Improving the accuracy of genomic predictions in numerically small populations

Sikkerhet ved genomisk prediksjon i små populasjoner

Philosophiae Doctor (PhD) Thesis

Oscar Okechukwu Michael Iheshiulor

Department of Animal and Aquacultural Sciences

Faculty of Veterinary Medicine and Biosciences

Norwegian University of Life Sciences

Ås 2016

Thesis number 2016:20

ISSN 1894-6402

ISBN 978-82-575-1349-8

PhD Supervisors

Prof. Theo H.E. Meuwissen
Department of Animal and Aquacultural Sciences
Norwegian University of Life Sciences
P.O. Box 5003, N-1432 Ås
Norway

Prof. John A. Woolliams
The Roslin Institute (Edinburgh) and Royal (DICK) School of Veterinary Studies
University of Edinburgh
EH25 9RG, Midlothian Easter Bush Campus
Scotland, United Kingdom

PhD Evaluation Committee

Prof. Jörn Bennewitz
Institute of Animal Husbandry and Animal Breeding
University of Hohenheim
D-70593 Stuttgart
Germany

Dr. Mario P.L. Calus
Animal Breeding and Genomic Centre
Wageningen University
Box 338, Wageningen
The Netherlands

Prof. Gunnar Klemetsdal
Department of Animal and Aquacultural Sciences
Norwegian University of Life Sciences
P.O. Box 5003, N-1432 Ås
Norway

ACKNOWLEDGEMENTS

This research work received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 289592 - Gene2Farm. Also acknowledged is Geno SA for providing the real data used in this thesis.

I would like to express my profound gratitude to my supervisors, Prof. Theo H.E. Meuwissen and Prof. John A. Woolliams for their guidance, support, and advice all through the period of this PhD study. Theo, I immensely appreciate your time, efforts and for being there whenever I needed you. John, it was nice visiting you in Edinburgh and thanks for your hospitality. I am also very grateful for all your input and thoughtful discussions. To you both, I must say thanks for inspiring me by your positive and constructive feedbacks.

I would also like to express my gratitude to all the administrative staff at IHA (including Inger Schult and Anne Golten who are retired) for their assistance especially during my early days in Norway. To my colleagues and friends in the Animal Breeding and Genetics Group, thanks for your friendship, encouragements and unfailing willingness to help whenever I knocked at your door.

To Assoc. Prof. Tormod Ådnøy, Dr. Jørgen Ødegård, Binyam Dagnachew, Solomon Antwi Boison, Xijiang Yu, Kahsay Nirea, Tu Luan, Nsa Eyo Dada, Tesfaye Kebede Belay, Gebreyohans Gebregiwergis, Borghild Hillestad, Cecilie Ødegård, Kristine Hov Martinsen, Sini Wallen, Katrine Haugaard, I say bravo for all your support and goodwill. Each one of you made this doctoral journey less cumbersome for me ☺

Finally, I wish to express my depth of appreciation to my parents, Mr. and Mrs. Hilary Alikeh Iheshiulor, for their constant encouragements and their self-sacrifices in allowing me to pursue my education. I also appreciate the encouragements from my mother in-law, Mrs. J.C. Onuma-Eleanya, and my siblings- Thankgod, Happiness, Chijioke, Ekene, Uzochi, Chinonye, Ugochi. To my beloved wife- Marvellous Onuma-Kalu and our dear son- Michael, thank you for your love, support, motivation, understanding and corporation. "You both rock!"

To God be the glory!

Ås, January, 2015

Oscar Okechukwu Michael Iheshiulor

TABLE OF CONTENTS

Papers I-III have individual page number

SUMMARY

Genomic selection (GS) is increasingly being implemented in livestock, especially in the dairy cattle sector. While dairy cattle populations such as Holsteins have benefitted greatly from GS in the accuracy of evaluation due to their large reference population size and low effective population size (Ne), the impact of GS on numerically small breeds, sometimes with small reference population size and larger Ne, is much less. The overall aim of this research work was to explore strategies to improve the accuracy of genomic prediction in numerically small cattle breeds.

Firstly, we conducted a computer simulation in order to assess how much predictive ability is gained by using WGS data under varied QTL density (45 or 132 QTL/Morgan) and heritability (0.07 or 0.30) compared to different SNP densities with emphasis on diverged dairy breeds of small populations with large Ne (>100). Also assessed was the relative performance of a non-variable selection method (SNP-BLUP) and a variable selection method (MixP). The use of WGS data for within-population predictions resulted in small to large increases in accuracy for lowly to moderately heritable traits when compared to the SNP densities. Depending on the heritability, marker and QTL density, the observed increase in accuracy was up to 13%. In comparison to the lowest simulated marker density, the increase was as much as 24-31%. The advantage of WGS data was more pronounced (7-92% increase in accuracy depending on the heritability, marker and QTL density, and time of divergence between populations) with a combined reference population across populations and when using MixP. While MixP outperformed SNP-BLUP at 45 QTL/Morgan, SNP-BLUP was as good as MixP when QTL density increased to 132 QTL/Morgan.

Secondly, we evaluated an iterative method (referred to as GBC) that combines relationship information using the G-BLUP approach and LD between QTL and neighbouring SNPs using the BayesC approach for genomic prediction. The performance of GBC was compared to other

evaluation methods. Two datasets were utilized for the analysis: 1) imputed medium-density (50k; MD) SNP dataset based on Illumina Bovine50k BeadChip, containing 48,249 SNPs and 3,244 records; and 2) imputed high-density (777k; HD) SNP dataset originating from the Illumina BovineHD chip and containing 539,665 SNPs and 3,164 records. With the MD SNP dataset, GBC showed an advantage over G-BLUP for all traits, while in comparison to MixP, accuracy was slightly lower. With the HD SNP dataset, GBC also performed better than G-BLUP and slightly below that of MixP except for fat yield where it gave higher prediction accuracy than both methods. The results show that incorporating aspects of G-BLUP and BayesC in a single model can improve accuracy of genomic prediction over the commonly used method: G-BLUP. On the other hand, MixP showed higher accuracies than G-BLUP for all traits studied and in most cases slightly higher than GBC.

Thirdly, we proposed and evaluated an approach to absorb phenotypic information on large numbers of ungenotyped animals into the mixed model equations of genotyped animals so that all information can be utilized in predictions. Predictions were then done using DYD (daughter yield deviations) or the resulting pseudo-records from absorption as response variables. These pseudo-records were calculated for the genotyped animals and their (ungenotyped) ancestors. The ungenotyped ancestors were included in the analysis by calculating their genotype probabilities. Two datasets were used for the analysis: 1) DYD dataset, i.e. a combination of DYD and genotype of 3,244 progeny-tested bulls; and 2) Absorbed dataset, i.e. a combination of absorbed records and genotype probabilities of 20,918 animals. With DYD as response variable, forward prediction accuracies ranged from 0.427 to 0.664 across the traits and evaluation methods. With absorbed records as response variable, accuracies ranged from 0.429 to 0.667 across the traits and evaluation methods. Comparison of performance between DYD and the Absorbed dataset showed that differences in accuracy between both datasets were not statistically significant, but were on average slightly higher for the Absorbed dataset for A-

BLUP, whilst the opposite was found for G-BLUP, GBC, and SS-GBLUP. In terms of bias, predictions based on the Absorbed dataset were generally less biased.

SAMMENDRAG

Genomisk seleksjon (GS) har i økende grad blitt implementert innen husdyravlen, særlig innen avl av melkefe. Populasjoner som Holstein har hatt gunstig effekt av GS gjennom bruk av data fra store referansepopulasjoner, kombinert med liten effektiv populasjonsstørrelse (Ne). På den annen side har den gunstige effekten av GS vært langt mindre uttalt for mindre tallrike populasjoner, men som potensielt kan ha større Ne enn Holstein. Hovedmålet med dette forskningsarbeidet var derfor å undersøke ulike strategier for å øke sikkerhet ved genomisk prediksjon for antallsmessig små storfepopulasjoner.

Det første arbeidet var en simuleringsstudie for å undersøke muligheter for å øke prediksjonsevne ved bruk av helgenom-data ved varierende QTL tettheter (45 eller 132 QTL/Morgan) og ulike arvegrader (0.07 eller 0.30) sammenliknet med bruk av SNP markører med varierende tetthet. Fokus var på divergerende melkeku-populasjoner med stor Ne (>100). Den prediktive evnen til enklere GS modeller uten variabelseleksjon (SNP-BLUP) ble sammenliknet med variabelseleksjons-modeller (MixP). Bruk av helgenom-data til innen-populasjon prediksjon ga liten til betydelig økning i sikkerhet for lav- til høyarvelige egenskaper sammenlignet med SNP tettheter. Avhengig av arvegrad, markør- og QTL tetthet, økte sikkerheten opp til 13%. Sammenliknet med de laveste markørtetthetene var økningen i sikkerhet ved bruk av helsekvens så mye som 24-31%. Fordelen ved helgenom-data var mest uttalt ved bruk av referanse-data over populasjoner basert på MixP modellen (7-92% økning i sikkerhet avhengig av markør- og QTL-tetthet, og grad av divergens mellom populasjonene). MixP modellen var bedre enn SNP-BLUP ved 45 QTL/Morgan, mens SNP-BLUP var like god som MixP dersom QTL tetthet økte til 132 QTL/Morgan.

I det andre arbeidet ble en iterativ metode (kalt GBC) evaluert. Metoden kombinerer slektskaps-informasjon (via G-BLUP) og LD mellom QTL and SNP loci i nærheten (med BayesC

tilnærming) i genomisk prediksjon uten anvendelse av Monte Carlo sampling metodikk. Prediktiv evne til GBC ble sammenliknet med andre modeller. To datasett ble brukt i analysen: 1) Imputert medium tetthet (50k, MD) SNP datasett basert på Illumina Bovine50k BeadChip, inneholdende 48,249 SNP loci og 3,244 fenotyper; og 2) Imputert høy-tetthets (777k; HD) SNP datasett basert på Illumina BovineHD chip, inneholdende 539,665 SNP loci og 3,164 fenotyper. Basert på MD SNP datasettet, ga GBC bedre sikkerhet enn G-BLUP for alle egenskaper, men noe lavere sikkerheter enn MixP. Basert på HD SNP datasettet, ga GBC også bedre sikkerheter enn G-BLUP, og litt lavere enn MixP unntatt for kg fett, der GBC ga bedre sikkerhet enn begge de to andre modellene. Resultatene viser at å inkorporere egenskaper ved modellene G-BLUP og BayesC i en enkelt modell (GBC) kan øke sikkerhet ved genomisk prediksjon over den mest brukte metoden: G-BLUP. På den annen side hadde MixP høyere sikkerhet enn G-BLUP for alle egenskaper som ble studert, og i de fleste tilfeller også noe høyere sikkerhet enn GBC.

