Heredity (2020) 124:37–49
https://doi.org/10.1038/s41437-019-0246-7

the geneticssociety

ARTICLE

# Improvement of genomic prediction by integrating additional single nucleotide polymorphisms selected from imputed whole genome sequencing data

Aoxing Liu[1,2] · Mogens Sandø Lund[1] · Didier Boichard[3] · Emre Karaman[1] · Sebastien Fritz[3] · Gert Pedersen Aamand[4] · Ulrik Sander Nielsen[5] · Yachun Wang[2] · Guosheng Su[1]

Received: 30 January 2019 / Revised: 11 May 2019 / Accepted: 17 June 2019 / Published online: 5 July 2019

## Abstract

The availability of whole genome sequencing (WGS) data enables the discovery of causative single nucleotide polymorphisms (SNPs) or SNPs in high linkage disequilibrium with causative SNPs. This study investigated effects of integrating SNPs selected from imputed WGS data into the data of 54K chip on genomic prediction in Danish Jersey. The WGS SNPs, mainly including peaks of quantitative trait loci, structure variants, regulatory regions of genes, and SNPs within genes with strong effects predicted with variant effect predictor, were selected in previous analyses for dairy breeds in Denmark–Finland–Sweden (DFS) and France (FRA). Animals genotyped with 54K chip, standard LD chip, and customized LD chip which covered selected WGS SNPs and SNPs in the standard LD chip, were imputed to 54K together with DFS and FRA SNPs. Genomic best linear unbiased prediction (GBLUP) and Bayesian four-distribution mixture models considering 54K and selected WGS SNPs as one (a one-component model) or two separate genetic components (a two-component model) were used to predict breeding values. For milk production traits and mastitis, both DFS (0.025) and FRA (0.029) sets of additional WGS SNPs improved reliabilities, and inclusions of all selected WGS SNPs generally achieved highest improvements of reliabilities (0.034). A Bayesian four-distribution model yielded higher reliabilities than a GBLUP model for milk and protein, but extra gains in reliabilities from using selected WGS SNPs were smaller for a Bayesian four-distribution model than a GBLUP model. Generally, no significant difference was observed between one-component and two-component models, except for using GBLUP models for milk.

## Introduction

Genomic prediction has been widely applied in dairy cattle breeding (Hayes et al. 2009). To achieve reliable prediction for breeding values of candidate animals, a reference population consisted of a large number of individuals with both phenotypes and genotypes is required (Karaman et al. 2016). Assembling such a sufficiently large reference population, however, may not be possible for traits that are hard to measure, such as feed intake (Berry et al. 2014), or for breeds that are numerically small, such as Danish Jersey (Lund et al. 2016).

To improve reliabilities of genomic prediction, especially for a numerically small breed, many approaches have been investigated (Brøndum et al. 2015; Lund et al. 2016; van den Berg et al. 2016b). One approach is to increase SNP density in order to construct stronger linkage disequilibrium between single nucleotide polymorphisms (SNPs) and causative mutations (de Roos et al. 2008). However,

**Supplementary information** The online version of this article (https://doi.org/10.1038/s41437-019-0246-7) contains supplementary material, which is available to authorized users.

✉ Yachun Wang
wangyachun@cau.edu.cn

✉ Guosheng Su
guosheng.su@mbg.au.dk

[1] Department of Molecular Biology and Genetics, Aarhus University, 8830 Tjele, Denmark

[2] Key Laboratory of Animal Genetics, Breeding and Reproduction, MARA; National Engineering Laboratory for Animal Breeding, College of Animal Science and Technology, China Agricultural University, 100193 Beijing, PR China

[3] GABI, INRA, AGROParisTech, Université Paris Saclay, 78350 Jouy-en-Josas, France

[4] Nordic Cattle Genetic Evaluation, 8200 Aarhus N, Denmark

[5] Seges, 8200 Aarhus N, Denmark

reliabilities from using the HD chip were only a little higher than those from using the 54K chip (Su et al. 2012a; VanRaden et al. 2013), while reliabilities from using imputed whole genome sequencing (WGS) data were quite similar to those from using the HD chip (van Binsbergen et al. 2015). One possible reason could be that only causative mutations or SNPs very close to causative mutations can improve reliabilities of genomic prediction (van den Berg et al. 2016a). Moreover, the inclusion of a large number of noncausative SNPs may bring only noise to genomic prediction (Pérez-Enciso et al. 2015).

Instead of using all WGS SNPs, an alternative approach is to integrate only causative mutations or SNPs in high linkage disequilibrium with causative mutations into the genotype data of the 54K chip in genomic prediction (Brøndum et al. 2015; van den Berg et al. 2016a). Although most causative mutations in cattle remain unknown, large numbers of causative SNPs or SNPs in high linkage disequilibrium with causative SNPs have been detected from WGS data via quantitative trait loci (QTL) mapping (Daetwyler et al. 2014; Mao et al. 2016) and bioinformatics analyses (Michot et al. 2016; Boussaha et al. 2015). Benefits of incorporating QTL SNPs selected from WGS data in genomic prediction have been widely verified (Brøndum et al. 2015; van den Berg et al. 2016a), but less has been done for WGS SNPs selected from bioinformatics analyses. A customized LD chip was designed under the project of EuroGenomics, covering SNPs in the standard LD chip (Boichard et al. 2012) and more importantly, some additional WGS SNPs selected from peaks of QTL or bioinformatics analyses (Boichard et al. 2018). The selected WGS SNPs in this customized LD chip could be promising in genomic prediction but has not yet been investigated previously.

Sharing the reference population with other populations is another cost-effective approach to improve reliabilities, especially for a numerically small breed with limited bulls to be used (Lund et al. 2011). The magnitudes of improvements on reliabilities depend on how well the information of causative mutations could be captured by using a joint reference population (Lund et al. 2014). A simulation study showed that directly using selected WGS SNPs close to causative mutations improved reliability in genomic prediction with a joint reference population, but reliability dropped quickly when using the WGS SNPs distant from causative mutations (van den Berg et al. 2016a). In real data, however, the option of using selected WGS SNPs has not yet been investigated in genomic prediction with a joint reference population. Another strategy that can increase the reference population size is to include cows in the reference. Although the phenotypic information of cows is less precise (more noise) than those of bulls with a large number of progenies, improvements of reliabilities can still be achieved since a large number of cows are available to be genotyped and included in the reference (Buch et al. 2012; Su et al. 2016). With the increase of reference population size, however, the improvement of reliabilities from using additional information from other sources (e.g., selected WGS SNPs) could be reduced (Daetwyler et al. 2008).

Different models in terms of different assumptions for SNP effects could influence the efficiency of using selected WGS SNPs in genomic prediction (Brøndum et al. 2015). The model with assumption closest to the true distribution of SNP effects can achieve the highest reliability. The Genomic best linear unbiased prediction (GBLUP) model implicitly assumes that SNP effects follow a normal distribution with a null mean and an equal variance, and therefore, a same amount of shrinkage is implicitly applied to all SNPs. The Bayesian mixture model such as Bayesian four-distribution mixture model (Erbe et al. 2012; Gao et al. 2013), however, assumes a prior distribution that SNP effects follow a mixture of four normal distributions with null means but varying variance parameters, resulting in different shrinkages for SNPs in different classes. Reliabilities from Bayesian mixture models are expected to be higher than GBLUP models for traits with major QTL, but similar to GBLUP models for traits affected by many QTL each with small effect. Furthermore, selected WGS SNPs can be considered as a genetic component together with standard chip SNPs (a one-component model) or as a separate genetic component (a two-component model) (Visscher et al. 2007; Brøndum et al. 2015). A two-component model will likely be more beneficial in a GBLUP model than in a Bayesian mixture model (Brøndum et al. 2015), since different variances between standard chip SNPs and selected WGS SNPs have already been allowed in a Bayesian mixture model but not in a GBLUP model.

Primary objectives of this study were to investigate the effects of selected WGS SNPs on genomic prediction in Danish Jersey. Besides, we investigated the effects of using selected WGS SNPs along with the increase of the reference population size by including US Jersey bulls and Danish Jersey cows. Furthermore, we assessed different models on their efficiency to use the information of selected WGS SNPs.

