600 out of 12 000 individuals (per generation). The different genotyping strategies are described in Table 2.

### Tracing relationships

Markers were traced through the pedigree by linkage analysis using the linkage disequilibrium multi-locus iterative peeling (LDMIP) method [7] based on information from all genotyped marker loci and the entire

**Table 2 Genotyping strategies for the different breeding program scenarios with classical pedigree-based selection (PED) and IBD-GS**

	PED	IBD5	IBD10	IBD20	IBDall
Number of families	100	100	100	100	100
Selection candidates per family	20	20	20	20	20
Number of pre-selected families	0	30	30	30	30
Genotyped selection candidates	0	600	600	600	600
Genotyped training animals per family	0	5+5	10+10	20+20	100
Total number genotyped	0	900	1200	1800	3600
Number of selected new parents	200	200	200	200	200
Minimum number of parental families	20	-	-	-	-

Only generation 5003 is considered.

pedigree (using parents from generation 5000 as base population). The output was used to calculate IBD probabilities at each genotyped marker locus [8] and the genome-wide IBD relationship matrix was produced by averaging over all loci. The IBD matrix also included IBD probabilities of non-genotyped animals, using information from relatives to estimate their genotypes [7]. Non-genotyped non-parents were consequently assumed to have standard pedigree relationships with all their sibs, but not necessarily with more distant relatives (as their relationships are influenced by the relationships among genotyped ancestors). Animals not related through the pedigree were considered unrelated, regardless of their marker genotypes.

#### Statistical models

Two statistical models, pedigree-based (PED) and IBD-GS, were applied to all datasets and scenarios to estimate BLUP (Best Linear Unbiased Prediction) breeding values with the DMU software package [9]. Both models had the following general characteristics:

$$\mathbf{y} = 1\mu + \mathbf{Z}\mathbf{a} + \mathbf{e}$$

where  $\mathbf{y}$  is a vector of all phenotypes,  $\mu$  is the overall mean,  $\mathbf{a}$  is a vector of additive genetic breeding values of all animals included in the pedigree,  $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$  is a vector of random residuals,  $\sigma_e^2$  is the residual variance and  $\mathbf{Z}$  is an appropriate incidence matrix. The two models only differed in their distributional assumptions for the additive breeding values:

$$\text{PED} : \mathbf{a} \sim N(\mathbf{0}, \mathbf{A}\sigma_a^2),$$

$$\text{IBD-GS} : \mathbf{a} \sim N(\mathbf{0}, \mathbf{G}_{\text{IBD}}\sigma_a^2),$$

where  $\sigma_a^2$  is the additive genetic variance,  $\mathbf{A}$  is the pedigree-based numerator relationship matrix, and  $\mathbf{G}_{\text{IBD}}$  is the

IBD-based genomic relationship matrix, calculated as described above. The true variance components were assumed known.

#### Validation

To evaluate the IBD-GS method, it was compared to classical pedigree-based selection methods. Animals were ranked based on their predicted breeding values from the PED and IBD-GS models, respectively. Predicted breeding values (either PED or IBD-GS) were used to select 200 parents for the next generation. Genetic gain in the first generation is expected to equal the average genetic level of the parents deviated from the population average in the previous generation. Hence, efficiency of classical pedigree-based selection vs. IBD-GS using selective genotyping was assessed by comparing the average true breeding values of the selected parents (from generation 5003) for the two models. Since none of the selection candidates had their own phenotype available (only sibs with data) classical selection based on PED implied family selection, i.e., all 200 selected individuals would originate from only 10 families (with 20 candidates each) unless restrictions are imposed. Hence, for classical selection, the number of selected offspring was restricted to 10 per family, and future parents were thus randomly sampled from the 20 top-ranking families. GS allowed individual selection also for non-phenotyped (albeit genotyped) animals. In practice, one would then select the best animals based on EBV across families. The lowest-ranking sibs within the best families are probably outperformed by the highest-ranking sibs among the second-best families, which implies that future parents are selected from a wider range of familial backgrounds. Hence, for GS, there was no need to put any restrictions on selection. Results showed that future parents were indeed selected from nearly all 30 pre-selected families.