In en tredje studie ble det foreslått og evaluert en ny metode for å absorbere fenotypisk informasjon på store antall ikke-genotyperte dyr inn i likningene for genotyperte dyr, slik at all informasjon kan bli utnyttet i genomiske prediksjoner. Genetiske effekter av fedre ble predikert, der enten DYD (døtrenes gjennomsnittlige fenotypiske prestasjoner korrigert for ikke-genetiske effekter og effekt av mødre) eller pseudo-observasjoner fra absorpsjonen av ikke-genotyperte dyr ble brukt som responsvariable. Pseudo-observasjoner ble beregnet for genotyperte dyr og deres (ikke-genotyperte) avkom. Ikke-genotyperte stamfedre ble inkludert i analysen gjennom beregned genotype-sannsynligheter. To datasett ble brukt i analysen: DYD datasett, dvs. en kombinasjon av DYD og genotyper for 3,244 avkomsgranskede okser; og 2) Absorbert datasett, dvs. en kombinasjon av absorberte pseudo-observasjoner og genotype sannsynligheter for 20,918 dyr. Med DYD som responsvariabel og genotyper av 3,244 avkomsgranskede okser varierte sikkerheten basert på fremtidige observasjoner fra 0.427 til 0.664 over egenskaper og evalueringsmetoder. Med absorberte fenotyper som responsvariabel varierte sikkerheten fra

0.429 til 0.667 over egenskaper og evalueringsmetoder. Forskjellene mellom DYD og det absorberte datasettet var ikke signifikante, men sikkerheten var i gjennomsnitt noe høyere for det absorberte datasettet for klassiske avlsverdiberegninger uten bruk av genetiske markører, men noe lavere for G-BLUP, GBC og SS-GBLUP (single-step G-BLUP). Prediksjoner basert på det absorberte datasettet hadde generelt mindre bias enn prediksjoner basert på DYD.

ABBREVIATIONS

A-matrix – Pedigree-based Relationship Matrix

CNV – Copy Number Variation

DGAT1 – Diacylglycerol O-acyltransferase 1

DNA - Deoxyribonucleic Acid

DYD – Daughter Yield Deviations

G-BLUP – Genomic Best Linear Unbiased Prediction

GEBV – Genomic Estimated Breeding Value

G-matrix – Genomic Relationship Matrix

GS – Genomic Selection

HD – High Density (777k)

INDEL – Insertion and Deletion

LD – Linkage Disequilibrium

LE-MAS – Linkage Equilibrium Based Marker Assisted Selection

MCMC – Markov Chain Monte Carlo

Ne – Effective Population Size

QTL – Quantitative Trait Loci

RR-BLUP – Ridge Regression Best Linear Unbiased Prediction

SNP – Single Nucleotide Polymorphism

SS-GBLUP – Single-step Genomic Best Linear Unbiased Prediction

TS – Traditional Selection

WGS – Whole Genome Sequence

LIST OF PAPERS

This thesis is based on the following manuscripts, which will be referred to in the text by their Roman numerals.

I. O. O. M. Iheshiulor, J. A. Woolliams, X. Yu, R. Wellmann, T. H. E. Meuwissen. **Within- and Across-breed Genomic Prediction Using Whole-genome Sequence and Single Nucleotide Polymorphism Panels**

(Submitted to Genetic Selection Evolution)

II. O. O. M. Iheshiulor, J. A. Woolliams, M. Svendsen, T. Solberg, and T. H. E. Meuwissen. **Comparison of Genomic Prediction Methods Using Medium- and High-density Single-nucleotide Polymorphism Datasets in Norwegian Red Cattle**

(Manuscript)

III. O. O. M. Iheshiulor, J. A. Woolliams, and T. H. E. Meuwissen. **The Absorption of Large numbers of Ungenotyped Descendants in Genomic Predictions**

(Manuscript)

1. GENERAL INTRODUCTION

1.1. Genomic Selection

Natural variations (environmental or at genome level) existing within livestock species, within a breed and within a population formed the basis for animal breeding and genetics (Eggen, 2012). For decades, breeding value estimation of selection candidates depended on phenotype and pedigree (TS) without emphasis on the underlying genes acting on the trait. TS was successful especially for easy-to-measure production traits with high heritability (e.g. milk yield) and in animals having short generation interval with large numbers of offspring (e.g. chicken, fish). However, for traits with very low or low heritability (e.g. fertility), not easy-to-measure traits (e.g. disease resistance), sex limited traits (e.g. calving ability), and animals having long generation interval, genetic progress is slow (Goddard and Hayes, 2007; Eggen, 2012; Meuwissen et al., 2013).

Animal breeding is today being revolutionised by Genomic Selection (GS). A selection method that allows selection of breeding animals on the basis of genetic value predicted with genetic markers covering the genome (Meuwissen et al., 2001). This advancement has been propelled by availability of statistical methodologies, discovery of large numbers of SNPs as well as availability of affordable dense genome-wide marker panels (Goddard et al., 2011; Meuwissen et al., 2013). On the other hand, its wide acceptance is due to its potential to improve selection accuracy while decreasing infrastructural costs, reducing generation intervals, and exploiting new sources of polymorphisms (Dekkers, 2004; Schaeffer, 2006) and thereby resulting in faster genetic progress. The unique features of GS that distinguishes it from other type of marker-assisted selection (e.g. LE-MAS) are: 1) it's based on markers covering the whole genome with an aim to use the best estimate of the effect of each marker and thereby resulting in the best estimate of the breeding value of individuals; 2) potentially all genetic variance are explained by markers; 3) assumes all QTL effects are explained by a set of markers that are in LD with

the QTL thereby maximizing the proportion of genetic variance explained by the SNP; 4) phenotypes of selection candidates (validation population) are not needed; 5) GEBV is the sum of SNP effects across the entire genome; (Goddard and Hayes, 2007; Meuwissen et al., 2013).

GS was pioneered in the dairy cattle sector because of its potential to achieve high accuracy for non-phenotyped bulls, thereby reducing generation intervals through reducing the need for progeny testing. This has been implemented using panels of SNPs distributed over the genome and various commercial bovine SNP chips are available with densities ranging from 3k to 777k (HD). So far, results of several GS studies in livestock show that GEBVs can be significantly more accurate than that from TS (see review, Meuwissen et al. (2013)). Dairy cattle populations such as Holsteins have benefitted greatly from GS in the accuracy of evaluation due to their large reference population and low Ne, but the impact of GS on numerically small breeds (e.g. Norwegian Red), sometimes with larger Ne, is much less. This is often exacerbated by a greater emphasis on functional traits in these breeds and typically, such traits have lower heritabilities than production traits.

1.2. Factors Driving the Accuracy of Genomic Predictions

GS enables the selection of young animals thereby reducing generation intervals especially in cattle breeding. Hence, accuracy of breeding values is of key importance for successful application of GS. Many factors influence the accuracy of GS. They are: 1) size of the reference population; 2) level of LD between markers and QTL (related to Ne i.e. small Ne, high LD and vice versa); 3) marker density; 4) level of relationship between reference and validation population; 5) heritability of the trait under investigation; 6) genetic architecture of the trait; and 7) prediction method; (Meuwissen et al., 2001; Calus et al., 2008; Solberg et al., 2008; Goddard, 2009; Hayes et al., 2009; Luan et al., 2009; Meuwissen, 2009; Coster et al., 2010; Daetwyler et al., 2010; Habier et al., 2010; Wientjes et al., 2013). The aforementioned authors and a lot of others have shown that improving most of these factors results in increased accuracy

of GS. It is also important to note that these factors differ between within- and across- breed predictions. Across-breed predictions suffer much more from low across population LD than within-breed predictions. The reasons why across-breed predictions are less accurate are explained in detail in the next section.

1.3. Across- and Multi- Breed Prediction

One of the major factors affecting accuracy of GS is the size of reference population (Hayes et al., 2009; Daetwyler et al., 2010). Accuracy increases with increasing numbers of reference animals (Goddard, 2009). Numerically small dairy populations are faced with the problem of small reference populations. To ameliorate the problem of insufficient reference population animals, pooling of related breeds has been proposed. However it has not been very successful due to: 1) non-persistent SNP-QTL LD across populations; 2) low genetic relationships between populations; 3) difference in allele substitution effects across populations which results in difference in genetic variance; 4) QTL segregating in one population may not be segregation in the other population, thereby resulting in differences in the genetic variance explained by that QTL across populations; (De Roos et al., 2009; Goddard and Hayes, 2009; Hayes et al., 2009; Habier et al., 2010; Clark et al., 2012; Wientjes et al., 2013; Wientjes et al., 2015). In studies where prediction equations have been setup in one population and then used to predict GEBVs of animals in another population, zero or close to zero accuracies have been reported. While in the case of multi-breed reference population using 50k or 777k, only slight improvements in prediction accuracy have been reported (e.g. (Erbe et al., 2012; Zhou et al., 2013; Hoze et al., 2014; Zhou et al., 2014a; Zhou et al., 2014b)). This has resulted in GS being focused more on within breeds/populations predictions. The use of WGS data for genomic prediction in these populations maybe a way to improve accuracy since it will contain all possible variants including the causal mutations.

1.4. Whole Genome Sequence Dataset

To date, the HD (777k) SNP panel is the most dense panel in the dairy cattle sector, however much increase in prediction accuracy has not been observed in the transition (by either genotyping or imputation) from 50k to HD. Solberg et al. (2011) and Su et al. (2012) compared the use of HD to the 50k SNP panel and reported small or no gain in accuracy or gains for only some traits. Hence, we need to move beyond SNPs to capturing all possible variations (rare or common) in a population. With the present trends of advancement in next generation sequencing technologies as well as decreasing cost of DNA re-sequencing, WGS data on large numbers of individuals maybe within reach. Currently the 1000 genome bull project provides a platform for possible imputation of sequence data on (densely) genotyped animals (Hayes et al., 2012). Its availability provides new opportunities for GS especially in the area of across- or multi- breed predictions. WGS data differs fundamentally from current dense SNP-panel data in that the causative mutations are included, and offers more than just SNPs, i.e. also Indels, CNV and other polymorphisms may be included. If all individuals in a population could be sequenced, all the genomic variants (including causal mutations) in the population could be captured which invariably explains more of the variation. Hence, predictions would no longer have to completely depend on LD between SNPs and QTL and this could lead to increased accuracy of GS (Meuwissen and Goddard, 2010; Hayes et al., 2014). In situations of incomplete LD, and considering the fact that marker density alongside LD between QTL and SNP has an important effect on the accuracy of GS, use of WGS data could improve GS since it does not need to rely on LD between flanking markers and QTL thereby providing better signals even in across diverged population prediction (Calus et al., 2008; De Roos et al., 2009; Goddard, 2009; Harris and Johnson, 2010). In the case of across-breed predictions, the use of WGS data would reduce or remove reliance on SNP-QTL associations, which may not persist across the breeds being evaluated (Hayes et al., 2014).

1.5. Genomic Prediction Methods

Unlike the traditional animal breeding where consensus exist on methods for estimating breeding values, the era of GS is confronted with a variety of methods and no consensus exist on which is the best approach. GS methods can be broadly classified into two groups: variable and non-variable selection methods (Daetwyler et al., 2010). G-BLUP which has been shown be equivalent to RR-BLUP (Meuwissen et al., 2001; Habier et al., 2007; VanRaden, 2008) is a typical example of the non- variable selection method while the Bayesian methods (e.g. BayesA/B/C etc.) and others make up the variable selection methods. A major difference between methods lies in their assumptions about the marker effects, and details of each method have been reviewed by (Neves et al., 2012; De Los Campos et al., 2013). While simulation studies have shown the variable selection methods to have an edge over the non-variable selection methods, much difference has not been observed in empirical studies except in cases where major genes control the trait (e.g. the DGAT1 gene in bovine, which controls fat percentage). As shown by Daetwyler et al. (2010) the numbers of QTL in relation to the structure of the genome plays a major role in this discrepancy.