## Materials and methods

### Data

#### Genotype

A total of 3745 Danish Jersey bulls, 1168 US Jersey bulls, and 28,678 Danish Jersey cows were genotyped. The

Danish Jersey bulls were mainly genotyped with the Illumina Bovine SNP50 chip (54K, Illumina, Inc). The US Jersey bulls were genotyped with either the Illumina Bovine SNP50 chip or the GeneSeek Genomic Profiler HD chip (777K, GeneSeek, Neogen Corporation), but SNPs which were not in the Bovine SNP50 chip were excluded in this study. For Danish Jersey cows, 3% were genotyped with Illumina Bovine SNP50 chip, 49% with standard Bovine LD Chip (standard LD, Illumina, Inc.), and 48% with EuroGenomics customized Illumina Bovine LD chip (Boichard et al. 2018). The EuroGenomics customized LD chip included SNPs in the standard LD chip together with 1754 WGS SNPs selected by Denmark–Finland–Sweden (DFS) and 4325 WGS SNPs selected by France (FRA). The DFS SNPs were peaks of QTL detected from imputed WGS data in Nordic Holsteins, Nordic Red, and Danish Jersey, selecting SNPs within each breed according to $P$-values of a single-marker regression model, functional annotations and linkage disequilibrium between SNPs (Brøndum et al. 2015). The FRA SNPs were WGS SNPs selected from one of the following categories: (i) literature; (ii) within genes with strong effect predicted with variant effect predictor (McLaren et al. 2016) (e.g., frameshift, stop gain, splicing site, and nonsynonymous substitution with strong predicted effect); (iii) regulatory regions of genes; (iv) peaks of QTL; and (v) breakpoints of structural SNPs (Boichard et al. 2018). Both DFS and FRA SNPs were discovered for milk production and functional traits (e.g., fertility, mastitis, calving, growth, and longevity). The SNPs on the sex chromosome or with unknown positions, monomorphism, multiple alleles and minor allele frequency (MAF) lower than 0.01 were excluded.

Animals genotyped with 54K and different versions of LD chips were imputed to 54K + DFS + FRA by a two-step family and population-based approach using the FImpute software (Sargolzaei et al. 2014). Firstly, individuals genotyped with different versions of LD chips were imputed to 54K. Secondly, individuals with 54K or imputed 54K were imputed to 54K + DFS + FRA. The SNP-wise imputation accuracy was measured as the Pearson correlation between observed and imputed genotypes (coded as 0, 1, or 2), and the proportion of correctly imputed genotypes to all imputed genotypes (i.e., concordance rate). Only SNPs with both correlation and concordance rate higher than 0.8 were used in genomic prediction. Ultimately, 39,803 SNPs in the 54K chip, 1270 DFS SNPs and 2427 FRA SNPs were kept for genomic prediction, with 28 SNPs overlapped between DFS and FRA SNPs. The correlations of imputation for SNPs used in genomic prediction were 97.0% for standard LD chip to 54K, 96.9% for DFS SNPs and 95.9% for FRA SNPs; while concordance rates were 98.3% for standard LD chip to 54K, 98.1% for DFS SNPs and 97.8% for FRA SNPs.

### Phenotype

The analyzed traits included milk, protein, fat, mastitis, and fertility. Deregressed proof (DRP) derived from official EBV was used as the pseudo phenotype in genomic prediction. To avoid double counting of the information from the derivation of DRP and the prediction of breeding values, two sets of DRP were used in this study. One derived from all genotyped Danish and US Jersey bulls ($\mathrm{DRP_b}$) to be used for genomic prediction with bulls as the reference population and as the validation population; the other derived from all genotyped Danish and US Jersey bulls and Danish Jersey cows ($\mathrm{DRP_{bc}}$) to be used for genomic prediction with the reference populations including cows and cows as the validation population. Thus, DRP of bulls in $\mathrm{DRP_b}$ set could include the information of the genotyped cows, while those in $\mathrm{DRP_{bc}}$ set not. Reliability of DRP was calculated as $r^2_{\mathrm{DRP}} = \frac{ERC_i}{ERC_i+\lambda}$, where $\lambda = \frac{1-h^2}{h^2}$. The $ERC_i$ was the effective record contribution of $i$th animal and $h^2$ was heritability which was 0.390 for milk, fat, and protein, 0.066 for mastitis, and 0.064 for fertility.

## Statistical models for predicting breeding values

A linear mixed model using pedigree (PBLUP) or genome-wide marker (GBLUP) based relationships and a Bayesian four-distribution mixture model were used for predicting breeding values. In GBLUP and Bayesian four-distribution mixture model, effects of selected WGS SNPs on genomic prediction were investigated by comparing four SNP scenarios: (i) 54K; (ii) 54K + DFS; (iii) 54K + FRA; and (iv) 54K + DFS + FRA. Furthermore, we assessed a one-component model and a two-component model on their efficiency to use the information of selected WGS SNPs. A one-component model considering all SNPs as one genetic component was applied for all four SNP scenarios; while a two-component model considering 54K SNPs and selected WGS SNPs as two separate genetic components was applied only for scenarios including selected WGS SNPs. It was assumed that SNPs captured all genetic variations, and therefore, the residual polygenic effect was not included in the models.

### PBLUP model

The PBLUP model is

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

where $\mathbf{y}$ is the vector of DRPs; $\mathbf{1}$ is the vector of ones; $\mu$ is the overall mean; $\mathbf{a}$ is the vector of additive genetic effects; $\mathbf{Z}$ is the incidence matrix relating $\mathbf{a}$ to phenotypes; and $\mathbf{e}$ is the vector of random residuals. It was assumed that $\mathbf{a} \sim N\left(0, \mathbf{A}\sigma^2_{\mathrm{a}}\right)$ and $\mathbf{e} \sim N\left(0, \mathbf{D}\sigma^2_e\right)$. The $\mathbf{A}$ is the additive relationship matrix constructed from the pedigree which traced genotyped animals three generations back. The $\mathbf{D}$ is the diagonal matrix with elements $d_{jj} = \left(1 - r^2_{\mathrm{DRP}}\right)/r^2_{\mathrm{DRP}}$ to account for heterogeneous residual variances ($\sigma^2_e$) due to different reliabilities of DRP ($r^2_{\mathrm{DRP}}$). The estimation of variance components, and the prediction of breeding values with PBLUP models were performed using the DMU software (Madsen and Jensen 2012).

### GBLUP model

The one-component GBLUP (G1) model is

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{a} + \mathbf{e}. \tag{2}$$

The two-component GBLUP (G2) model is

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{a}_{\mathbf{54K}} + \mathbf{Z}\mathbf{a}_{\mathbf{WGS}} + \mathbf{e}, \tag{3}$$

where $\mathbf{y}$ is the vector of DRPs; $\mathbf{1}$ is the vector of ones; $\mu$ is the overall mean; $\mathbf{a}$, $\mathbf{a}_{\mathbf{54K}}$, and $\mathbf{a}_{\mathbf{WGS}}$ are vectors of additive genetic effects accounted by all SNPs in the model, by 54K SNPs and by selected WGS SNPs; $\mathbf{Z}$ is the incidence matrix relating $\mathbf{a}$, $\mathbf{a}_{\mathbf{54K}}$, and $\mathbf{a}_{\mathbf{WGS}}$ to phenotypes; and $\mathbf{e}$ is the vector of random residuals. It is assumed that $\mathbf{a} \sim N\left(0, \mathbf{G}\sigma^2_{\mathrm{a}}\right)$, $\mathbf{a}_{\mathbf{54K}} \sim N(0, \mathbf{G}_{54\mathrm{K}}\sigma^2_{\mathrm{a}_{54\mathrm{K}}})$, $\mathbf{a}_{\mathbf{WGS}} \sim N(0, \mathbf{G}_{\mathbf{WGS}}\sigma^2_{\mathrm{a}_{\mathrm{WGS}}})$, and $\mathbf{e} \sim N\left(0, \mathbf{D}\sigma^2_e\right)$. The $\mathbf{G}$, $\mathbf{G}_{\mathbf{54K}}$, and $\mathbf{G}_{\mathbf{WGS}}$ are genomic relationship matrices constructed from all SNPs in the model, from 54K SNPs and from selected WGS SNPs using method 1 in VanRaden (2008); $\sigma^2_{\mathrm{a}}$, $\sigma^2_{\mathrm{a}_{54\mathrm{K}}}$ and $\sigma^2_{\mathrm{a}_{\mathrm{WGS}}}$ are additive genetic variances explained by SNPs in $\mathbf{G}$, $\mathbf{G}_{\mathbf{54K}}$, and $\mathbf{G}_{\mathbf{WGS}}$, respectively. The $\mathbf{D}$ is the diagonal matrix with elements $d_{jj} = \left(1 - r^2_{\mathrm{DRP}}\right)/r^2_{\mathrm{DRP}}$ to account for heterogeneous residual variances ($\sigma^2_e$) due to different reliabilities of DRP ($r^2_{\mathrm{DRP}}$). In the G2 model, proportions of total variances explained by 54K SNPs and selected WGS SNPs are estimated from the data, while the covariance between 54K SNPs and selected WGS SNPs is assumed to be zero. The estimation of variance components, and the prediction of breeding values with GBLUP models were performed using the DMU software (Madsen and Jensen 2012).