Selection based on the two models was compared based on the expected rate of inbreeding, which was approximated as [10]:

$$\Delta F \approx \frac{1}{2} \sum_{i=1}^k c_i^2,$$

where  $c_i$  is the relative genetic contribution of parent  $i$  (appearing as parent for generation  $t$ ) to generation  $t+1$ , and  $k = 200$  (total number of parents per generation).

The two models were also assessed for bias by regression of true breeding values on predicted breeding values. A regression coefficient equal to 1 indicates proper scaling,  $< 1$  indicates inflation and  $> 1$  indicates deflation of the predictions.

## Results

IBD-GS led to a substantial increase in genetic gain over one generation, compared to classical selection (Figure 1), especially with higher heritability. For a moderately heritable trait ( $h^2 = 0.3$ ), the increase in genetic gain with IBD-GS compared to classical pedigree-based selection was substantial, with relative increases of 23, 28, 31 and 32% by genotyping the phenotypic most extreme (top/bottom) 5, 10, 20%, and all training sibs, respectively. Hence, most of the extra gain with IBD-GS was already obtained by genotyping the top and bottom 5% training sibs for the 30 pre-selected families. For a low heritable trait ( $h^2 = 0.1$ ), the extra gain with IBD-GS was lower (13 to 21%) but, again, most of the extra gain was achieved with genotyping of the top and bottom 5% of training sibs of the pre-selected families.

Although restrictions with respect to inbreeding were imposed only for the classical selection scheme, the unrestricted IBD-GS scheme had lower average rates of inbreeding (1.04 to 1.13% vs. 1.25%) and new parents were recruited from more (26 to 29 vs. 20) of the 30 pre-selected families (Table 3).

The average regression coefficients across replicates (standard deviations) of true breeding values on predicted breeding values are in Table 4. Across replicates, the pedigree-based model had an average regression coefficient close to the expectation (1.0) and was thus unbiased. IBD-GS was also unbiased when all members of preselected favorable families were genotyped (IBDall). However, selective within-family genotyping of phenotypically extreme sibs resulted in slight, albeit significant bias at the highest heritability, in terms of inflated variance of genomic EBV (Table 4). Nevertheless, when looking at single replicates, the pedigree-based model

frequently deviated more from the expected regression coefficient than the IBD-GS model.