Presently, a good number of GS methods focus on genotyped individuals and involves multiple-step procedures in estimation of breeding values (Legarra et al., 2009; Christensen and Lund, 2010; Meuwissen et al., 2011). However, it is not common that all individuals in a given population are genotyped. The need to accommodate both genotyped and un-genotyped individuals led to the single-step method that combines pedigree and genomic information via a special relationship matrix called **H**-matrix (i.e. a combination of the **A**-matrix and G-matrix) (Legarra et al., 2009; Christensen and Lund, 2010). Legarra et al. (2014) have reviewed the performances so far as well as the drawbacks of single-step approach in real data.

Generally, in practical implementation of GS, G-BLUP is used for estimation of breeding values because of its simplicity and low computational demands. However, methods that are more efficient are needed in-order to take advantage of: 1) the different information sources in genomic data; 2) thousands of phenotyped but ungenotyped animals in the population; 3) multi-breed reference populations; and 4) WGS dataset since it would contain much more variants than the present SNP dataset.

2. AIM AND OUTLINE OF THIS THESIS

The overall aim of this thesis was to explore strategies to improve the accuracy of genomic prediction in numerically small cattle breeds. The specific objectives were to:

- Evaluate the benefit of WGS data relative to different SNP densities under varied genetic architectures and genetic models for the estimation of breeding values in small populations
- Investigate the effect of simultaneously exploiting relationship information and linkage disequilibrium on accuracy of genomic prediction compared to focusing on relationship information or linkage disequilibrium
- Evaluate how to utilize ungenotyped animals in the population for genomic prediction

Following the specific objectives:

Paper I assessed how much predictive ability is gained by using WGS data under varied QTL density and heritability compared to different SNP densities with emphasis on diverged dairy breeds of small populations with large Ne (>100). Also assessed was the relative performance of a non-variable selection method and a variable selection method.

Paper II evaluated an iterative method (referred to as GBC) that combines relationship information using the G-BLUP approach and LD between QTL and neighbouring SNPs using the BayesC approach for genomic prediction. GBC performance was compared to other evaluation methods using 50k and HD SNP panels.

Paper III proposed and evaluated an approach to absorb phenotypic information on large numbers of ungenotyped animals into mixed model equations of genotyped animals so that all information can be utilized in predictions. Predictions were then done using either DYD's or the resulting pseudo-records from absorption as response variables.

3. BRIEF SUMMARY OF PAPERS

3.1. PAPER I

Within- and Across-breed Genomic Prediction Using Whole-genome Sequence and Single Nucleotide Polymorphism Panels

With decreasing costs, and current advances in next generation sequencing technologies, WGS data on large number of individuals is within reach. Its availability provides new opportunities for GS and these need to be explored. Hence, this simulation study investigated how much predictive ability is gained by using WGS data under varied QTL density (45 or 132 QTL/Morgan) and heritability (0.07 or 0.30) compared to different SNP densities with emphasis on diverged dairy cattle breeds of small populations. Relative performance of SNP-BLUP, and MixP (a method that fits a mixture of two normal distributions for SNP effects using the Pareto principle) was also evaluated. Genomic predictions were based on within and across population predictions as well as using multi-breed reference populations.

Main results:

- WGS data for within-population genomic prediction resulted in small to large increases in accuracy for lowly - moderately heritable traits.
- Up to 13% increase in accuracy was observed depending on the heritability, marker and QTL density. In comparison to the lowest simulated marker density, the increase was as much as 24-31%.
- The advantage of WGS data was more pronounced with reference populations combined across breeds and when using MixP.
- While MixP outperformed SNP-BLUP at 45 QTL/Morgan, SNP-BLUP was as good as MixP when QTL density increased to 132 QTL/Morgan.

Conclusion

Genomic prediction in numerically small cattle populations could benefit from a combination of WGS data, multi-breed reference populations, and variable selection methods.

3.2. PAPER II:

Comparison of Genomic Prediction Methods Using Medium-density and High-density Single-nucleotide Polymorphism Datasets in Norwegian Red Cattle

GS enables the selection of young animals thereby reducing generation intervals especially in cattle breeding programs. Hence, accuracy of breeding values is of key importance for successful application of GS. Presently, many methods are available for genomic prediction, and they can be broadly classified into two groups: variable and non-variable selection methods. Both methods are presently treated as distinct approaches. Thus, this study evaluated an iterative method (called GBC) that incorporates aspects of both methods and compares its performance to A-BLUP, G-BLUP, and MixP. Prediction methods were evaluated using Imputed 50k and HD SNP dataset.

Main results:

- With the 50k SNP dataset, GBC was more accurate than G-BLUP for all traits while in comparison to MixP, it was slightly less accurate.
- With the HD SNP dataset, GBC also performed better than G-BLUP and slightly better than MixP, except for fat yield where it performed better than both methods.
- MixP outperformed G-BLUP in all traits studied and was slightly better than GBC in most cases.
- A-BLUP, which is pedigree-based, yielded significantly less accurate results in comparison to the genomic methods.
- Generally, the results show that incorporating aspects of both non-variable and variable selection methods can improve accuracy of genomic prediction over commonly used method, G-BLUP.

Conclusion

GBC is quite a flexible tool in the sense that it simultaneously incorporates aspects of variable and non-variable models, thereby exploiting family relationship while also accounting for genes of large effects. MixP on the hand seems to strike a good balance between genes of large and small effects using the Pareto principle. The application of both methods in genomic prediction merits further exploration.

3.3. PAPER III

The Absorption of Large Numbers of Ungenotyped Descendants in Genomic Predictions

The utilization of all available information could lead to more accurate and less biased predictions. SS-GBLUP exploits all available information, however, its extension to the variable selection models is not straightforward. Thus, we evaluated an absorption approach that absorbs phenotypic information of large numbers of ungenotyped animals into the mixed model equations of genotyped animals in-order to enable the utilization of all information in variable selection methods of genomic prediction. Various prediction methods (including variable selection method) were implemented using DYD's or the resulting pseudo-records from absorption as response variables.

Main results:

- With DYD dataset, the evaluation methods performed as follows: A-BLUP (0.427 – 0.491)< G-BLUP (0.575 – 0.652)< GBC (0.580 – 0.664).
- With Absorbed dataset, the evaluation methods performed as follows: A-BLUP (0.429 – 0.511)< SS-GBLUP (0.558 – 0.656)<GBC (0.561 – 0.665)<G-BLUP (0.565 – 0.667).
- Comparison of performance between using DYD and Absorbed dataset showed that differences in accuracy between the both datasets were not statistically significant, but were on average slightly higher for the Absorbed dataset for A-BLUP, whilst the opposite was found for G-BLUP, GBC, and SS-GBLUP. In terms of bias, predictions based on the Absorbed dataset were generally less biased.

Conclusion

An approach to absorb phenotypic information on large numbers of ungenotyped animals into mixed model equations of genotyped animals was proposed and evaluated. On the Absorbed

dataset obtained from absorption, the methods performed as follows: A-BLUP<SS-GBLUP<GBC<G-BLUP. Implementation of a variable selection method for genomic prediction on the Absorbed dataset did not show an extra advantage. Overall, the utilization of all available information led to less biased predictions.

4. GENERAL DISCUSSION

In recent years, GS has been implemented in livestock, especially in the dairy cattle sector. While dairy cattle populations such as Holsteins have benefitted greatly from GS in the accuracy of evaluation due to their large reference population size and low Ne, the impact of GS on numerically small breeds, sometimes with small reference population size and larger Ne, is much less. This thesis explored possibilities to increase accuracy of genomic prediction in numerically small dairy cattle breeds. However, the findings can be applied to other breeds or populations. In **Paper I,** a combination of WGS data, multi-breed reference populations, and variable selection methods, were found to give substantial increases in accuracy of genomic prediction compared to the SNP panels. In **Paper II**, GBC (a method that incorporates aspects of G-BLUP and BayesC approaches) improved accuracy of genomic prediction over the commonly used method, G-BLUP. MixP (a method that fits a mixture of two normal distributions for SNP effects using the Pareto principle) outperformed G-BLUP for all traits studied and performed slightly better than GBC in most cases. In **Paper III**, absorption of millions of phenotypic information on ungenotyped animals into mixed model equations of genotyped animals and their ancestor did not result in an extra gain in accuracy but led to less biased predictions.

This general discussion will address in addition to the evaluated possibilities, other way(s) to increase accuracy of genomic prediction in numerically small breeds, including their possible challenges.

4.1. Maximizing the Accuracy of Genomic Prediction

Achieving a prediction accuracy close to 1 is the target for genomic prediction. This is yet to be achieved, even with the increasing marker densities (50k to HD and in the near future WGS data on large numbers of individuals will be available). For production traits (often moderately to highly heritable), prediction accuracies have not exceeded 0.84 (Pryce and Daetwyler, 2012;

Ducrocq and Wiggans, 2014), while for health and functional traits (often lower heritability), prediction accuracies are even much lower. Marker densities have increased drastically, and numbers of genotyped animals are rising sharply. Increasing marker density results in higher LD between SNPs and QTL, however, it also results in an increased number of uninformative SNPs and a linear function of the uninformative SNPs may predict random errors in the reference phenotypes (Harris and Johnson, 2010). So in order to maximize accuracy, reference population size has to be increased considerably. This is quite important because having large numbers of reference animals would mean having sufficient phenotypic data to detect causative mutations and to distinguish their effects from random noise (De Roos, 2011). The Holstein breed due to their global presence is more fortunate than the numerically small breeds with respect to reference population size. And this has resulted in higher prediction accuracies than achieved by numerically small breeds. However, there is still a need for improvement in prediction accuracy in both the Holstein population and the numerically small breeds. Increased reference population size among other factors will be instrumental to improving or achieving a prediction accuracy close to 1.

Assuming that marker density is high enough to capture all genetic variance, the amount of phenotyping and genotyping needed to maximize or obtain a prediction accuracy close to 1 under varied numbers of reference population size, heritability, and Ne is shown in Figure 1. Calculations are based on theoretical expectations of Daetwyler et al. (2008) and Daetwyler et al. (2010), i.e. $r = \sqrt{Nh^2/(Nh^2 + M_e)}$, where N is the size of the reference population (i.e. phenotyped and genotyped individuals), h^2 is the heritability of the trait and M_e is the effective number of chromosome segments calculated as $M_e = 2N_eL/\ln(4N_eL)$, (Goddard, 2009), where N_e is the effective population size and L is the length of chromosome. From Figure 1, it can be seen generally that with increasing numbers of phenotyped and genotyped animals, accuracy of prediction increased too. The degree of increase in accuracy of prediction also

depended on N_e and h^2 of the trait. The scenario of N_e = 100 represents the Holstein population and N_e >100 represents the numerically small breeds. While about 5,000 phenotyped and genotyped animals are required to obtain appreciable accuracies of prediction for the moderate to highly heritable traits, 10,000 or more are required to obtain an accuracy close to 1. In the case of lowly heritable traits, a lot more phenotyped and genotyped animals are required. Such huge reference population sizes are not easy to assemble especially in numerically small breeds. Hence, the need for strategies to increase their reference population size and possibly increase accuracy of predictions. Proposed strategies are discussed in subsequent sections.