### Bayesian four-distribution mixture model

The one-component Bayesian four-distribution mixture (B1) model is

$$\boldsymbol{y} = 1\mu + \mathbf{X}\mathbf{g} + \boldsymbol{e}. \tag{4}$$

The two-component Bayesian four-distribution mixture (B2) model is

$$\boldsymbol{y} = 1\mu + \mathbf{X}_{54\mathrm{K}}\mathbf{g}_{\mathbf{54K}} + \mathbf{X}_{\mathbf{WGS}}\mathbf{g}_{\mathbf{WGS}} + \mathbf{e}, \tag{5}$$

where $\mathbf{y}$ is the vector of DRPs; $\mathbf{1}$ is the vector of ones; $\mu$ is the overall mean; $\mathbf{g}$, $\mathbf{g}_{\mathbf{54K}}$, and $\mathbf{g}_{\mathbf{WGS}}$ are vectors of effects for all SNPs in the model, 54K SNPs, and selected WGS SNPs;$\mathbf{X}$, $\mathbf{X}_{\mathbf{54K}}$, and $\mathbf{X}_{\mathbf{WGS}}$ are genotype matrices for all SNPs in the model, 54K SNPs, and selected WGS SNPs; and $\mathbf{e}$ is the vector of random residuals. It is assumed that the distribution of marker effects ($\mathbf{g}$, $\mathbf{g}_{\mathbf{54K}}$, or $\mathbf{g}_{\mathbf{QTL}}$) follows a mixture of four normal distributions:

$$\mathbf{g}_{\mathbf{i}} \sim \pi_{i1}N\left(0, \mathbf{I}\sigma^2_{i1}\right) + \pi_{i2}N\left(0, \mathbf{I}\sigma^2_{i2}\right) + \pi_{i3}N\left(0, \mathbf{I}\sigma^2_{i3}\right) + \pi_{i4}N\left(0, \mathbf{I}\sigma^2_{i4}\right),$$

where $i$ is the $i$th genetic component in the model; $\pi_{ij}$ ($j =$ 1,2,3 and 4) is the probability of an SNP belongs to the $j$th distribution within the $i$th component, and $\sigma^2_{ij}$ is the variance for $j$th distribution within the $i$th component. In the present study, $\pi_{ij}$ is sampled from the Dirichlet distribution $\pi_{ij} =$ ($\pi_{i1}$, $\pi_{i2}$, $\pi_{i3}$, $\pi_{i4}$) ~ dir (125, 25, 5, 1) with prior $\pi_{i1} = 0.889$, $\pi_{i2} = 0.1$, $\pi_{i3} = 0.01$, and $\pi_{i4} = 0.001$, where $\sigma^2_{ij}$ is updated from the data with $1000\sigma^2_{i1} = 100\sigma^2_{i2} = 10\sigma^2_{i3} = \sigma^2_{i4}$. The ratios among $\sigma^2_{i1}$, $\sigma^2_{i2}$, $\sigma^2_{i3}$, and $\sigma^2_{i4}$ are fixed, thus, only one of them is required to be estimated within each genetic component. It is assumed that $\mathbf{e} \sim N\left(0, \mathbf{D}\sigma^2_e\right)$, where $\mathbf{D}$ is the diagonal matrix with elements $d_{jj} = \left(1 - r^2_{\mathrm{DRP}}\right)/r^2_{\mathrm{DRP}}$ to account for heterogeneous residual variances ($\sigma^2_e$) due to different reliabilities of DRP ($r^2_{\mathrm{DRP}}$). In the B2 model, proportions of total variances explained by 54K SNPs and selected WGS SNPs are estimated from the data, while the covariance between 54K SNPs and selected WGS SNPs is assumed to be zero. Each of the Bayesian four-distribution mixture model was run as a single chain with a total length of 50,000 Markov chain samples, where the first 10,000 iterations were discarded as burn-in. Ultimately, every 20th sample of the remaining 40,000 iterations were saved for the posterior analysis. The analyses with Bayesian four-distribution mixture models were performed using the Bayz software (http://www.bayz.biz).

## Validation of genomic prediction

### Reference and validation populations

To investigate the effects of using selected WGS SNPs on genomic prediction along with the change of the reference population size by including US Jersey bulls and Danish Jersey cows, five reference populations were tested in this study: (i) Danish bulls (DK); (ii) Danish and US bulls (DKUS); (iii) Danish cows (COW); (iv) Danish bulls and cows (DKCOW); and (v) Danish and US bulls and Danish cows (DKUSCOW). It is improper to include bulls in the validation population given cows in the reference, since most genotyped cows were daughters or sibs for genotyped bulls. Therefore, cows were used as the validation population. To avoid strong sib-relationships between reference and validation populations, a strategy similar to that in Su et al. (2016) was used to create validation and reference populations. Genotyped cows born in and after 2014 and their genotyped paternal female half-sibs born after July 1st 2008 were considered as the validation cows. Half-sib families with size larger than 500 were removed from the validation set and kept in the reference set to avoid a large reduction of the reference population size. Finally, the validation population included 5829 validation cows from 155 paternal half-sib families. For the reference populations, validation cows' maternal female and male half-sibs born after July 1st 2008 and progenies of validation cows and the sibs of these progenies were excluded. For fertility, however, DRP was only available in bulls. The validation population for fertility was 281 Danish bulls born in and after 2005, while reference populations were: (i) 1029 Danish bulls born before 2005 (DK), and (ii) a combination of 1029 Danish bulls and 1153 US bulls (DKUS). Only animals with the reliability of DRP higher than 0.20 were used as the validation population. Validation and reference scenarios as well as numbers of animals in validation and reference populations for all traits are presented in Table 1. Reliabilities of DRP in reference and validation populations are presented in Table 2.

### Reliability, bias, and stability

The predictability for estimating breeding values was assessed by reliability, bias, and stability. The reliability of prediction was measured as the squared correlation between estimated breeding values and DRP divided by the average reliability of DRP for the animals in the validation population. The bias of prediction was measured as the regression coefficient of DRP on the estimated breeding values for the animals in the validation population. The stability of prediction was measured as the correlation between breeding values estimated from the reduced dataset (using DRP from the reference population) and the full dataset (using DRP from both reference and validation populations) for the animals in the validation population, which was performed in scenarios of 54K and 54K + DFS + FRA using GBLUP models.

We compared reliability (or bias) among different SNP sets given the same reference population and model, among different reference populations given the same SNP set and model, and among different models given the same reference population and SNP set. A nonparametric bootstrap sample, with an equal size as the validation population, was obtained by randomly sampling with replacement from validation animals. We repeated the bootstrap procedure to get 10,000 bootstrap samples. The standard deviation of reliability (or bias) and the contrasts from 10,000 bootstrap samples were used as the standard error of reliability (or bias) and the contrasts. A two-tailed paired *t*-test was used to compare reliability (or bias) between a pair of scenarios. A Bonferroni correction was used to control the false positive caused by multiple comparisons.

## Comparison of genomic prediction with or without the selected WGS SNPs while keeping the same number of SNPs

The SNP density increased by adding additional selected WGS SNPs. For example, the number of SNPs used for genomic prediction increased 3669 by adding DFS + FRA SNPs. We hypothesized that changes in reliabilities (or bias) after adding selected WGS SNPs was due to these SNPs being or linking closely to causative mutations instead of the increase in SNP density. To test this hypothesis, we randomly removed 3669 SNPs from 54K and created a new SNP scenario together with DFS and FRA SNPs (54Kminus + DFS + FRA). Thus, the number of SNPs in 54Kminus+DFS+FRA was the same as that in 54K. We repeated this procedure five times. The average reliability from five replicates of 54Kminus+DFS+FRA for each model and each reference population was compared with the reliability from the 54K using the same model and the same reference population.

**Table 1** Number of animals in reference and validation populations

| Traits | Validation population | Reference population | | | |
|---|---|---|---|---|---|
| | | Scenarios | Danish bulls | US bulls | Danish cows |
| Milk, protein, fat, and mastitis | 5829 Danish cows | DK | 1282 | – | – |
| | | DKUS | 1282 | 1148 | – |
| | | COW | – | – | 8763 |
| | | DKCOW | 1282 | – | 8602 |
| | | DKUSCOW | 1282 | 1148 | 8602 |
| Fertility | 281 Danish bulls | DK | 1029 | – | – |
| | | DKUS | 1029 | 1157 | – |

*DK* Danish bulls as the reference population, *DKUS* Danish and US bulls as the reference population, *COW* Danish cows as the reference population, *DKCOW* Danish bulls and cow as the reference population, *DKUSCOW* Danish and US bulls and Danish cows as the reference population

**Table 2** Reliability of DRP ($r^2_{\mathrm{DRP}}$) in reference and validation populations

| Trait | $h^2$ | $r^2_{\mathrm{DRP}}$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | DK | DKUS | COW | DKCOW | DKUSCOW | Validation population |
| Milk | 0.390 | 0.921 | 0.846 | 0.487 | 0.543 | 0.565 | 0.446 |
| Protein | 0.390 | 0.921 | 0.846 | 0.487 | 0.543 | 0.565 | 0.446 |
| Fat | 0.390 | 0.921 | 0.846 | 0.487 | 0.543 | 0.565 | 0.446 |
| Mastitis | 0.066 | 0.806 | 0.738 | 0.233 | 0.307 | 0.341 | 0.240 |
| Fertility[a] | 0.064 | 0.624 | 0.609 | – | – | – | 0.661 |