## Discussion

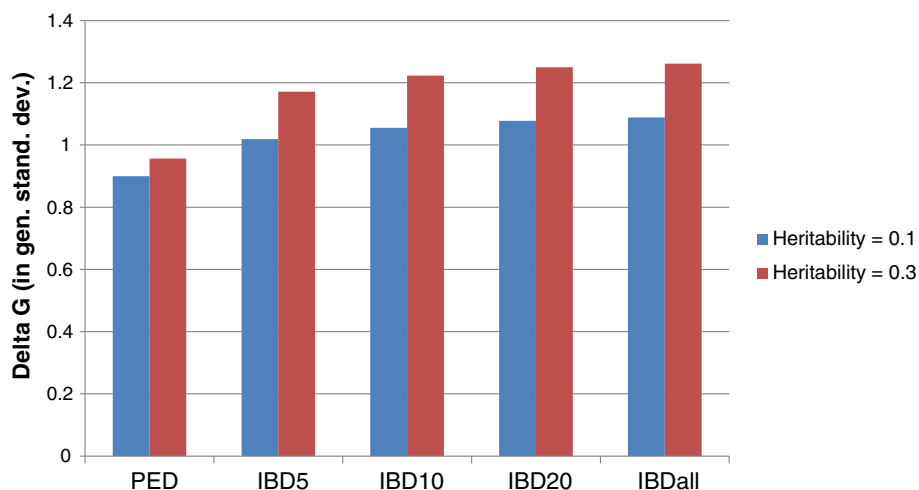
### Genetic gain

The study shows that for traits that are evaluated on sibs of selection candidates, substantial increases in genetic gain can be achieved through IBD-GS using sparse marker genotyping on small sub-samples of carefully selected individuals, e.g., 900 to 1800 animals per generation from a population of 12 000 individuals. Using this approach, only the best 30% of families were genotyped (for both selection candidates and training animals). In most cases, no candidates were selected from the lowest ranking preselected family. Hence, genotyping more families would probably not change selection decisions for this population structure.

Most of the potential advantage of IBD-GS relative to classical selection was already obtained when the 5% top and bottom training sibs from the pre-selected families were genotyped. Furthermore, there was little practical difference in response to selection between genotyping the 20% top and bottom vs. all training sibs. However, when analyzing real data, extreme observations may be artifacts and selective genotyping strategies using small and highly selected samples are likely to be more vulnerable to such errors. Hence, careful quality control of the data is particularly important when combined with selective genotyping. Genotyping larger fractions (i.e. 10 to 20%) would increase robustness of the IBD-GS analysis.

### Population structure and family production

In this study, both models required pedigree information. For aquaculture species for which tagging is impossible at



**Figure 1 Genetic gain in GS-IBD and PED selection schemes.** Genetic gain in genetic standard deviation units for classical pedigree-based selection schemes (PED) and GS-IBD using sparse and selective genotyping for different numbers of genotyped phenotypically extreme sibs (IBD# = genotyping the # phenotypically best and worst training sibs of each family, IBDall = genotyping all training sibs).

**Table 3 Average number of contributing families and estimated rate of inbreeding for classical pedigree-based selection (PED) and IBD-GS**

Heritability	0.10		0.30	
	Number of families	$\Delta F$ (%)	Number of families	$\Delta F$ (%)
PED	20.0	1.25	20.0	1.25
IBD5	26.1	1.13	28.6	1.07
IBD10	27.0	1.12	28.9	1.06
IBD20	27.4	1.11	29.2	1.04
IBDall	27.6	1.11	29.1	1.05

young ages, it may be necessary to separate families until they reach tagging size, which increases the costs of family production and creates physical limitations in the number of families that can be produced, and potentially introduces common environmental family (tank) effects. Still, genome-wide SNP markers (as in IBD-GS) can be used to trace parentage of individuals with high accuracy, even with communal rearing e.g. [11,12]. However, when within-family selective genotyping is to be used, prior knowledge of the pedigree is required, i.e., favoring separate rearing of families. Fortunately, separate rearing is already common practice in many existing aquaculture selection programs for traits that are evaluated on sibs of selection candidates (e.g., in *Salmonidae*), and thus the structure necessary to apply selective genotyping exists. In contrast, for mass-spawning species, for which artificial stripping is difficult to apply, e.g., gilthead seabream (*Sparus aurata*), communal rearing may be the only practical alternative. Furthermore, mass-spawning species typically have rather chaotic family structures (many small families, and uneven contributions of parents), which makes within-family selective genotyping difficult to apply, even when family background can be deduced (e.g., through an initial parentage assignment based on microsatellites). Alternatively, parentage testing and sparse genotyping for IBD-GS may be combined into one genotyping test but then additional saving from selective genotyping is not being used.

**Table 4 Average ( $\beta$ ) and standard deviation (SD) across replicates of regression coefficients of true breeding values on predicted breeding values for classical pedigree-based selection (PED) and IBD-GS**

Heritability	0.10		0.30	
	$\beta$	SD	$\beta$	SD
PED	0.963	0.198	0.991	0.126
IBD5	0.964	0.155	0.961*	0.088
IBD10	0.965*	0.145	0.947***	0.085
IBD20	0.971	0.145	0.948***	0.080
IBDall	0.993	0.146	1.020	0.086

\* $P < 0.05$  for  $\beta < 1$ ; \*\* $P < 0.01$  for  $\beta < 1$ ; \*\*\* $P < 0.001$  for  $\beta < 1$ .