4.1.1. Multi-breed Reference Populations

The first possibility of increasing reference population size in numerically small breeds is by combining reference populations of breeds that are genetically related (De Roos et al., 2009). For instance in the Nordic countries, a combination of Norwegian Red, Swedish Red, Finnish Ayrshire, and Danish Red since they have been reported to share relationships from previous semen exchange programs (Bett et al., 2010; Olsen et al., 2011; Zhou et al., 2014a). Although combining related breeds sounds quite appealing, it has not been very successful compared to within-breed predictions when using SNP panels due to the reasons mentioned in the general introduction (section 1.3). As reported by several studies (e.g. Zhou et al. (2013); Hoze et al. (2014); Zhou et al. (2014a); Zhou et al. (2014b)) and **Paper I,** only slight or no improvement in prediction accuracy was achieved using the SNP panels. In addition, **Paper I** of this thesis under different scenarios showed that higher prediction accuracy could be obtained when using a combination of WGS data and multi-breed reference population in comparison to using the SNP panels.

WGS data differs from SNP data in the sense that it contains all genomic variants (causative mutations included) and this makes it less dependent on SNP-QTL associations, which may not

persist across-breeds. The presence of the causative mutations in WGS data enhances the possibility of picking up similar causative mutations segregating between populations and also having comparable effects (Hayes et al., 2014), while combining related breeds increases the reference population size. Hence, the observed substantial increase in accuracy of prediction. The simulated dataset in **Paper I** was based on the scaling argument of (Meuwissen, 2009; Meuwissen and Goddard, 2010), hence, only 1 chromosome of 1 Morgan was simulated. So to translate the results to real application, larger numbers of reference animals would be required. For numerically small populations, this can be achieved through combining reference populations of related breeds. Since the sequencing of thousands of animals is still not cheap, it is recommended to sequence key ancestors or the most prominent animals and then impute the rest of the population that are sparsely genotyped up to sequence (Hayes et al., 2013). Strategies for selecting individuals to be sequenced or densely genotyped can be found in Druet et al., 2014 and Xijiang et al., 2014.

4.1.2. Combined Bull and Cow Reference Populations

In dairy cattle breeding, the bulls have more impact than the cows and provide high predictive accuracy as a result of the large amount of information from their daughters averages (Calus, 2010; Jimenez-Montero et al., 2012). To this effect, GS as well as genotyping has focused more on bulls. However, the number of bulls available especially in the numerically small breeds are likely not enough to constitute a sufficiently large reference population. Possibly all available bulls have been genotyped but there is still a need to increase the reference population. The cow population is often larger than that of the bulls and thus, can be used to make up the reference population and possibly increase prediction accuracy. Inclusion of cows could also provide more information for the lowly heritable traits such as health and functional traits as well as novel traits (Calus et al., 2013a; Egger-Danner et al., 2014). However, to maximize the expected

gain from a combination of bulls and cows in the reference population, some issues have to be resolved.

Cow evaluations are much less accurate than those of progeny-tested bulls (Wiggans et al., 2011; Ding et al., 2013; Su et al., 2015). In a cow's evaluation, yield deviations are often used as the phenotypic information while in the case of the bulls, DYD's or de-regressed proofs are often used. While yield deviations are based on a weighted average of the cow's own performances adjusted for all effects other than the genetic effect, DYD's on the other hand are based on the average performance of each bull's daughters, adjusted for all fixed and non-genetic random effects of the daughters and genetic effects of their mates (VanRaden and Wiggans, 1991; Liu et al., 2004). Hence, the first issue that arises when combining bulls and cows in the reference population is how to combine both information sources. Measures deployed thus far to handle this issue includes: 1) considerably increasing the information from cows by genotyping and including a large number of cows to the bull reference population (Su et al., 2015); 2) fitting a bivariate model where the analyzed trait is measured only on a cow or a bull reference population, or recorded on both (for details see Calus et al. (2013b)); 3) adjustment of the mean and variance of cow Mendelian sampling component (i.e. EBV minus parent average) to be similar to those of bulls (for details see Wiggans et al. (2011)). The approach of Su et al. (2015) has been evaluated in Danish Jersey population, that of Calus et al. (2013b) in Holstein population from 4 countries, and that of Wiggans et al. (2011) in Holstein and Jersey population in USA. All 3 studies and others reported an increased prediction accuracy when genotyped cows were added to bulls reference population. The absorption approach in **Paper III** is also a way to utilize bulls and cows information in a single reference population. The absorbed records which results from the absorption process are based on traditional EBV's and their reliabilities and information sources are weighted accordingly. Thus, the issue of difference in information between the bulls and cows is properly accounted

for. Further investigations possibly comparing these mentioned strategies to clarify the optimal way to jointly utilize bulls and cows information in a single reference population, would be beneficial, both for increasing prediction accuracy as well as producing unbiased genomic predictions.

A second issue that might arise when using combined cow and bull reference population is double counting of contributions of the cows. This situation arises when cows are included with their phenotypes in the analyses, and at the same time contribute to the DYD of the bulls that are included in the analysis (Calus et al., 2013b). Such situation could result in biased or overestimation of predictions (Calus et al., 2013b; Su et al., 2015). As a remedy, both authors recommended that either daughters of included bulls should be removed from the data or DYD should be estimated excluding information of daughters that are included as cows in the analysis.

Although genotyping costs are decreasing, genotyping all or thousands of animals in the population is still not cheap and economically viable, and the genotyping of historical animals may be impossible, if no DNA was preserved. Therefore, the third issue that arises with including cows to the reference population is which cow(s) should be genotyped. Jimenez-Montero et al. (2012) evaluated five different female-selective genotyping strategies (i.e. random selection, two-tailed selection by yield deviations, two-tailed selection by breeding value, top yield deviation selection, and top breeding value selection) to increase the accuracy of genomic prediction in populations that have a limited number of bulls with a large number of progeny. They concluded that for small cattle populations, the two-tailed selection strategies (i.e. genotyping cows on both tails of the distribution based on preferably yield deviations but breeding values can also be used in the case of high heritability traits) were advantageous while the random selection strategy was advised for larger populations. A two-tailed selection strategy enables a better representation of the entire herd or population and removes any possible bias

that might arise because of preferential treatment (Buch et al., 2012; Calus et al., 2013b; Thomasen et al., 2014). Jimenez-Montero et al. (2013) demonstrated that the selection and inclusion of cows with high estimated breeding values or yield deviations resulted in the lowest prediction accuracy, while Dassonneville et al. (2012) demonstrated that inclusion of elite females in the reference population led to overestimated predictions for production traits due to biased phenotypes, although this may depend on the breeding value estimation method. So, in essence, it may be advantageous that the cows genotyped should be a representation of the entire population. In addition to that, it is also important that the relationship between animals in the reference population be minimized, while the relationship between animals in the reference and selection candidates should be maximized (Buch et al., 2012; Pszczola et al., 2012; Thomasen et al., 2014). Maximizing the genetic relationship between reference and selection candidates results in increased accuracy of genomic prediction (Habier et al., 2007).

An alternative to densely genotyping all selected cows, would be to densely genotype some cows while the remaining are sparsely genotyped and then imputed up to high density. This approach will lead to reduced genotyping costs and enable that more cows are genotyped. Several strategies for prioritizing animals for dense genotyping have been evaluated by Xijiang et al. (2014). They concluded that methods such as MCA and MCG, which minimize the conditional genetic variance of the target animals, using either the pedigree-based relationship matrix (MCA), or a genomic relationship matrix based on sparse marker genotypes (MCG) were optimal procedure for prioritizing animals for dense genotyping.

4.1.3. Utilization of Ungenotyped Individuals

An area of genomic prediction currently receiving much attention is the utilization of information from ungenotyped animals alongside genotyped animals in prediction. This has become of interest because: 1) it is not common that all animals in a population are genotyped; 2) there are a lot more phenotyped animals than genotyped animals; and 3) exploiting all

available information could lead to more accurate and less biased predictions since no information is lost. On the other hand, for lowly heritable traits, and numerically small populations the utilization of all available information is beneficial. SS-GBLUP enables the utilization of all available information (Legarra et al., 2009; Christensen and Lund, 2010). It simultaneously combines information of both genotyped and ungenotyped animals by integrating genomic, pedigree, and phenotype information. Quite a number of studies have shown SS-GBLUP to result in slightly higher prediction accuracy and most importantly, less biased predictions. In **Paper III**, SS-GBLUP yielded similar results to other genomic prediction methods on an absorbed dataset.

Just like SS-GBLUP, the absorption approach proposed and evaluated in **Paper III**, enables the utilization of all available information. As an extra, variable and non-variable selection methods based genomic prediction can be implemented on the resulting data from the absorption process. The absorption approach being a prelude to genetic analysis enables us to circumvent dealing with thousands or millions of records from ungenotyped animals during genetic analysis. Thus, it reduces the computational burden in the sense that it enables genetic analysis to focus only on genotyped animals and their ancestors while still making use of the all information from their descendants.

4.2. Genomic Prediction Methods

The importance of prediction methods in GS cannot be overemphasized considering that the effects of thousands of SNPs (millions in the case of WGS data) have to be estimated accurately with a much smaller number of phenotypic records. While G-BLUP, which is commonly used because of it uncomplicated nature and low computational demand assumes that the a-priori variance of SNP effects is equal, the variable selection methods (such as BayesA/B/C/R etc.) assume that most SNPs have small or zero effects and a few have large effects.

As we move towards across-breed prediction (i.e. using a multi-breed reference population) especially for the numerically small breeds and the availability of WGS data on large numbers of individuals, prediction methods would prove very useful in achieving the intended aim, which is increased prediction accuracy. In the case of across-breed prediction, relationships as well as LD is expected to weaken depending on how long populations have diverged. While in the case of using WGS data for genomic prediction, the effects of millions of variants would need to be accurately estimated. In both cases, G-BLUP is unlikely to be optimal since: 1) it focuses more on exploiting family relationships in a given population (Habier et al., 2007; Odegard and Meuwissen, 2014; Odegard et al., 2014); and 2) it's a-priori assumption of equal variance makes it difficult for a single SNP to capture the effect of a causative mutation rather the effects are distributed across many SNPs not minding whether they are informative or uninformative (Hayes et al., 2014).

Alternatively, variable selection methods are expected to be more accurate since: 1) they are much more able to utilize LD information than G-BLUP (Habier et al., 2007); and 2) their a-priori assumption of SNP effects allows that most SNPs have small or zero effects and only a few have large effects. This approach not only makes that the focus is on picking up and utilizing SNPs with large effects or the actual causative mutation but it also enables that the effects of causative mutations are not distributed across several SNPs in moderate LD with the causative mutation (Hayes et al., 2014). Studies using either a multi-breed reference population or WGS data have reported an increased prediction accuracy from variable selection methods over G-BLUP (Meuwissen and Goddard, 2010; Clark et al., 2011; Erbe et al., 2012; Hoze et al., 2014; MacLeod et al., 2014; Zhou et al., 2014a). In **Paper I,** we showed that variable selection method performed better than G-BLUP using WGS data and a multi-breed reference population.