*DK* Danish bulls as the reference population, *DKUS* Danish and US bulls as the reference population, *COW* Danish cows as the reference population, *DKCOW* Danish bulls and cow as the reference population, *DKUSCOW* Danish and US bulls and Danish cows as the reference population

[a]Bulls were used as the validation population

## Results

Reliabilities of predicting breeding values using a PBLUP model and a G1 model are presented in Table 3. Using relationships derived from SNPs yielded much higher reliabilities (0.388 average across all SNP sets) compared with using relationships derived from pedigree (0.173) for all traits. For PBLUP, standard errors of reliabilities for mastitis (~1.5 times more) and fertility (~2.5 times more) were much larger than those for milk production traits, indicating that the estimates of reliabilities for mastitis and fertility were less precise than those for milk production traits. Compared with using 54K SNPs alone, adding additional selected WGS SNPs led to significant improvements of reliabilities for milk and protein, small improvements for fat and mastitis, whereas no improvement for fertility. For milk production and mastitis, both DFS and FRA sets of additional WGS SNPs improved reliability (an average gain of 0.025 and 0.029, respectively), and the inclusion of all selected WGS SNPs generally achieved the highest reliabilities (an average gain of 0.034). Generally, reliabilities improved along with the increase of the reference population size when using the same SNP set. Including US Jersey bulls in the reference (DKUS) led to significant improvements (0.068 over all SNP sets) of reliabilities compared with only using a DK reference for milk, protein, and fat. Including cows in the reference population led to significant improvements in reliabilities compared with only using bulls as reference. By expanding the reference population from the DK reference to the DKCOW reference, the average improvement of reliability across all traits over all SNP sets was 0.148, and from the DKUS reference to the DKUSCOW reference was 0.109. Furthermore, for milk production and mastitis, improvements of reliabilities by integrating selected WGS SNPs ranged from 0.023 to 0.037 when using different reference populations. The gains in order were DKUSCOW<COW<DKCOW<DK<DKUS.

Reliabilities in genomic prediction from different models for 54K and 54K + DFS + FRA are presented in Table 4. When using 54K, the B1 model was significantly superior to a G1 model for milk, protein, and fat, but equal to a G1 model for mastitis and fertility. In this case, the average improvement of reliability from a G1 model to a B1 model for milk, protein, and fat was 0.051. When using 54K +DFS + FRA, a B1 model was significantly superior to a G1 model for milk and protein but equal to a G1 model for fat, mastitis, and fertility. In this case, the average improvement of reliability from a G1 model to a B1 model for milk and protein was 0.023. Although a B1 model was better than a G1 model for milk and protein, the extra gain in reliability from using selected WGS SNPs (DFS+FRA) was smaller for a B1 model (0.011) than a G1 model (0.059). Regarding the comparisons between a one-component model and a two-component model when using 54K + DFS + FRA, significant differences were observed only for milk with DK, DKUS, and DKUSCOW as reference. For milk, a G2 model was generally superior to a G1 model (0.016), whereas a B2 model was equal to a B1 model.

Variance components estimated from a G1 model and a G2 model are presented in additional Table 1. In the G1 model, differences between variances components estimated before and after adding selected WGS SNPs were small. In the G2 model, proportions of genetic variances explained by selected WGS SNPs were 40.2% for milk, 31.2% for protein, 23.6% for fat, 17.5% for mastitis, and 25.5% for fertility, average on all scenarios with different reference populations and SNP sets. Total additive genetic variances (the sum of two genetic components) from the G2 model were in general slightly smaller than those from the G1 model, suggesting the covariance between the two genetic components could not be zero. The variances estimated from different sets of phenotypic data were somewhat different due to different structures in different data sets.

Bias for predicting breeding values using a PBLUP model and a G1 model is presented in Table 5. When cows were included in the reference population, regression

**Table 3** Reliabilities from a PBLUP[a] model and a G1[b] model using different SNP scenarios[c] with significance tests[d]

| Trait | Reference | PBLUP | G1 | | | |
|---|---|---|---|---|---|---|
| | | | 54K | 54K + DFS | 54K + FRA | 54K + DFS + FRA |
| Milk | DK | 0.132 (0.013) | $^{e}0.320_{d}$ | $^{e}0.397_{c}$ | $^{e}0.403_{b}$ | $^{e}0.424_{a}$ |
| | DKUS | 0.174 (0.015) | $^{d}0.426_{d}$ | $^{d}0.510_{c}$ | $^{d}0.519_{b}$ | $^{d}0.533_{a}$ |
| | COW | 0.105 (0.012) | $^{c}0.577_{d}$ | $^{c}0.637_{c}$ | $^{c}0.645_{b}$ | $^{c}0.649_{a}$ |
| | DKCOW | 0.149 (0.014) | $^{b}0.619_{c}$ | $^{b}0.679_{b}$ | $^{b}0.684_{b}$ | $^{b}0.691_{a}$ |
| | DKUSCOW | 0.189 (0.016) | $^{a}0.655_{d}$ | $^{a}0.704_{c}$ | $^{a}0.710_{b}$ | $^{a}0.715_{a}$ |
| Protein | DK | 0.175 (0.015) | $^{d}0.268_{d}$ | $^{d}0.291_{c}$ | $^{d}0.298_{b}$ | $^{d}0.306_{a}$ |
| | DKUS | 0.209 (0.016) | $^{c}0.326_{d}$ | $^{c}0.351_{c}$ | $^{c}0.361_{b}$ | $^{c}0.368_{a}$ |
| | COW | 0.116 (0.013) | $^{c}0.366_{c}$ | $^{c}0.389_{b}$ | $^{c}0.397_{a}$ | $^{c}0.399_{a}$ |
| | DKCOW | 0.150 (0.014) | $^{b}0.402_{c}$ | $^{b}0.427_{b}$ | $^{b}0.432_{b}$ | $^{b}0.435_{a}$ |
| | DKUSCOW | 0.178 (0.015) | $^{a}0.429_{c}$ | $^{a}0.449_{b}$ | $^{a}0.454_{ab}$ | $^{a}0.456_{a}$ |
| Fat | DK | 0.199 (0.016) | $^{c}0.265_{c}$ | $^{c}0.277_{ab}$ | $^{c}0.276_{b}$ | $^{c}0.281_{a}$ |
| | DKUS | 0.214 (0.016) | $^{b}0.296_{c}$ | $^{b}0.308_{b}$ | $^{b}0.313_{b}$ | $^{b}0.316_{a}$ |
| | COW | 0.148 (0.014) | $^{b}0.333_{c}$ | $^{b}0.340_{b}$ | $^{b}0.341_{ab}$ | $^{b}0.343_{a}$ |
| | DKCOW | 0.222 (0.017) | $^{a}0.376_{a}$ | $^{a}0.380_{a}$ | $^{a}0.381_{a}$ | $^{a}0.382_{a}$ |
| | DKUSCOW | 0.231 (0.017) | $^{a}0.384_{a}$ | $^{a}0.389_{a}$ | $^{a}0.389_{a}$ | $^{a}0.391_{a}$ |
| Mastitis | DK | 0.205 (0.035) | $^{c}0.254_{a}$ | $^{c}0.256_{a}$ | $^{b}0.256_{a}$ | $^{b}0.258_{a}$ |
| | DKUS | 0.196 (0.034) | $^{bc}0.267_{a}$ | $^{bc}0.268_{a}$ | $^{b}0.268_{a}$ | $^{b}0.269_{a}$ |
| | COW | 0.168 (0.031) | $^{abc}0.295_{a}$ | $^{abc}0.297_{a}$ | $^{ab}0.298_{a}$ | $^{ab}0.299_{a}$ |
| | DKCOW | 0.154 (0.030) | $^{ab}0.321_{b}$ | $^{ab}0.323_{b}$ | $^{a}0.331_{a}$ | $^{a}0.332_{a}$ |
| | DKUSCOW | 0.146 (0.030) | $^{a}0.323_{b}$ | $^{a}0.323_{b}$ | $^{a}0.333_{a}$ | $^{a}0.333_{a}$ |
| Fertility[e] | DK | 0.161 (0.050) | $^{a}0.301_{a}$ | $^{a}0.298_{a}$ | $^{a}0.291_{a}$ | $^{a}0.288_{a}$ |
| | DKUS | 0.189 (0.053) | $^{a}0.303_{a}$ | $^{a}0.302_{a}$ | $^{a}0.294_{a}$ | $^{a}0.293_{a}$ |

*DK* Danish bull as the reference population, *DKUS* Danish and US bull as the reference population, *COW* Danish cows as the reference population, *DKCOW* Danish bull and cows as the reference population, *DKUSCOW* Danish and US bulls and Danish cows as the reference population