### Multi-trait selection

This study considered selection on a single trait but real selection programs are often multi-trait and can include more than one trait measured on sibs of selection candidates. In such cases, selection of the families to be genotyped would be based on a total merit index. In some cases, traits measured on sibs are recorded on different subsets of training sibs. This would still allow the proposed strategy for within-family selective genotyping to be used for each trait separately, but would require genotyping more animals. If multiple traits are recorded on the same group of training sibs, sibs for genotyping could be selected based on their total merit index. These animals are not necessarily the most phenotypic extremes for each trait but they are expected to be the most extreme with respect to the aggregated genotype, which is what we want to predict and genetically improve.

### Bias by selective genotyping

When genotyping is restricted to phenotypically extreme sibs, estimates of genetic variation may be inflated, resulting in biased EBV. However, results from a previous study indicate that when non-genotyped sibs are included in the analysis through pedigree information (as in the current study), accurate estimates of genetic variance can still be obtained for larger selectively genotyped samples [13]. In the current study, the true genetic variance was assumed known (i.e., BLUP), which reduces the risk of inflated variation of EBV. Still, a statistically significant but small bias in terms of inflated variance of EBV was detected with IBD-GS, but the limited magnitude of the inflation indicates that the bias would not be a problem in practice. The bias detected was of similar magnitude as previously reported for other GS approaches e.g. [14,15]. Furthermore, the pedigree-based model (which was unbiased when averaged over replicates) showed larger between-replicate variation. Hence, for real data, pedigree-based evaluation may actually give larger scaling errors than IBD-GS using selective genotyping.

### Inbreeding

Although there were no restrictions on inbreeding and genetic gain was faster, inbreeding rates were systematically lower with IBD-GS than with pedigree-based selection schemes. Here, the expected rate of inbreeding was estimated over only one generation of selection. In the longer term, differences between the two methods are expected to be more substantial and in favor of IBD-GS because pedigree-based selection for traits measured on the sibs will favor co-selection of relatives. Furthermore, imposed recruitment of parents of less favorable EBV, as was done for the pedigree-based program, can reduce

inbreeding in the short term but the long term contribution of such parents may be low. In the IBD-GS scheme parents were solely selected based on EBV, which is expected to give more even long-term genetic contributions of the selected parents compared to a classical selection scheme imposing selection of genetically inferior parents.

### Genetic variability

It has been stated that GS is particularly useful for traits with low heritability [16], which is indeed likely to be true for traits that are recorded directly on selection candidates as well as on sibs. With classical selection for traits of high heritability, individual phenotypes will compensate for the lack of precision achieved with sib data. For a trait measured only on sibs, the extra genetic gain from IBD-GS over pedigree-based selection is more likely to increase with increasing heritability. This is because, with pedigree-based selection, the EBV of all selection candidates is set equal to their mid-parent EBV, regardless of heritability, while IBD-GS allows estimation of Mendelian deviations from this mean, even without individual phenotypes. In the current study, mid-parent means are expected to be accurately estimated in both models, even for traits with low heritability because of the large number of phenotyped offspring (100 per family), while accuracies of the estimated Mendelian sampling deviations from these means (predicted in IBD-GS) are expected to be greater with high heritability. As a consequence, the extra gain from IBD-GS increases with increasing heritability of the traits evaluated on sibs. The greater benefit of IBD-GS for traits with higher heritability is evident from our study; the classical pedigree-based model showed similar genetic gains (6% different) for traits with a low or moderate heritability, while for IBD-GS, genetic gain was substantially greater (15 to 16%) for traits with moderate versus low heritability.