Despite the fact that studies have shown the variable selection methods to give (slightly) higher prediction accuracy than G-BLUP, they are not commonly used in routine genetic evaluations. The reason being that most variable selection methods are based on MCMC algorithms, which makes them quite time consuming and computational demanding. Alternatives to the MCMC based variable selection methods are iterative methods such as fastBayesB (Meuwissen et al., 2009), MixP (Yu and Meuwissen, 2011), and emBayesR (Wang et al., 2015). The authors have tested the methods on both simulated and real data. All methods were reported to perform better than G-BLUP in terms of prediction accuracy and similar to BayesB and BayesR, respectively. In terms of computational time, the non-MCMC based were much faster than the MCMC based methods, however, comparable to G-BLUP. In this thesis, MixP outperformed G-BLUP in **Paper I** and **Paper II**. Results from **Paper II and Paper III** also shows that incorporating aspects of G-BLUP and BayesC into a single model improved accuracy of prediction over G-BLUP in some cases. Hence, we conclude that for routine genomic evaluations, the iterative variable selection methods can be considered, although they need to be further developed for multi-trait evaluations since routine genomic evaluations are often multi-trait.

5. CONCLUSIONS

Genomic selection offers great opportunities to further increase the rate of genetic progress in livestock and plants. Accurate predictions are essential for its successful implementation. Improving or maximizing the accuracy of genomic prediction is possible, however, its success depends on a combination of factors.

To maximize the accuracy of genomic predictions in numerically small breeds:

- Increased reference population size is crucial and this can be done by combining reference populations of genetically related breeds or including genotyped cows to the reference population. The utilization of phenotyped but ungenotyped animals is also an option.
- Higher marker density such as WGS data will be essential for the use of combined reference populations, since the across population LD between markers and QTL extends only across short distances. WGS data will remove or reduce dependencies on marker - QTL association since the causative mutations are in the data.
- Variable selection methods are highly needed in this era of increasing marker density and more animals being genotyped since G matrices hardly improve by marker densities beyond ~1000 SNPs per Morgan, whereas variable selection methods focus on causal variants or those in very high LD.

6. FURTHER RESEARCH

Paper I was based on simulated WGS data, however, soon, imputed WGS data on large numbers of individuals will be available. This large volume of data will be both computationally demanding and statistically challenging. Hence, more robust and efficient genomic prediction methods are required.

Large reference populations are a necessity for high accuracy of genomic prediction and pooling of related breeds is an option to achieve this. A combination of multi-breed reference and WGS data was shown to be quite beneficial in this thesis. This needs to be empirically evaluated as WGS data become available across related cattle breeds.

Including genotyped cows to the reference population may also increase the reference population. A challenge is that cow evaluations are much less accurate than progeny-tested bulls. Therefore, further investigations possibly comparing the strategies mentioned under general discussion (section 4.1.2.) to identify the optimal way to jointly utilize bulls and cows information in a single reference population, would be beneficial, both for increasing prediction accuracy as well as producing unbiased genomic prediction.

Genomic prediction utilizes genetic relationship among individuals and LD between SNPs and QTL as information sources. **Paper II** showed that methods explicitly accounting for both information sources slightly improved accuracy of prediction. This should be further investigated especially in situations where reference populations and validation individuals are distantly related.

Simultaneous use of all available information is gradually becoming the trend in genomic prediction. **Paper III** on average did not show an extra gain in accuracy but less biased predictions were obtained using all available information. In situations, with very different

selection histories of the selection candidates, biases of GEBV estimations may themselves reduce the accuracy of GS, since the GEBV of some candidates may be differently biased than that of others. Further empirical studies are needed on how best to utilize the rapidly growing number of genotyped animals and millions of ungenotyped animals in variable selection based genomic prediction.

7. REFERENCES

Bett, R. C., K. Johansson, E. Zonabend, B. Malmfors, J. Ojango, M. Okeyo, and J. Philipsson. 2010. Trajectories of evolution and extinction in the Swedish cattle breeds. In 9th world congress on genetics applied to livestock production. Leipzig, Germany.

Buch, L. H., M. Kargo, P. Berg, J. Lassen, and A. C. Sorensen. 2012. The value of cows in reference populations for genomic selection of new functional traits. Animal (6):880-886.

Calus, M. P. 2010. Genomic breeding value prediction: methods and procedures. Animal 4:157-164.

Calus, M. P., Y. de Haas, M. Pszczola, and R. F. Veerkamp. 2013a. Predicted accuracy of and response to genomic selection for new traits in dairy cattle. Animal 7(2):183-191.

Calus, M. P., Y. de Haas, and R. F. Veerkamp. 2013b. Combining cow and bull reference populations to increase accuracy of genomic prediction and genome-wide association studies. J Dairy Sci 96:6703-6715.

Calus, M. P., T. H. E. Meuwissen, A. P. de Roos, and R. F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553-561.

Christensen, O. F. and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet Sel Evol 42:2.

Clark, S. A., J. M. Hickey, H. D. Daetwyler, and J. H. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet Sel Evol 44:4.

Clark, S. A., J. M. Hickey, and J. H. van der Werf. 2011. Different models of genetic variation and their effect on genomic evaluation. Genet Sel Evol 43:18.

Coster, A., J. W. Bastiaansen, M. P. Calus, J. A. van Arendonk, and H. Bovenhuis. 2010. Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. Genet Sel Evol 42:9.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021-1031.

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3:e3395.

Dassonneville, R., A. Baur, S. Fritz, D. Boichard, and V. Ducrocq. 2012. Inclusion of cow records in genomic evaluations and impact on bias due to preferential treatment. Genet Sel Evol 44:40.

De Los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. Calus. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327-345.

De Roos, A. P. W. 2011. Genomic selection in dairy cattle. in Animal Breeding and Genomic Centre. Vol. Ph.D. Wageningen University, the Netherlands, Wageningen.

De Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of genomic predictions across multiple populations. Genetics 183:1545-1553.

Dekkers, J. C. M. 2004. Commercial application of marker- and gene-assisted selection in livestock Strategies and lessons. J Anim Sci 82:E313 - E328.

Ding, X., Z. Zhang, X. Li, S. Wang, X. Wu, D. Sun, Y. Yu, J. Liu, Y. Wang, Y. Zhang, S. Zhang, Y. Zhang, and Q. Zhang. 2013. Accuracy of genomic prediction for milk production traits in the Chinese Holstein population using a reference population consisting of cows. J Dairy Sci 96:5315-5323.

Druet, T., I. M. Macleod, and B. J. Hayes. 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 112:39-47.

Ducrocq, V. and G. Wiggans. 2014. Genetic improvement of dairy cattle. Pages 370–395 in Genetics of Cattle (2nd). D. J. Garrick and A. Ruvinsky, ed. CAB International, Wallingford, UK.

Eggen, A. 2012. The development and application of genomic selection as a new breeding paradigm. Anim Front 2:10-15.

Egger-Danner, C., J. B. Cole, J. E. Pryce, N. Gengler, B. Heringstad, A. Bradley, and K. F. Stock. 2014. Invited review: overview of new traits and phenotyping strategies in dairy cattle with a focus on functional traits. Publications from USDA-ARS / UNL Faculty:1489.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci 95:4114-4129.

Goddard, M. E. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245-257.

Goddard, M. E. and B. J. Hayes. 2007. Genomic selection. J Anim Breed Genet 1245:323 - 330.

Goddard, M. E. and B. J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet 10:381-391.

Goddard, M. E., B. J. Hayes, and T. H. Meuwissen. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet 128:409-421.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389 – 2397.

Habier, D., J. Tetens, F. R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol 42:5.

Harris, B. L. and D. L. Johnson. 2010. The impact of high density SNP chips on genomic evaluation in dairy cattle. Interbull Bull 42:40-43.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol 41:51.

Hayes, B. J., R. Fries, M. S. Lund, D. A. Boichard, P. Stothard, R. F. Veerkamp, C. V. Tassell, C. Anderson, I. Hulsegge, B. Guldbrandtsen, D. Rocha, D. Hinirichs, A. Bagnato, M. Georges, R. Spelman, J. Reecy, A. L. Archibald, E. G. Mike, and B. Gredler. 2012. 1000 Bull Genomes Consortium Project [Abstract]. In Proceedings of Plant and Animal Genome XX Conference. San Diego, CA, United States.

Hayes, B. J., H. A. Lewin, and M. E. Goddard. 2013. The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. Trends Genet 29(4):206-214.

Hayes, B. J., I. M. MacLeod, H. D. Daetwyler, P. J. Bowman, A. J. Chamberlain, C. J. Vander Jagt, A. Capitan, H. Pausch, P. Stothard, X. Liao, C. Schrooten, E. Mullaart, R. Fries, B. Guldbrandtsen, M. S. Lund, D. A. Boichard, R. F. Veerkamp, C. P. VanTassell, B. Gredler, T. Druet, A. Bagnato, J. Vilkki, D. J. deKoning, E. Santus, and M. E. Goddard. 2014. Genomic prediction from whole genome sequence in livestock - the 1000 Bull Genomes Project. In Proceedings of 10th World Congress of Genetics Applied to Livestock Production. Vancouver, Canada.

Hoze, C., S. Fritz, F. Phocas, D. Boichard, V. Ducrocq, and P. Croiseau. 2014. Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. J Dairy Sci 97(6):3918-3929.

Jimenez-Montero, J. A., O. Gonzalez-Recio, and R. Alenda. 2012. Genotyping strategies for genomic selection in small dairy cattle populations. Animal 6:1216-1224.

Jimenez-Montero, J. A., O. Gonzalez-Recio, and R. Alenda. 2013. Comparison of methods for the implementation of genome-assisted evaluation of Spanish dairy cattle. J Dairy Sci 96(1):625-634.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J Dairy Sci 92(9):4656-4663.

Legarra, A., O. F. Christensen, I. Aguilar, and I. Misztal. 2014. Single Step, a general approach for genomic selection. Livest Sci 166:54-65.

Liu, Z., F. Reinhardt, A. Bunger, and R. Reents. 2004. Derivation and calculation of approximate reliabilities and daughter yield-deviations of a random regression test-day model for genetic evaluation of dairy cattle. J Dairy Sci 87:1896–1907.

Luan, T., J. A. Woolliams, S. Lien, M. Kent, M. Svendsen, and T. H. E. Meuwissen. 2009. The accuracy of Genomic Selection in Norwegian red cattle assessed by cross-validation. Genetics 183:1119-1126.

MacLeod, I. M., B. J. Hayes, and M. E. Goddard. 2014. The effects of demography and long-term selection on the accuracy of genomic prediction with sequence data. Genetics 198:1671-1684.

Meuwissen, T. H. E. 2009. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet Sel Evol 41:35.

Meuwissen, T. H. E. and M. E. Goddard. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185:623-631.

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

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2013. Accelerating improvement of livestock with genomic selection. Annu Rev Anim Biosci 1:221-237.

Meuwissen, T. H. E., T. Luan, and J. A. Woolliams. 2011. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. J Anim Breed Genet 128(6):429-439.

Meuwissen, T. H. E., T. R. Solberg, R. Shepherd, and J. A. Woolliams. 2009. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genet Sel Evol 41:2.

Neves, H. H., R. Carvaheiro, and S. A. Queiroz. 2012. A comparison of statistical methods for genomic selection in a mice population. BMC Genetics 13:100.