[a] PBLUP: pedigree BLUP

[b] G1: one-component GBLUP

[c] 54K: SNPs in 54K chip; 54K+DFS: SNPs in 54K chip together with WGS SNPs selected by analysis of data from major dairy breeds in Denmark–Finland–Sweden; 54K + FRA: SNPs in 54K chip together with WGS SNPs selected by analysis of data from major dairy breeds in France; 54K + DFS + FRA: SNPs in 54K chip together with WGS SNPs selected by analysis of data from major dairy breeds in Denmark–Finland–Sweden and France

[d] Letters in the superscript were for comparisons among reference populations using the same SNP scenario and model; letters in subscript were for comparisons among SNP scenarios using the same reference populations and model. Same letters denote no significant difference; while different letters denote significant difference at $P = 0.05$ after Bonferroni correction

[e] Bulls were used as the validation population

coefficients were lower from GBLUP models than from PBLUP models. When only bulls were in the reference population, no clear trend in the regression coefficient was observed between GBLUP and PBLUP models. Using selected WGS SNPs generally led to equal or significantly increased regression coefficients compared with using only 54K SNPs. Regression coefficients further deviated from unity indicated more bias. Regression coefficients increased from a DK to a DKUS reference and from a DKCOW to a DKUSCOW reference for all traits, except for from a DKCOW to a DKUSCOW reference for mastitis. Besides, the inclusion of cows in the reference led to equal or significantly decreased regression coefficients when using genomic information. Bias of genomic prediction from different models for 54K and 54K + DFS + FRA is presented in Table 6. Generally, there was no clear trend for bias observed from different models.

Stabilities of predicting breeding values for 54K and 54K + DFS + FRA using GBLUP models are presented in additional Table 2. The average correlation between predictions from full and reduced datasets was 0.834 using 54K SNPs with G1 models, while it was 0.851 and 0.856 using 54K + DFS + FRA SNPs with G1 and G2 models, respectively. In general, adding selected WGS SNPs

**Table 4** Reliabilities for 54K[a] and 54K + DFS + FRA[b] using different models[c], with significance tests[d]

| Trait | Reference | 54K | | 54K+DFS+FRA | | | |
|---|---|---|---|---|---|---|---|
| | | G1 | B1 | G1 | B1 | G2 | B2 |
| Milk | DK | $0.320_{b}$ | $0.463_{a}$ | $0.424_{c}$ | $0.484_{a}$ | $0.463_{b}$ | $0.489_{a}$ |
| | DKUS | $0.426_{b}$ | $0.549_{a}$ | $0.533_{c}$ | $0.576_{a}$ | $0.552_{b}$ | $0.586_{a}$ |
| | COW | $0.577_{b}$ | $0.667_{a}$ | $0.649_{b}$ | $0.674_{a}$ | $0.652_{b}$ | $0.675_{a}$ |
| | DKCOW | $0.619_{b}$ | $0.704_{a}$ | $0.691_{c}$ | $0.721_{a}$ | $0.703_{b}$ | $0.723_{a}$ |
| | DKUSCOW | $0.655_{b}$ | $0.732_{a}$ | $0.715_{b}$ | $0.741_{a}$ | $0.722_{b}$ | $0.742_{a}$ |
| Protein | DK | $0.268_{b}$ | $0.299_{a}$ | $0.306_{b}$ | $0.314_{a}$ | $0.309_{ab}$ | $0.305_{ab}$ |
| | DKUS | $0.326_{b}$ | $0.378_{a}$ | $0.368_{b}$ | $0.383_{a}$ | $0.371_{b}$ | $0.375_{ab}$ |
| | COW | $0.366_{b}$ | $0.396_{a}$ | $0.399_{a}$ | $0.400_{a}$ | $0.394_{a}$ | $0.401_{a}$ |
| | DKCOW | $0.402_{b}$ | $0.440_{a}$ | $0.435_{b}$ | $0.444_{a}$ | $0.439_{ab}$ | $0.446_{a}$ |
| | DKUSCOW | $0.429_{b}$ | $0.465_{a}$ | $0.456_{b}$ | $0.468_{a}$ | $0.457_{b}$ | $0.468_{a}$ |
| Fat | DK | $0.265_{b}$ | $0.278_{a}$ | $0.281_{a}$ | $0.284_{a}$ | $0.276_{a}$ | $0.277_{a}$ |
| | DKUS | $0.296_{b}$ | $0.314_{a}$ | $0.316_{ab}$ | $0.318_{a}$ | $0.311_{ab}$ | $0.312_{b}$ |
| | COW | $0.333_{b}$ | $0.343_{a}$ | $0.343_{a}$ | $0.347_{a}$ | $0.345_{a}$ | $0.348_{a}$ |
| | DKCOW | $0.376_{a}$ | $0.382_{a}$ | $0.382_{a}$ | $0.384_{a}$ | $0.382_{a}$ | $0.384_{a}$ |
| | DKUSCOW | $0.384_{b}$ | $0.398_{a}$ | $0.391_{ab}$ | $0.398_{a}$ | $0.391_{b}$ | $0.397_{ab}$ |
| Mastitis | DK | $0.254_{a}$ | $0.252_{a}$ | $0.258_{a}$ | $0.243_{ab}$ | $0.250_{a}$ | $0.226_{b}$ |
| | DKUS | $0.267_{a}$ | $0.267_{a}$ | $0.269_{a}$ | $0.262_{a}$ | $0.261_{ab}$ | $0.245_{b}$ |
| | COW | $0.295_{a}$ | $0.296_{a}$ | $0.299_{a}$ | $0.298_{a}$ | $0.296_{a}$ | $0.296_{a}$ |
| | DKCOW | $0.321_{a}$ | $0.323_{a}$ | $0.332_{a}$ | $0.328_{a}$ | $0.329_{a}$ | $0.327_{a}$ |
| | DKUSCOW | $0.323_{a}$ | $0.328_{a}$ | $0.333_{a}$ | $0.334_{a}$ | $0.332_{a}$ | $0.337_{a}$ |
| Fertility[e] | DK | $0.301_{a}$ | $0.292_{a}$ | $0.288_{ab}$ | $0.276_{a}$ | $0.267_{ab}$ | $0.243_{b}$ |
| | DKUS | $0.303_{a}$ | $0.300_{a}$ | $0.293_{a}$ | $0.292_{a}$ | $0.288_{a}$ | $0.281_{a}$ |

*DK* Danish bull as the reference population, *DKUS* Danish and US bull as the reference population, *COW* Danish cows as the reference population, *DKCOW* Danish bull and cows as the reference population. *DKUSCOW* Danish and US bulls and Danish cows as the reference population

[a] 54K: SNPs in the 54K chip

[b] 54K + DFS + FRA: SNPs in 54K chip together with WGS SNPs selected by analysis of data from major dairy breeds in Denmark–Finland–Sweden and France

[c] G1: one-component GBLUP model; G2: two-component GBLUP model; B1: one-component Bayesian four-distribution mixture model; B2: two-component Bayesian four-distribution mixture model

[d] Letters in the right lower position were for comparisons among models using the same reference population and SNP scenario. Same letters denote no significant difference; while different letters denote significant difference

[e] Bulls were used as the validation population

improved prediction stabilities. The G2 model yielded better stabilities than the G1 model for milk, fat, and protein but not for mastitis. For scenarios using different reference populations, stabilities increased with the increase of the reference population size, ranging from 0.764 for using a DK reference to 0.909 for using a DKUSCOW reference.

Differences between 54Kminus+DFS+FRA and 54K in reliabilities and bias using G1 and B1 models are presented in additional Tables 3 and 4. Generally, differences in reliabilities (or bias) between 54Kminus+DFS+FRA and 54K were similar to those between 54K + DFS + FRA and 54K. Improvements in reliabilities from 54K to 54Kminus +DFS+FRA were 0.082 for milk, 0.034 for protein, 0.012 for fat, and 0.006 for mastitis using G1 models, while those were 0.016 for milk, 0.006 for protein, 0.002 for fat, and no improvement for mastitis using B1 models. There was no improvement for fertility using both G1 and B1 models. Differences in bias between 54Kminus+DFS+FRA and 54K were no more than 0.05 using both G1 and B1 models except for milk from DK (0.10) and DKUS reference (0.10) using G1 models.