This study considered genetic gain over only one generation, starting from a previously unselected population. In practice, selection is usually continued over several generations, and the Bulmer effect would (as long as selection continues) reduce between-family variation in the population [17]. As a consequence, the ability to use within-family (Mendelian) genetic variation will be more important for future generations (and more so if restrictions are imposed on inbreeding). Thus, because classical sib-selection does not use within-family variation the relative advantage of GS is expected to increase if selective breeding is applied over multiple generations.

### Sensitivity to assumptions

Genomic selection simulation studies are typically based on several idealized assumptions, including: (1) QTL effects are strictly additive; (2) all heritable variation is

explained by the additive QTL effects, (3) all alleles originate from a single base population, which implies consistent population-wide LD (no segmentation of the population), and (4) marker and QTL alleles follow similar frequency distributions. In practice, these assumptions do not necessarily apply: many populations have mixed origins, QTL and markers do not necessarily follow the same distributions (e.g., QTL loci are often subjected to selection) and non-additive (dominance and epistatic) genetic effects exist. Through sexual reproduction and recombination, non-additive effects are often not passed on to future generations, which reduces their relevance for selective breeding. However, epistatic effects between linked loci are expected to be more sustained over generations and evidence for epistatic interactions between linked loci has been reported for different species [18,19]. Within families, haplotypes of linked loci are likely inherited as one block, i.e., as if they were a single locus [20]. Similarly, transgenerational epigenetic inheritance [21], such as chromatin marking (including methylation), can also contribute to inherited differences and thus contribute to the heritable (“additive”) variation, i.e., the variance available for selective breeding. Such heritable factors, may partly explain the so-called “missing heritability problem” that is typically observed with IBS relationships [22]. These effects of local epistasis and transgenerational epigenetics (chromatin marking) may, however, be effectively captured by genomic IBD relationships. Thus, the IBD-GS model is expected to take such effects into account.

One of the main advantages of the IBD-GS method is that it can effectively use sparse marker genotypes and selective genotyping. In contrast, IBS-GS methods are intended for dense marker data [1] and sparse genotyping would require a sparse-to-dense imputation step (i.e., dense marker data must be available for preceding generations). Furthermore, IBD-GS is well suited for the joint analysis of data from non-genotyped and genotyped individuals (e.g., as in selective genotyping), which is likely to introduce bias in IBS-based analyses. The IBD-GS method is thus easier to implement for analysis of selectively genotyped and sparse marker data and of phenotypic data that do not necessarily fit the idealized conditions that are typically assumed in stochastic simulation studies.

### Conclusions

The IBD-GS was substantially more effective than classical selection, even when based on very few markers, and when combined with selective genotyping of small proportions of the population, i.e., 900 to 1800 out of 12 000 individuals. The study shows that a low-cost GS program can be successfully performed by combining

sparse and selective genotyping with pedigree and linkage information. Thus, the IBD-GS method is suitable for cost-effective genomic evaluation using sparse markers, for selective genotyping of training animals, and for phenotypic data that do not necessarily fit the idealized conditions that are typically assumed in stochastic simulation studies.

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

JØ had the original idea, performed the simulations and statistical analyses and was main responsible for writing the manuscript and THEM was responsible for developing the linkage analysis methodology (LDMIP software). All authors read and approved the final manuscript.

#### Acknowledgements

This study was partly financed by the Research Council of Norway through projects no. 186862 and 225181. Dr. Nina Santi, AquaGen is acknowledged for valuable scientific input to the study. Two anonymous reviewers are gratefully acknowledged for their valuable comments.

#### Author details

<sup>1</sup>AquaGen AS, P.O. Box 1240, Sluppen, NO-7462 Trondheim, Norway.

<sup>2</sup>Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432 Ås, Norway.