Odegard, J. and T. H. Meuwissen. 2014. Identity-by-descent genomic selection using selective and sparse genotyping. Genet Sel Evol 46:3.

Odegard, J., T. Moen, N. Santi, S. A. Korsvoll, S. Kjoglum, and T. H. Meuwissen. 2014. Genomic prediction in an admixed population of Atlantic salmon (Salmo salar). Front Genet 5:402.

Olsen, H. G., B. J. Hayes, M. P. Kent, T. Nome, M. Svendsen, A. G. Larsgard, and S. Lien. 2011. Genome-wide association mapping in Norwegian Red cattle identifies quantitative trait loci for fertility and milk production on BTA12. Anim Genet 42:466-474.

Pryce, J. E. and H. D. Daetwyler. 2012. Designing dairy cattle breeding schemes under genomic selection: a review of international research. Anim Prod Sci 52:107.

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

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet 123:218-223.

Solberg, T. R., B. Heringstad, M. Svendsen, H. Grove, and T. H. E. Meuwissen. 2011. Genomic predictions for production and functional traits in Norwegian Red from BLUP analyses of imputed 54K and 777K SNP data. Interbull Bull 44:240-243.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen. 2008. Genomic selection using different marker types and densities. J Anim Sci 86:2447-2454.

Su, G., R. F. Brondum, P. Ma, B. Guldbrandtsen, G. P. Aamand, and M. S. Lund. 2012. Comparison of genomic predictions using medium-density (approximately 54,000) and high-density (approximately 777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. J Dairy Sci 95:4657-4665.

Su, G., P. Ma, U. S. Nielsen, G. P. Aamand, G. Wiggans, B. Guldbrandtsen, and M. S. Lund. 2015. Sharing reference data and including cows in the reference population improve genomic predictions in Danish Jersey. Animal 2:1-9.

Thomasen, J. R., A. C. Sorensen, M. S. Lund, and B. Guldbrandtsen. 2014. Adding cows to the reference population makes a small dairy population competitive. J Dairy Sci 97(9):5822-5832.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J Dairy Sci 91:4414-4423.

VanRaden, P. M. and G. R. Wiggans. 1991. Derivation, Calculation, and Use of National Animal Model Information. J Dairy Sci 74:2737-2746.

Wang, T., Y. P. Chen, M. E. Goddard, T. H. Meuwissen, K. E. Kemper, and B. J. Hayes. 2015. A computationally efficient algorithm for genomic prediction using a Bayesian model. Genet Sel Evol 47:34.

Wientjes, Y. C., R. F. Veerkamp, and M. P. Calus. 2013. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193(2):621-631.

Wientjes, Y. C. J., R. F. Veerkamp, P. Bijma, H. Bovenhuis, C. Schrooten, and M. P. L. Calus. 2015. Empirical and deterministic accuracies of across-population genomic prediction. Genet Sel Evol 47:5.

Wiggans, G. R., T. A. Cooper, P. M. Vanraden, and J. B. Cole. 2011. Technical note: adjustment of traditional cow evaluations to improve accuracy of genomic predictions. J Dairy Sci 94(12):6188-6193.

Yu, X. and T. H. Meuwissen. 2011. Using the Pareto principle in genome-wide breeding value estimation. Genet Sel Evol 43:35.

Yu, X., J. A. Woolliams, T. H. E. Meuwissen. 2014. Prioritizing animals for dense genotyping in order to impute missing genotypes of sparsely genotyped animals. Genet Sel Evol 46:46.

Zhou, L., X. Ding, Q. Zhang, Y. Wang, M. S. Lund, and G. Su. 2013. Consistency of linkage disequilibrium between Chinese and Nordic Holsteins and genomic prediction for Chinese Holsteins using a joint reference population. Genet Sel Evol 45:7.

Zhou, L., B. Heringstad, G. Su, B. Guldbrandtsen, T. H. Meuwissen, M. Svendsen, H. Grove, U. S. Nielsen, and M. S. Lund. 2014a. Genomic predictions based on a joint reference population for the Nordic Red cattle breeds. J Dairy Sci 97:4485-4496.

Zhou, L., M. S. Lund, Y. Wang, and G. Su. 2014b. Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. J Anim Breed Genet 131(4):249-257.

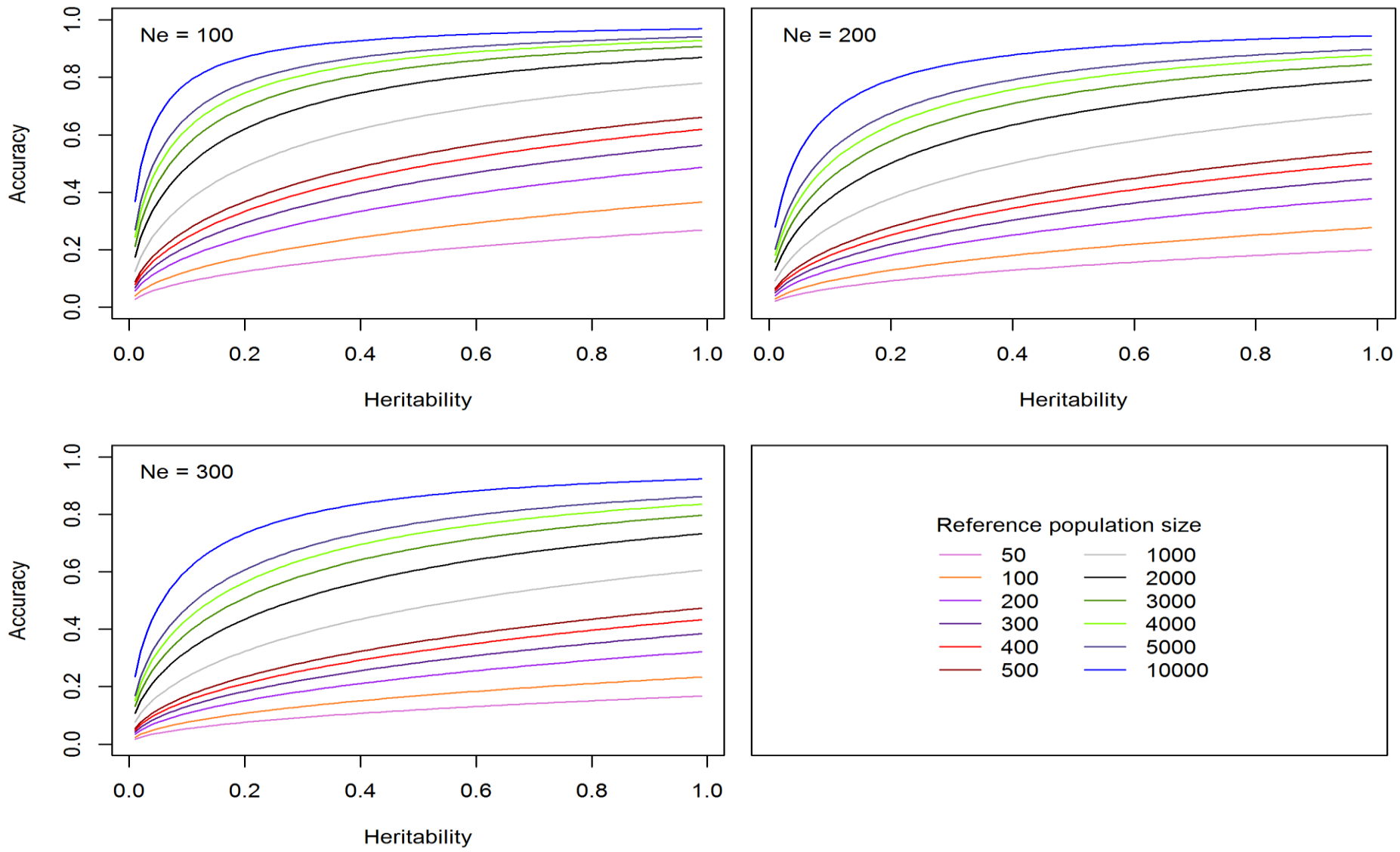

Figure 1 Expected accuracy of prediction under different effective population size (Ne), reference population size, and heritabilities

Paper I

Within- and across-breed genomic prediction using whole-genome sequence and single nucleotide polymorphism panels

O. O. M. Iheshiulor, J. A. Woolliams, X. Yu, R. Wellmann, and T. H. E. Meuwissen

Submitted to Genetic Selection Evolution

Within- and across-breed genomic prediction using whole-genome sequence and single nucleotide polymorphism panels

Oscar OM Iheshiulor[1]*, John A Woolliams[1,2], Xijiang Yu[1], Robin Wellmann[3], Theo HE Meuwissen[1]

[1]Department of Animal and Aquaculture Sciences, Norwegian University of Life Sciences, N-1432 Ås, Norway

[2]The Roslin Institute (Edinburgh), Royal (DICK) School of Veterinary Studies, University of Edinburgh, Midlothian, EH25 9RG, Scotland, UK

[3]Institute of Animal Husbandry and Animal Breeding, University of Hohenheim, Germany

*Corresponding author

Email addresses:

OOMI: oscar.iheshiulor@nmbu.no

JAW: john.woolliams@roslin.ed.ac.uk

XY: xijiang.yu@nmbu.no

RW: r.wellmann@uni-hohenheim.de

THEM: theo.meuwissen@nmbu.no

Abstract

Background

Currently, genomic prediction in cattle is largely based on panels of about 54k single nucleotide polymorphisms (SNPs). However with the decreasing costs of and current advances in next-generation sequencing technologies, whole-genome sequence (WGS) data on large numbers of individuals is within reach. Availability of such data provides new opportunities for genomic selection, which need to be explored.

Methods

This simulation study investigated how much predictive ability is gained by using WGS data under scenarios with QTL (quantitative trait loci) densities ranging from 45 to 132 QTL/Morgan and heritabilities ranging from 0.07 to 0.30, compared to different SNP densities, with emphasis on divergent dairy cattle breeds with small populations. The relative performances of best linear unbiased prediction (SNP-BLUP) and of a variable selection method with a mixture of two normal distributions (MixP) were also evaluated. Genomic predictions were based on within-population, across-population, and multi-breed reference populations.

Results

The use of WGS data for within-population predictions resulted in small to large increases in accuracy for low to moderately heritable traits. Depending on heritability of the trait, and on SNP and QTL densities, accuracy increased by up to 31 %. The advantage of WGS data was more pronounced (7 to 92 % increase in accuracy depending on trait heritability, SNP and QTL densities, and time of divergence between populations) with a combined reference population and when using MixP. While MixP outperformed SNP-BLUP at 45 QTL/Morgan, SNP-BLUP was as good as MixP when QTL density increased to 132 QTL/Morgan.

Conclusions

Our results show that, genomic predictions in numerically small cattle populations would benefit from a combination of WGS data, a multi-breed reference population, and a variable selection method.