## Discussion

In this study, we investigated the effects of integrating selected WGS SNPs in the EuroGenomics customized LD chip (Boichard et al. 2018) to 54K SNPs on genomic

**Table 5** Regression coefficients of DRP on prediction from a PBLUP[a] model and a G1[b] model using different SNP scenarios[c] with significance tests[d]

| Trait | Reference | PBLUP | G1 | | | |
|---|---|---|---|---|---|---|
| | | | 54K | 54K + DFS | 54K + FRA | 54K + DFS + FRA |
| Milk | DK | 1.03 (0.05) | $^{ab}1.07_{c}$ | $^{a}1.16_{b}$ | $^{a}1.17_{b}$ | $^{a}1.18_{a}$ |
| | DKUS | 1.12 (0.05) | $^{a}1.12_{c}$ | $^{a}1.19_{b}$ | $^{a}1.20_{ab}$ | $^{a}1.21_{a}$ |
| | COW | 1.10 (0.06) | $^{c}0.94_{a}$ | $^{d}0.92_{b}$ | $^{d}0.92_{ab}$ | $^{d}0.92_{ab}$ |
| | DKCOW | 1.12 (0.05) | $^{b}1.05_{a}$ | $^{c}1.04_{a}$ | $^{c}1.04_{a}$ | $^{c}1.04_{a}$ |
| | DKUSCOW | 1.21 (0.05) | $^{a}1.07_{a}$ | $^{b}1.07_{a}$ | $^{b}1.07_{a}$ | $^{b}1.07_{a}$ |
| Protein | DK | 1.00 (0.04) | $^{cd}0.87_{c}$ | $^{cd}0.90_{b}$ | $^{cd}0.91_{ab}$ | $^{bc}0.91_{a}$ |
| | DKUS | 1.12 (0.04) | $^{ab}0.93_{c}$ | $^{ab}0.96_{b}$ | $^{ab}0.97_{ab}$ | $^{a}0.98_{a}$ |
| | COW | 1.09 (0.06) | $^{abcd}0.95_{a}$ | $^{abcd}0.93_{a}$ | $^{abcd}0.94_{a}$ | $^{abc}0.94_{a}$ |
| | DKCOW | 0.93 (0.04) | $^{bd}0.91_{a}$ | $^{bd}0.92_{a}$ | $^{bd}0.92_{a}$ | $^{c}0.92_{a}$ |
| | DKUSCOW | 1.02 (0.04) | $^{ac}0.93_{a}$ | $^{ac}0.94_{a}$ | $^{ac}0.94_{a}$ | $^{ab}0.94_{a}$ |
| Fat | DK | 0.96 (0.04) | $^{bd}0.81_{a}$ | $^{ab}0.82_{a}$ | $^{bd}0.82_{a}$ | $^{ab}0.82_{a}$ |
| | DKUS | 1.06 (0.04) | $^{ac}0.85_{a}$ | $^{ab}0.84_{a}$ | $^{ac}0.85_{a}$ | $^{ab}0.85_{a}$ |
| | COW | 0.94 (0.05) | $^{ab}0.85_{a}$ | $^{a}0.86_{a}$ | $^{ab}0.86_{a}$ | $^{a}0.86_{a}$ |
| | DKCOW | 1.00 (0.04) | $^{cd}0.82_{a}$ | $^{b}0.82_{a}$ | $^{cd}0.82_{a}$ | $^{b}0.82_{a}$ |
| | DKUSCOW | 1.04 (0.04) | $^{abcd}0.83_{a}$ | $^{ab}0.83_{a}$ | $^{abcd}0.83_{a}$ | $^{ab}0.83_{a}$ |
| Mastitis | DK | 1.37 (0.12) | $^{ab}1.12_{a}$ | $^{a}1.12_{a}$ | $^{a}1.12_{a}$ | $^{a}1.12_{a}$ |
| | DKUS | 1.39 (0.12) | $^{a}1.17_{a}$ | $^{a}1.17_{a}$ | $^{a}1.16_{a}$ | $^{a}1.16_{a}$ |
| | COW | 1.46 (0.14) | $^{bc}1.00_{a}$ | $^{ab}1.00_{a}$ | $^{abc}1.00_{a}$ | $^{abc}1.00_{a}$ |
| | DKCOW | 1.21 (0.12) | $^{cd}0.95_{a}$ | $^{b}0.95_{a}$ | $^{b}0.96_{a}$ | $^{b}0.96_{a}$ |
| | DKUSCOW | 1.07 (0.12) | $^{d}0.90_{a}$ | $^{c}0.90_{a}$ | $^{c}0.91_{a}$ | $^{c}0.91_{a}$ |
| Fertility[e] | DK | 0.90 (0.15) | $^{a}1.10_{a}$ | $^{a}1.09_{a}$ | $^{a}1.08_{a}$ | $^{a}1.07_{a}$ |
| | DKUS | 0.92 (0.14) | $^{a}1.05_{a}$ | $^{a}1.05_{a}$ | $^{a}1.02_{a}$ | $^{a}1.03_{a}$ |

*DK* Danish bull as the reference population, *DKUS* Danish and US bull as the reference population, *COW* Danish cows as the reference population, *DKCOW* Danish bull and cows as the reference population, *DKUSCOW* Danish and US bulls and Danish cows as the reference population

[a] PBLUP: pedigree BLUP

[b] G1: one-component GBLUP

[c] 54K: SNPs in 54K chip; 54K + DFS: SNPs in 54K chip together with WGS SNPs selected by analysis of data from major dairy breeds in Denmark–Finland–Sweden; 54K + FRA: SNPs in 54K chip together with WGS SNPs selected by analysis of data from major dairy breeds in France; 54K +DFS + FRA: SNPs in 54K chip together with WGS SNPs selected by analysis of data from major dairy breeds in Denmark–Finland–Sweden and France

[d] Letters in the left upper positions were for comparisons among reference populations using the same SNP scenario and model; letters in the right lower position were for comparisons among SNP scenarios using the same reference populations and model. Same letters denote no significant difference, while different letters denote significant difference

[e] Bulls were used as the validation population

prediction in Danish Jersey, which is a numerically small breed. Using selected WGS SNPs improved reliabilities and stabilities for milk production and mastitis, which was consistent with the results of integrating QTL SNPs in genomic prediction for Nordic Holsteins and Red (Brøndum et al. 2015). Besides, differences between 54Kminus+DFS +FRA and 54K were similar to those between 54K + DFS + FRA and 54K, supporting the hypothesis that improvements of reliabilities from integrating selected WGS SNPs were due to selected WGS SNPs being or linking closely to causative mutations, not due to the increase of SNP density. For fertility, however, using selected WGS SNPs did not improve the reliability, which was in line with the previous study in Nordic Holsteins and Red (Brøndum et al. 2015) where the improvement of reliability for fertility was limited when integrating selected WGS SNPs. Due to the polygenic nature of fertility traits (Liu et al. 2017) and the small reference population size, the power of detecting causative mutations from WGS data could be limited in Danish Jersey. Moreover, more WGS SNPs were selected for milk production traits (about two times more) than for fertility in DFS SNPs, since the number of WGS SNPs to be selected was determined by the economic weight of a specific trait in Nordic selection index (Brøndum et al. 2015).

**Table 6** Regression coefficients of DRP on prediction from 54K[a] and 54K + DFS + FRA[b] using different models[c], with significance tests[d]

| Trait | Reference | 54K | | 54K +DFS + FRA | | | |
|---|---|---|---|---|---|---|---|
| | | G1 | B1 | G1 | B1 | G2 | B2 |
| Milk | DK | $1.07_{b}$ | $1.20_{a}$ | $1.18_{ab}$ | $1.23_{a}$ | $1.16_{c}$ | $1.20_{b}$ |
| | DKUS | $1.12_{b}$ | $1.19_{a}$ | $1.21_{bc}$ | $1.22_{ab}$ | $1.20_{c}$ | $1.23_{a}$ |
| | COW | $0.94_{a}$ | $0.91_{b}$ | $0.92_{a}$ | $0.91_{a}$ | $0.91_{a}$ | $0.91_{a}$ |
| | DKCOW | $1.05_{a}$ | $1.04_{a}$ | $1.04_{a}$ | $1.04_{a}$ | $1.04_{a}$ | $1.04_{a}$ |
| | DKUSCOW | $1.07_{a}$ | $1.08_{a}$ | $1.07_{a}$ | $1.08_{a}$ | $1.07_{a}$ | $1.08_{a}$ |
| Protein | DK | $0.87_{b}$ | $0.92_{a}$ | $0.91_{ab}$ | $0.92_{a}$ | $0.89_{bc}$ | $0.87_{c}$ |
| | DKUS | $0.93_{b}$ | $0.98_{a}$ | $0.98_{ab}$ | $0.98_{a}$ | $0.97_{ab}$ | $0.96_{b}$ |
| | COW | $0.95_{a}$ | $0.92_{b}$ | $0.94_{a}$ | $0.91_{b}$ | $0.92_{b}$ | $0.91_{b}$ |
| | DKCOW | $0.91_{a}$ | $0.92_{a}$ | $0.92_{a}$ | $0.92_{a}$ | $0.92_{a}$ | $0.92_{a}$ |
| | DKUSCOW | $0.93_{b}$ | $0.95_{a}$ | $0.94_{a}$ | $0.95_{a}$ | $0.94_{a}$ | $0.95_{a}$ |
| Fat | DK | $0.81_{a}$ | $0.82_{a}$ | $0.82_{a}$ | $0.81_{a}$ | $0.79_{b}$ | $0.79_{b}$ |
| | DKUS | $0.85_{a}$ | $0.83_{a}$ | $0.85_{a}$ | $0.84_{ab}$ | $0.83_{bc}$ | $0.82_{c}$ |
| | COW | $0.85_{a}$ | $0.85_{a}$ | $0.86_{a}$ | $0.86_{a}$ | $0.87_{a}$ | $0.86_{a}$ |
| | DKCOW | $0.82_{a}$ | $0.82_{a}$ | $0.82_{a}$ | $0.82_{a}$ | $0.82_{a}$ | $0.82_{a}$ |
| | DKUSCOW | $0.83_{a}$ | $0.84_{a}$ | $0.83_{a}$ | $0.84_{a}$ | $0.83_{a}$ | $0.84_{a}$ |
| Mastitis | DK | $1.12_{a}$ | $1.12_{a}$ | $1.12_{a}$ | $1.09_{a}$ | $1.09_{a}$ | $1.01_{b}$ |
| | DKUS | $1.17_{a}$ | $1.15_{a}$ | $1.16_{a}$ | $1.13a_{b}$ | $1.13_{b}$ | $1.07_{c}$ |
| | COW | $1.00_{a}$ | $1.00_{a}$ | $1.00_{a}$ | $1.00_{a}$ | $1.00_{a}$ | $1.00_{a}$ |
| | DKCOW | $0.95_{a}$ | $0.94_{a}$ | $0.96_{a}$ | $0.95_{a}$ | $0.96_{a}$ | $0.95_{a}$ |
| | DKUSCOW | $0.90_{b}$ | $0.93_{a}$ | $0.91_{b}$ | $0.94_{a}$ | $0.91_{b}$ | $0.95_{a}$ |
| Fertility[e] | DK | $1.10_{a}$ | $1.07_{a}$ | $1.07_{a}$ | $1.03_{ab}$ | $1.02_{b}$ | $0.92_{c}$ |
| | DKUS | $1.05_{a}$ | $1.04_{a}$ | $1.03_{ab}$ | $1.03_{a}$ | $1.01_{ab}$ | $1.00_{b}$ |