Received: 15 May 2013 Accepted: 3 December 2013

Published: 20 January 2014

#### References

1. Meuwissen THE, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157**:1819–1829.
2. Luan T, Woolliams JA, Ødegård J, Dolezal M, Roman-Ponze SI, Bagnato A, Meuwissen THE: **The importance of identity-by-state information for the accuracy of genomic selection.** *Genet Sel Evol* 2012, **44**:28.
3. Habier D, Fernando RL, Dekkers JCM: **Genomic selection using low-density marker panels.** *Genetics* 2009, **182**:343–353.
4. Wientjes YCJ, Veerkamp RF, Calus MPL: **The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction.** *Genetics* 2013, **193**:621–631.
5. Habier D, Fernando RL, Dekkers JCM: **The impact of genetic relationship information on genome-assisted breeding values.** *Genetics* 2007, **177**:2389–2397.
6. Sargolzaei M, Schenkel FS: **QMSim: a large-scale genome simulator for livestock.** *Bioinformatics* 2009, **25**:680–681.
7. Meuwissen THE, Goddard ME: **The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole-genome sequence density genotypic data.** *Genetics* 2010, **185**:1441–1449.
8. Fernando R, Grossman M: **Marker assisted selection using best linear unbiased prediction.** *Genet Sel Evol* 1989, **21**:467–477.
9. Madsen P, Jensen J: *DMU: a user's guide. A package for analysing multivariate mixed models. Version 6, release 5.0.* Tjele, Denmark: University of Aarhus, Faculty of Agricultural Sciences (DJF), Dept of Genetics and Biotechnology; 2010.
10. Robertson A: **Inbreeding in artificial selection programmes.** *Genet Res* 1961, **2**:189–194.
11. Hauser L, Baird M, Hillborn R, Seeb LW, Seeb JE: **An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (*Oncorhynchus nerka*) population.** *Mol Ecol Resour* 2011, **11**(Suppl 1):150–161.
12. Hayes B, Sonesson AK, Gjerde B: **Evaluation of three strategies using DNA markers for traceability in aquaculture species.** *Aquaculture* 2005, **250**:70–81.
13. Ødegård J, Meuwissen T: **Estimation of heritability from limited family data using genome-wide identity-by-descent sharing.** *Genet Sel Evol* 2012, **44**:16.
14. Vitezica ZG, Aguilar I, Misztal I, Legarra A: **Bias in genomic predictions for populations under selection.** *Genet Res* 2011, **93**:357–366.
15. Luan T, Woolliams JA, Lien S, Kent M, Svendsen M, Meuwissen THE: **The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation.** *Genetics* 2009, **183**:1119–1126.
16. Calus MPL, Meuwissen THE, de Roos APW, Veerkamp RF: **Accuracy of genomic selection using different methods to define haplotypes.** *Genetics* 2008, **178**:553–561.
17. Bulmer MG: **The effect of selection on genetic variability.** *Am Nat* 1971, **105**:201–211.
18. Gaertner BE, Parmenter MD, Rockman MV, Kruglyak L, Phillips PC: **More than the sum of its parts: a complex epistatic network underlies natural variation in thermal preference behavior in *Caenorhabditis elegans*.** *Genetics* 2012, **192**:1533–1542.
19. Duthie C, Simm G, Doeschl-Wilson A, Kalm E, Knap PW, Roehe R: **Epistatic analysis of carcass characteristics in pigs reveals genomic interactions between quantitative trait loci attributable to additive and dominance genetic effects.** *J Anim Sci* 2010, **88**:2219–2234.
20. Haig D: **Does heritability hide in epistasis between linked SNPs?** *Eur J Hum Genet* 2011, **19**:123–123.
21. Jablonka E, Raz G: **Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution.** *Q Rev Biol* 2009, **84**:131–176.
22. Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, Cornes BK, Montgomery GW, Martin NG: **Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings.** *PLoS Genet* 2006, **2**:e41.

doi:10.1186/1297-9686-46-3

**Cite this article as:** Ødegård and Meuwissen: Identity-by-descent genomic selection using selective and sparse genotyping. *Genetics Selection Evolution* 2014 **46**:3.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
www.biomedcentral.com/submit