Background

Genomic selection (GS) is becoming the standard approach to generate genetic progress in livestock. It was pioneered in the dairy cattle sector because of its potential to achieve high accuracy for non-phenotyped animals, thereby reducing generation intervals by reducing the need for progeny-testing. It has been implemented through the use of panels of SNPs (single-nucleotide polymorphisms) that are distributed over the whole genome, and various commercial bovine SNP chips are available, with densities ranging from 3k to 777k (high-density (HD) panels). So far, results of several GS studies in livestock that have been summarized in [1] show that genomic estimated breeding values (GEBV) can be significantly more accurate than EBV based on phenotypes. Dairy cattle such as the Holstein have greatly benefitted from GS through increases in the accuracy of GEBV resulting from a large reference population and low effective population size (Ne). However, the impact of GS on numerically small breeds, sometimes with larger effective population sizes, is much less significant. This is often exacerbated by the fact that greater emphasis is put on functional traits in these breeds, which typically have lower heritabilities than production traits. Solberg et al. [2] and Su et al. [3] have compared the use of HD to 54k SNP chips and reported small or no increases in accuracy of GEBV or increases only for some traits. Pooling animals from related breeds has been proposed as an option to overcome the small size of reference populations but has not been very successful due to non-persistent associations between SNPs and QTL (quantitative trait loci) across breeds (populations), or inconsistent linkage disequilibrium (LD) between SNPs and QTL across populations [4-6].

As a consequence of these results, progress in the genomic evaluation of dairy cattle has been largely based on the use of 54k SNP panels and primarily restricted to evaluations within breeds. However, new opportunities for GS will arise from the rapid advances in next-generation sequencing technologies e.g. [7], with whole-genome sequence (WGS) data becoming available for large numbers of individuals. Such data need to be explored for their potential in across-breed evaluations. WGS data differ fundamentally from the current data obtained with dense SNP chips because all variants, such as SNPs, indels, copy number variants (CNV), etc., are included. Since all variants, both rare and common, are captured for a population, WGS data could provide more precise signals for causative mutations, both within and across families; hence predictions would no longer have to completely rely on linkage disequilibrium (LD) between SNPs and QTL. Consequently, WGS data could lead to more accurate genomic predictions. In the case of across-breed predictions, the use of WGS data could reduce or remove the need to rely on associations between SNPs and QTL which may not persist across the breeds being evaluated [8].

Although Meuwissen and Goddard [9] reported an advantage of WGS data over dense SNP data using simulated data, their results were restricted to within-population predictions and to a small number of QTL/Morgan. Therefore, the advantages of WGS data for across-breed predictions and divergent small populations remain largely unknown. Hence, the first objective of this study was to assess how much predictive ability is gained by using WGS data under varied QTL densities and trait heritabilities compared to different SNP densities, with emphasis on divergent dairy breeds with small populations and large effective population sizes (> 100). Secondly, we assessed the relative performance of the use of a non-variable selection method (SNP-based best linear unbiased prediction (SNP-BLUP)) and a variable selection method (MixP; [10]) for genomic prediction.

Methods

Scaling of the simulated populations

GS generally requires large training populations, and thus the simulation and analysis of WGS data (with millions of SNPs), which comprise many large and replicated populations over many generations, are computationally prohibitive. In this study, we followed the scaling argument used and tested by [11, 9], which is based on the equation developed by [12, 13] for expected accuracy of genomic prediction. According to [12, 13], the accuracy of genomic prediction depends on the parameter $\lambda = Th^2/ML$, where T is the number of individuals with genotypes and phenotype in the training data, h^2 is the heritability of the trait, M is the effective number of loci per Morgan (~2Ne), and L is the genome size in Morgan. If we scale the genome size down from 30 to 1 Morgan, and simultaneously reduce the training population size by a factor of 30, λ remains constant and, thus, the accuracy of prediction will not be affected. E.g. a large-scale simulation with 6000 training animals with 30-Morgan genomes, yields approximately the same accuracy of genomic prediction as a training population of 200 animals with a 1-Morgan genome, which requires less computer resources.

Simulation of whole-genome sequence data

The parameters used in the simulation of the genome and population structure are summarized in Table 1. The simulation was based on a forward-in-time approach. A Fisher-Wright idealized population was simulated [14], with a mutation rate of 10^{-8} per bp per meiosis, assuming 1 Mb per cM and a historical effective population (Ne) size of 200. The simulation was conducted for a minimum of 1950 or 1990 generations, to create a mutation-drift-recombination equilibrium. Previous studies have shown that this is sufficient to establish equilibrium [15, 16]. At each generation, a breeding pool of 100 males and 100 females was generated, and mating

proceeded with random sampling of a sire and a dam for each offspring. Therefore, mutation and drift were the only two evolutionary forces considered.

To simulate two diverged populations, the population was split into two to represent separate breeding populations, Population A and Population B. This divergence took place after either 1950 or 1990 generations and each of the populations was simulated for a further T generations (T = 50 or 10), such that each population was propagated for 2000 generations. Hence, SNP mutations that occurred after 1950 or 1990 generations were specific to each population. After the divergence, only within-subpopulation matings occurred (i.e. there was no exchange of genetic material between sub-populations), with a Ne of 200 for each sub-population. At generation 2000, census size was increased to 500 individuals in each population. Among these 500 individuals, 200 were randomly sampled and phenotypes were generated to form a reference population for estimation of marker effects, while the remaining 300 individuals were used to form a validation population, for which breeding values were predicted from their genotypes.

QTL densities and datasets

The mutation-drift process resulted on average 4648 variants that were distributed across a 1-Morgan chromosome with a minor allele frequency (MAF) higher than 0.02 and a standard deviation of the number of variants of 125. All variants were treated as SNPs. Among the SNPs generated, 45 or 132 loci were randomly sampled and designated as causative QTL, which resulted in 4603 or 4516 remaining SNPs. For each population, there were on average 45 (132) randomly sampled QTL SNPs. All SNPs, including the QTL represented the WGS data. Different SNP densities were then created by randomly sampling without replacement from the non-QTL loci. These panels contained 3000, 2000, 1000, or 200 SNPs and were named data

data3000, data2000, data1000, or data200, respectively. These densities are equivalent to 90K, 60K, 30K, and 6K SNP panels for a 30-Morgan bovine genome.

Genetic and phenotypic values

Two traits, with heritabilities of 0.30 and 0.07, were simulated for each scenario. Since for quantitative traits, a large proportion, typically more than half, of the total genetic variance is additive and responsible for most of the genetic variation within a population [14, 17], we assumed only an additive genetic model. An allelic effect (a_j') was assigned to the reference allele (allele "1") of each QTL by sampling effects from a normal distribution. After sampling, their effects were standardized to achieve a total genetic variance of 1, by $a_j = a_j' / \sqrt{\sum_k 2p_k(1-p_k)(a_k')^2}$, where subscripts *k* and *j* denote the *k-th* and *j-th* QTL, the summation is over all QTL, and p_k is the within-population frequency of allele "1" of the *k*-th QTL. Then, the total genetic value for individual *i* was calculated as:

$$g_i = \sum_{j=1}^{N_{QTL}} x_{ij} a_j ,$$

where x_{ij} is the number of alleles "1" that individual *i* carries at locus *j*. Phenotypes were generated by adding environmental effects drawn from a normal distribution with a mean of zero and variance such that heritability was 0.30 or 0.07 in generation 2000 within each population.

Design of evaluations with reference and validation populations

In Scenario 1, predictions were calculated within population A using its own reference population of 200 individuals, with the remaining 300 individuals from the same population used for validation. In Scenario 2, across-population predictions were calculated for population

B using the reference population of population A to predict the breeding value of individuals in the validation population of B. In Scenario 3, predictions were based on a multi-breed reference population after combining the reference populations of A and B to reach a total number of 400 individuals. Validation was performed within-breed, using 300 individuals from population A. This Scenario 3 can occur in practice when breeders of populations A and B combine their reference populations, with the aim of increasing accuracy for their own population. In Scenario 4, we checked the impact of increasing the reference population of A in Scenarios 1 and 2 by increasing its size to 400 and validating it by predicting breeding values in either A or B and comparing it to multi-breed estimation of SNP effects.

Estimation methods and data analysis

Two methods were used to estimate SNP effects in the reference population: SNP-BLUP and MixP [10]. SNP-BLUP estimates the effects of SNPs by best linear unbiased prediction [18] and was implemented using a ridge regression model of the SNP effects that is equivalent to model 2 of VanRaden [19]. MixP is similar to BayesC [20] except that SNPs with small effects are assumed to explain part of the genetic variance instead of having no effect. Therefore, MixP assumes that SNP effects come from a mixture of two normal distributions [10], i.e. one with a large variance (σ_1^2) and one with a small variance (σ_2^2). The distribution of the total genetic variance (Vg) over the 'large' SNPs and the 'small' SNPs followed the Pareto principle (hence the P in MixP), such that x% of the SNPs with the largest effects are responsible for $(100 - x)$% of the genetic variance. Given the prior for the mixing frequency ($\pi = x/100$) and using the Pareto principle, the variances of the large and small SNP effects are respectively:

$$\left.\begin{aligned}\sigma_1^2 &= \frac{(1-\pi)V_g}{\pi N_m}\\ \sigma_2^2 &= \frac{\pi V_g}{(1-\pi)N_m}\end{aligned}\right\},$$

where N_m is the total number of genotyped SNPs, such that $N_m(\pi\sigma_1^2 + (1-\pi)\sigma_2^2) = V_g$. The π value used for MixP was set to N_{QTL}/N_m (i.e. number of QTL simulated vs. number of SNPs used). A preliminary study on the optimal values of π revealed that values around N_{QTL}/N_m were close to optimal and that deviations from this value hardly affected the accuracy of genomic prediction.

The linear model used to estimate SNP effects for both SNP-BLUP and MixP approaches was as follows:

$$\mathbf{y} = \mu + \sum_{j=1}^{N_m} \boldsymbol{X_j} b_j + \mathbf{e},$$

where **y** is a Nx1 vector of phenotypes; μ is the overall mean; N_m is the total number of genotyped SNPs; $\mathbf{X_j}$ is a Nx1 vector of the N standardized SNP genotypes, i.e. $\mathbf{X_j} = \frac{-2p_j}{\sqrt{2p_j(1-p_j)}}, \frac{1-2p_j}{\sqrt{2p_j(1-p_j)}}, \; or \; \frac{2(1-p_j)}{\sqrt{2p_j(1-p_j)}}$ depending on the genotype of individuals *i* "0 0", "1 0", or "1 1", respectively and p_j is the allele frequency of SNP *j*; b_j is the effect of the *j-th* SNP genotype; and **e** is a Nx1 vector of environmental effects assumed to be distributed as N(0, $\mathbf{I}\sigma_e^2$). For the SNP-BLUP approach, b_j is assumed to follow the distribution N(0, σ_b^2), where σ_b^2 is the SNP variance (σ_g^2/N_m), and for the MixP, each $b_j$is N(0, σ_1^2) with probability π, or b_j is N(0, σ_2^2) with a probability (1-π). The simulated genetic and environmental variances were used as parameters in SNP-BLUP and MixP. MixP used the Iterative Conditional Expectation (ICE) algorithm of Meuwissen et al. [21]. Full details of MixP are provided in [10].

After estimating SNP effects, the GEBV ($\hat{g}_i$)of the validation individuals (i.e. the individuals having only genotypic records) was predicted as:

$$\hat{g}_i = \sum_{j=1}^{N_m} X_{ij}\hat{b}_j ,$$

where X_{ij} is the standardized SNP genotype of individual *i* for SNP *j*; and $\widehat{b_j}$ is the estimate of the SNP effect. The correlation between true (g_i) and estimated genetic value ($\widehat{g_i}$) was used as a measure of the accuracy of prediction.