*DK* Danish bull as the reference population, *DKUS* Danish and US bull as the reference population, *COW* Danish cows as the reference population, *DKCOW* Danish bull and cows as the reference population. DKUSCOW Danish and US bulls and Danish cows as the reference population

[a] 54K: SNPs in the 54K chip

[b] 54K + DFS + FRA: SNPs in 54K chip together with WGS SNPs selected by analysis of data from major dairy breeds in Denmark–Finland–Sweden and France

[c] G1: one-component GBLUP model; G2: two-component GBLUP model; B1: one-component Bayesian four-distribution mixture model; B2: two-component Bayesian four-distribution mixture model

[d] Letters in the right lower position were for comparisons among models using the same reference population and SNP scenario. Same letters denote no significant difference; while different letters denote significant difference

[e] Bulls were used as the validation population

To select a model with high efficiency to use the information of selected WGS SNPs, we compared a GBLUP model with a Bayesian four-distribution mixture model, and treating the selected WGS SNPs and 54K SNPs as the same or different genetic components. The B1 model used in this study was similar to the BayesR model which has been verified to be equal or superior to a G1 model in various species (Erbe et al. 2012). Improvements of reliabilities from selected WGS SNPs, however, were smaller for B1 models compared with G1 models since B1 models can better identify SNPs linking closely to causative mutations than G1 models when using 54K SNPs. Furthermore, for milk, a G2 model was generally superior to a G1 model which was in line with Brøndum et al. (2015), but no significant difference was observed between reliabilities from B1 and B2 models. The Bayesian mixture models already allowed to account for different variances between SNPs in the 54K chip and selected WGS SNPs when using a B1 model, and thus a B2 model worked similarly to a B1 model. The idea of putting more emphases on informative SNPs by a two-component model was similar to a weighted GBLUP model (Zhang et al. 2010), which has been verified to be better than the regular GBLUP model (Su et al. 2014). The weighted GBLUP model has similar computation costs as the regular GBLUP model but is able to reach similar reliabilities as the Bayesian mixture model, therefore, could be easily implemented in the industry. In addition, integrating DFS and FRA SNPs in a single-step model with weights on the genomic relationship matrix could further improve reliabilities for both genotyped and nongenotyped animals. A simulation study showed that a weighted single-step model efficiently used the information of causative SNPs when weighted by realistic effects or estimated effects from association studies (Fragomeni et al. 2017). Furthermore, selected WGS SNPs used in this study were discovered based on different strategies, e.g., QTL detection and bioinformatics analysis. The model which can incorporate prior biological information by defining classes of SNPs, e.g., BayesRS (Brøndum et al. 2012) and BayesRC (MacLeod et al. 2016), is likely to better use the information of causative mutations and further improve the reliability.

To test whether the benefits of integrating selected WGS SNPs still existed when increasing the reference population size, we applied multiple reference populations in genomic prediction. Reliabilities for prediction improved after including US Jersey in the reference for all traits. A joint reference with genetically related populations led to improvements in reliabilities since linkage disequilibrium persisted over long distances. For example, large improvements of reliabilities were observed when pooling Danish Jersey with US Jersey (Wiggans et al. 2015; Su et al. 2016), whereas limited or no improvement of reliability was observed when pooling Jersey with Holsteins (Lund et al. 2014). To date, the information of US Jersey bulls has been included in the routine genomic evaluation for Danish Jersey. Improvements of reliabilities for milk production traits from integrating selected WGS SNPs were almost the same when the reference population increased from DK to DKUS. The most pronounced advantage of using WGS data is in genomic prediction across populations (Iheshiulor et al. 2016), and it has been reported that selected WGS SNPs can improve genomic prediction across populations (van den Berg et al. 2016a).

Including cows in the reference population led to significant improvements of reliabilities compared with only using bulls as reference for all traits, which was in line with results from previous studies using both real (Su et al. 2016) and simulated data (Buch et al. 2012). Although the phenotypic information of cows is less informative compared with those of bulls with a large number of progenies, improvements of reliabilities were still considerable since a large number of genotyped cows were added to the reference population. In this study, reliabilities of using COW as a reference population were higher than those from using DKUS as a reference population, this could be due to that genotyped Jersey cows were much less selected than genotyped Jersey bulls. It has been reported that selective animals could underestimate reliabilities of genomic prediction (Uimari and Mäntysaari 1993; Su et al. 2012b). Improvements of reliabilities for milk production traits from using selected WGS SNPs slightly dropped after including cows in the reference. It has been reported that improvements of reliabilities from using additional information from other sources depended on the size of the original reference population (Daetwyler et al. 2008). Therefore, improvements of reliabilities from using selected WGS SNPs were dropped given the reference population dramatically enlarged by including a large number of cows. Furthermore, even in the largest reference population (i.e., DKUSCOW), benefits of using selected WGS SNPs for milk and protein remained significant. This indicated that selected WGS SNPs could also benefit for populations or breeds with relatively large population sizes.

An alternative strategy to use information from other populations is to preselect informative WGS SNPs from the population with large population size or from multiple breeds (van den Berg et al. 2016b), because of large power to detect informative SNPs. Besides, FRA SNPs performed slightly better than DFS SNPs although with poorer imputation accuracy. However, we cannot conclude that SNPs detected from bioinformatics analyses (e.g., FRA SNPs) were more meaningful in genomic prediction than those from QTL mapping (e.g., DFS SNPs) since the number of FRA SNPs was around two times of DFS SNPs. A combination of DFS and FRA SNPs yielded the highest reliabilities, which indicated that more reliable predictions as well as faster genetic gains could be achieved if the information of selected WGS SNPs can be shared across countries. In addition, other novel strategies, e.g., machine learning (Long et al. 2007), could be promising for the preselection of WGS SNPs for genomic prediction.

A total of 117 DFS SNPs and 568 FRA SNPs with MAF higher than 0.01 were removed from analyses due to low imputation accuracy. Although selected WGS SNPs used in genomic prediction (after quality control) had relatively high imputation accuracy, the imputation accuracy was still much lower than unity. Compared with true genotypes, imputed genotypes decreased reliabilities and unbiasedness of genomic prediction, where the degree of decrement was influenced by the imputation accuracy (van den Berg et al. 2017). Improvement of reliability in genomic prediction from using selected WGS SNPs could be enlarged if more animals were directly genotyped with customized LD chip instead of imputation.

In summary, based on the results of this study, the efforts for developing and implementing the customized SNP chip with WGS SNPs selected from QTL mapping and/or bioinformatics are worthwhile in the industry, especially for the numerically small breed. Firstly, improvements in reliabilities from integrating selected WGS SNPs are promising in general. Secondly, the inclusion of thousands of additional WGS SNPs would not lead to a large increase of the computational burden, and therefore easy to implement in the industry. Thirdly, benefiting from the development of genotyping technologies, the costs of the customized SNP chip (e.g., customized LD chip) is similar to that of the standard SNP chip (e.g., standard LD chip), which makes the improvements of reliabilities from integrating selected WGS SNPs being an additional bonus. Since genetic progress by selection is linearly related to accuracy of genetic evaluation, considering a large dairy cattle population, even a small improvement in reliability is important for breeding and production.

## Conclusion

Integrating additional selected WGS SNPs to the genotype data of 54K chip led to significant improvements of reliabilities for milk and protein, small improvements for fat and mastitis, and no improvement for fertility. Reliabilities improved along with the increase of the reference population size for all traits, and benefits from using selected WGS SNPs for milk and protein remained significant in the scenario of the largest reference population. A Bayesian four-distribution model yielded higher reliabilities than a GBLUP model for milk and protein, but extra gains in reliabilities from using selected WGS SNPs was smaller for a Bayesian four-distribution model than a GBLUP model. No significant difference was observed between considering 54K SNPs and selected WGS SNPs as one or two genetic components, except for using GBLUP models for milk.

**Acknowledgements** We greatly appreciate the editor and reviewers for the careful work and valuable comments, which are very helpful for improving the manuscript. We thank Viking Genetics, the Danish Cattle Federation and Nordic Cattle Genetic Evaluation for providing phenotypic and genomic data of Danish Jersey. We thank Xiaowei

Mao for the help on imputation and revising the manuscript. We thank Esa Mantysaari, Per Madsen and Peipei Ma for nice suggestions and discussions during data analyses. We thank Goutam Sahana and Bernt Guldbrandtsen for helping on preparing genotype and pedigree data. This work was supported by the Genomic Selection in Animals and Plants (GenSAP) research project financed by the Danish Council of Strategic Research (Aarhus, Denmark). The first author acknowledges the scholarship provided by the China Scholarship Council (CSC).

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## References

Berry DP, Coffey MP, Pryce JE, De Haas Y, Løvendahl P, Krattenmacher N et al. (2014) International genetic evaluations for feed intake in dairy cattle through the collation of data from multiple sources. J Dairy Sci 97(6):3894–3905

Boichard D, Chung H, Dassonneville R, David X, Eggen A, Fritz S et al. (2012) Design of a bovine low-density SNP array optimized for imputation. PLoS ONE 7(3):e34130

Boichard D, Boussaha M, Capitan A, Rocha D, Hozé C, Sanchez MP et al. (2018) Experience from large scale use of the EuroGenomics custom SNP chip in cattle. Proceedings of the World Congress on Genetics Applied to Livestock Production, Vol Molecular Genetics 4:675

Boussaha M, Esquerré D, Barbieri J, Djari A, Pinton A, Letaief R et al. (2015) Genome-wide study of structural variants in bovine Holstein, Montbeliarde and Normande dairy breeds. PLos ONE 10 (8):e0135931

Brøndum RF, Su G, Lund MS, Bowman PJ, Goddard ME, Hayes BJ (2012) Genome position specific priors for genomic prediction. BMC Genom 13(1):543

Brøndum RF, Su G, Janss L, Sahana G, Guldbrandtsen B, Boichard D et al. (2015) Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. J Dairy Sci 98(6):4107–4116

Buch LH, Kargo M, Berg P, Lassen J, Sørensen AC et al. (2012) The value of cows in reference populations for genomic selection of new functional traits. Animal 6(6):880–886

Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3(10):e3395

Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF et al. (2014) Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 46(8):858–867

de Roos APW, Hayes BJ, Spelman RJ, Goddard ME (2008) Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle. Genetics 179(3):1503–1512

Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM (2012) Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci 95(7):4114–4129

Fragomeni BO, Lourenco DA, Masuda Y, Legarra A, Misztal I (2017) Incorporation of causative quantitative trait nucleotides in single-step GBLUP. Genet Sel Evol 49(1):59

Gao H, Su G, Janss L, Zhang Y, Lund MS (2013) Model comparison on genomic predictions using high-density markers for different groups of bulls in the Nordic Holstein population. J Dairy Sci 96 (7):4678–4687

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92(2):433–443

Iheshiulor OO, Woolliams JA, Yu X, Wellmann R, Meuwissen TH (2016) Within-and across-breed genomic prediction using whole-genome sequence and single nucleotide polymorphism panels. Genet Sel Evol 48(1):15

Karaman E, Cheng H, Firat MZ, Garrick DJ, Fernando RL (2016) An upper bound for accuracy of prediction using GBLUP. PLoS ONE 11(8):e0161054

Liu A, Wang Y, Sahana G, Zhang Q, Liu L, Lund MS et al. (2017) Genome-wide association studies for female fertility traits in Chinese and Nordic Holsteins. Sci Rep 7(1):8487

Long N, Gianola D, Rosa GJ, Weigel KA, Avendano S (2007) Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. J Anim Breed Genet 124(6):377–389

Lund MS, de Roos APW, de Vries AG, Druet T, Ducrocq V, Fritz S et al. (2011) A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet Sel Evol 43(1):43

Lund MS, Su G, Janss L, Guldbrandtsen B, Brøndum RF (2014) Genomic evaluation of cattle in a multi-breed context. Livest Sci 166:101–110

Lund MS, van den Berg I, Ma P, Brøndum RF, Su G (2016) How to improve genomic predictions in small dairy cattle populations. Animal 10(6):1042–1049

MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ (2016) Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genom 17(1):144

Madsen P, Jensen J (2012) A User's Guide to DMU. Version 6, release 5.1

Mao X, Sahana G, de Koning DJ, Guldbrandtsen B (2016) Genome-wide association studies of growth traits in three dairy cattle breeds using whole-genome sequence data. J Anim Sci 94 (4):1426–1437

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A et al. (2016) The ensembl variant effect predictor Genome Biol 17 (1):122

Michot P, Chahory S, Marete A, Grohs C, Dagios D, Donzel E (2016) A reverse genetic approach identifies an ancient frameshift mutation in RP1 causing recessive progressive retinal degeneration in European cattle breeds. Genet Sel Evol 48(1):56

Pérez-Enciso M, Rincón JC, Legarra A (2015) Sequence-vs. chip-assisted genomic selection: accurate biological information is advised. Genet Sel Evol 47(1):43

Sargolzaei M, Chesnais JP, Schenkel FS (2014) A new approach for efficient genotype imputation using information from relatives. BMC Genom 15(1):478

Su G, Brøndum RF, Ma P, Guldbrandtsen B, Aamand GP, Lund MS (2012a) Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. J Dairy Sci 95(8):4657–4665

Su G, Madsen P, Nielsen US, Mäntysaari EA, Aamand GP, Christensen OF et al. (2012b) Genomic prediction for Nordic Red Cattle using one-step and selection index blending. J Dairy Sci 95 (2):909–917

Su G, Christensen OF, Janss L, Lund MS (2014) Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. J Dairy Sci 97(10):6547–6559

Su G, Ma P, Nielsen US, Aamand GP, Wiggans G, Guldbrandtsen B et al. (2016) Sharing reference data and including cows in the reference population improve genomic predictions in Danish Jersey. Animal 10(6):1067–1075

Uimari P, Mäntysaari EA (1993) Repeatability and bias of estimated breeding values for dairy bulls and bull dams calculated from animal model evaluations. Anim Sci 57(2):175–182

van Binsbergen R, Calus MP, Bink MC, Eeuwijk FA, Schrooten C, Veerkamp RF (2015) Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol 47(1):71

van den Berg I, Boichard D, Guldbrandtsen B, Lund MS (2016a) Using sequence variants in linkage disequilibrium with causative mutations to improve across-breed prediction in dairy cattle: a simulation study. G3 6(8):2553–2561

van den Berg I, Boichard D, Lund MS (2016b) Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. Genet Sel Evol 48(1):83

van den Berg I, Bowman PJ, MacLeod IM, Hayes BJ, Wang T, Bolormaa S et al. (2017) Multi-breed genomic prediction using Bayes R with sequence data and dropping variants with a small effect. Genet Sel Evol 49(1):70

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91(11):4414–4423

VanRaden PM, Null DJ, Sargolzaei M, Wiggans GR, Tooker ME, Cole JB et al. (2013) Genomic imputation and evaluation using high-density Holstein genotypes. J Dairy Sci 96(1):668–678

Visscher PM, Macgregor S, Benyamin B, Zhu G, Gordon S, Medland S et al. (2007) Genome partitioning of genetic variation for height from 11,214 sibling pairs. Am J Hum Genet 81(5):1104–1110

Wiggans GR, Su G, Cooper TA, Nielsen US, Aamand GP, Guldbrandtsen B et al. (2015) Improving accuracy of Jersey genomic evaluations in the United States and Denmark by sharing reference population bulls. J Dairy Sci 98(5):3508–3513

Zhang Z, Liu J, Ding X, Bijma P, de Koning DJ, Zhang Q (2010) Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS ONE 5(9): e12648