Replication strategy

Simulation procedures were replicated 30 times. Propagation of populations A and B over 2000 generations was repeated 30 times with T = 10 and 45 QTL, and 30 times with T = 10 and 132 QTL. A further 60 full replications of the populations were carried out with T = 50, equally divided between 45 and 132 QTL. Genomic evaluation procedures were then carried out on each of these 120 replicates. The same 120 replicates were used for h^2 = 0.07 and 0.30 by resampling the phenotypes. Thus, the results are means of 30 replicates for each T (10 or 50) by QTL number (45 or 132) by heritability (0.07 or 0.30) combination. Standard errors were computed as the standard deviation of the accuracies across the 30 replicates, divided by $\sqrt{30}$.

Results

Simulated populations

Allele frequencies and QTL variances ($2pqa^2$) differed between populations (A and B). As an example, Fig. 1 shows the distribution of allele frequencies for one of the simulated replicates (scenario of 45 QTL/Morgan and T = 10 or 50) for both populations, while Fig. 2 and 3 show the QTL variances. Some SNPs and QTL were fixed in both populations (result not shown). Figure 4 shows the average LD in the WGS dataset, measured as the squared correlation (r^2) between adjacent SNPs and the persistency of LD phase of adjacent SNPs between the two populations at different times of divergence, measured as the correlation between the two

populations of the phased LD, r, of marker pairs. LD ranged from ~0.36 to 0.40 at genomic distances of 0 to 50 kb, respectively, and this trend was similar in both populations. At a genomic distance of 100 kb, LD dropped to about 0.25. As expected, LD decreased further with increasing genomic distance between SNPs (Fig. 4). Persistence of r for adjacent SNPs between populations was equal to ~0.85 at 50 kb for the scenario with T = 10 and ~0.65 for T = 50. This implies that LD of very close SNPs was more persistent between the two populations at T = 10 than at T = 50. A gradual decline in r was observed with increasing genomic distance. For data3000 to data1000, r^2 and r were lower (especially for data1000) as a result of the decreasing SNP density and increasing inter-SNP distance (result not shown). This also affected r^2 and r results for data200. The average inter-SNP distances were equal to 21, 33, 50, 100 and 496 kb for the WGS data and data3000 to data200, respectively.

Scenario 1: Genomic predictions within population A

Accuracies of predictions for population A based on different datasets, trait heritabilities (h^2), and QTL densities, using SNP-BLUP and MixP are in Tables 2 and 3.

Effect of dataset

Use of WGS data resulted in a 1 to 31 % increase in accuracy across the different SNP densities. In all cases, the lowest accuracies were found with the lowest SNP density, i.e. data200, while the highest accuracies were obtained with WGS data. The observed differences in accuracy between the WGS data and data1000 up to data3000 were quite small compared to that between the WGS data and data200.

Effect of heritability and QTL density

Accuracy was higher for the trait with a high heritability (0.30) and when QTL density was low. Increasing the QTL density from 45 to 132 per Morgan led to a decrease in accuracy regardless

of trait heritability, however, this decrease was greater for the trait with a heritability of 0.07 (Tables 2 and 3).

Effect of evaluation method

The relative superiority of the two methods depended on QTL density and on trait heritability. At 45 QTL/Morgan, MixP was slightly superior to the SNP-BLUP method, however, the differences between both methods became smaller when the number of QTL increased to 132. The two methods achieved very similar accuracy for the trait with a low heritability (0.07) and 132 QTL/Morgan.

Tables 2 and 3 also show the relative performance of the two methods when using WGS data. Higher accuracies were observed when using MixP, especially at low QTL density. At a density of 45 QTL/Morgan, accuracy increased by 6.0 and 4.4 % with MixP compared to SNP-BLUP for heritabilities of 0.30 and 0.07, respectively, while at a density of 132 QTL/Morgan, accuracy increased only slightly by 0.9 and 0.3 %, respectively. Based on these results, it follows that the predictive ability of MixP decreases as the QTL densities increase, because MixP is not able to fully identify SNPs with larger effects, while the predictive ability of SNP-BLUP is less affected by increasing QTL densities because it assumes that all SNPs have equal variance.

Scenario 2: Across-population genomic predictions

Table 4 summarizes the accuracies obtained for the different datasets using SNP-BLUP and MixP when the populations have diverged for T generations. The prediction equation from population A was used to estimate GEBV for population B. In summary, Table 4 shows that, for population A, prediction accuracy was significant only when WGS data was used and accuracies were close to zero with the SNP densities.

At T = 10, prediction accuracies were significantly lower compared to those obtained when reference and validation individuals originated from the same population. Depending on the evaluation method and QTL density, accuracies ranged from 0.39 to 0.48 when heritability was 0.30 and when WGS data were used, while accuracies were close to zero with the SNP densities, (Table 4). For the trait with heritability equal to 0.07, accuracies ranged from 0.21 to 0.29 when WGS data were used, while accuracies were again close to zero with the SNP densities.

At T = 50, prediction accuracies were also significantly lower compared to those obtained when reference and validation originated from the same population and at T = 10. Accuracies ranged from 0.24 to 0.36 with heritability equal to 0.30 and from 0.17 to 0.20 for heritability equal to 0.07 when WGS data were used in the different scenarios, while accuracies were close to zero when using the SNP densities. Accuracy was highest (0.36) when the trait had heritability 0.30 and when using MixP at a density of 45 QTL/Morgan.

In general, increasing QTL density from 45 to 132 per Morgan, as well as time of divergence from 10 to 50 generations, led to a decrease in accuracy, which was even greater for a trait with a low heritability. MixP was relatively superior to SNP-BLUP when using WGS data in all cases with a density of 45 QTL/Morgan but SNP-BLUP was as good as MixP for greater QTL densities.

Scenario 3: Multi-breed genomic predictions

Accuracies of multi-breed predictions when populations diverged for T generations are in Table 5. The reference population comprised 200 individuals from each population A and B. Generally, adding 200 individuals from population B to the reference of population A led to a substantial increase in accuracy when using WGS data. Across SNP densities, accuracies were more or less similar to those obtained for within-population predictions (see Tables 2 and 3). Prediction accuracies with data200 were always lower than those of the other datasets; a very

small increase in accuracy was observed when going from data1000 to data2000 and data3000, while it was much larger with WGS data. The dataset, heritability, QTL density, time of divergence, and evaluation method, all had an effect on the accuracy of prediction.

At T = 10, a multi-breed reference population using WGS data led to a greater increase in accuracy than that obtained for within- or across-population predictions. Depending on the evaluation method and QTL density, accuracies ranged from 0.65 to 0.71 with heritability equal to 0.30 when WGS data were used, while they ranged from 0.41 to 0.60 with the SNP densities. For the trait with heritability equal to 0.07, accuracies ranged from 0.41 to 0.53 when WGS data were used, while they ranged from 0.22 to 0.41with the SNP densities.

At T = 50, a multi-breed reference population using WGS data also led to an increase in accuracy compared to within- or across-population predictions but the increase in accuracy was lower than that obtained at T = 10. With WGS data, accuracies ranged from 0.61 to 0.68 with heritability equal to 0.30 and from 0.37 to 0.48 with heritability equal to 0.07 for the different scenarios. With the SNP densities, accuracies ranged from 0.41 to 0.59 and from 0.24 to 0.41 for heritabilities of 0.30 and 0.07, respectively.

In summary, increasing QTL density from 45 to 132 per Morgan and time of divergence from 10 to 50 generations led to a decrease in accuracy, although its magnitude depended on trait heritability and the evaluation method. With a density of 45 QTL/Morgan, MixP was superior to SNP-BLUP when using WGS data but SNP-BLUP was as good as MixP when the QTL density increased to 132 QTL/Morgan. Accuracy was highest, i.e. 0.71 for the trait with heritability equal 0.30 at 45 QTL/Morgan and using MixP when populations had diverged for 10 generations.

Scenario 4: Impact of size of the single-breed versus the multi-breed reference population

Table 6 shows the impact of using either a single- (200 or 400 individuals) or multi- breed reference population on the accuracy of genomic prediction. The impact of increasing the reference population size in the case of across-breed prediction was also analyzed but since the results followed the same trend as those in Table 4, they were not included in Table 6. Multi-breed estimation of SNP effects with WGS data at either 10 or 50 generations of divergence resulted in higher accuracy than when using 200 single-breed individuals. However, single-breed predictions resulted in higher accuracies than multi-breed predictions when equal numbers of reference individuals (400) were used.

Discussion

This study examined the accuracy of genomic prediction for within-population, across-population, and multi-breed reference populations. WGS data, various SNP densities, QTL densities, trait heritabilities, and GEBV estimation methods were used. Results show that the use of WGS data would lead to increased accuracy of genomic prediction for low to moderately heritable traits. Increases in accuracy from using WGS data were much greater with a multi-breed reference population but remained substantial in across-population scenarios. For within-population predictions, the use of WGS data compared to data3000 increased accuracy by 3.1 and 3.3 % (SNP-BLUP) and by 5.7 and 5.9 % (MixP) for traits with heritabilities equal to 0.30 and 0.07, respectively, at 45 QTL/Morgan. With 132 QTL/Morgan, these figures decreased to: 1.2 and 4.5 % (SNP-BLUP) and to 1.4 and 4.5 % (MixP). Expanding the reference population with animals from another population (breed) and using WGS data resulted in a remarkable increase in accuracy compared to that achieved by within-population prediction. With SNP densities, only a minor or no increase in accuracy was observed. In general, MixP had an advantage over SNP-BLUP at low QTL density, but this advantage decreased as QTL density

increased, probably due to the many QTL with smaller effects. With many QTL with smaller effects, the MixP method was less able to pinpoint the SNP(s) that best explained the QTL.

Within-population genomic predictions

WGS data are not yet available on large numbers of individuals but it should lead to greater accuracies of genomic prediction. Presently, the 1000-bull genomes project [22] makes it possible to impute WGS data on (densely) genotyped animals. It is expected that with WGS data, possibly all variants (including causative mutations) in a population will be captured, which means that genomic prediction does not need to rely completely on LD between SNPs and causative mutations. In all the scenarios investigated, the use of WGS data showed a clear advantage over the use of different SNP densities. Depending on trait heritability, marker density (data1000 to data3000), and QTL density, the observed increase in accuracy when using WGS data for within-population prediction was as high as 13 %. For the lowest SNP density (i.e. data200), the increase reached 24 to 31 %. These results follow the same upward trend in accuracy as observed by Meuwissen and Goddard [9].

In their simulation study, VanRaden et al. [23] excluded the causative mutations but increased the numbers of SNPs from 54k to 500k and reported a gain in accuracy of only 1.6 %. A simulation study by Druet et al. [24] under the neutral model (i.e. when QTL allele frequencies followed the same distribution as other variants in the sequence) found accuracies to increase by only 1.4 % when comparing WGS data with a SNP panel. When the same authors, assumed that the causative mutations all had low MAF, using the WGS data (real/imputed) improved the accuracy of genomic prediction by up to 30 %. Thus, depending on the scenario assumed, there are agreements and differences of our results with the study of [24]. These differences, in results especially for the neutral model, may be because Druet et al. [24] used a much denser SNP panel (14 2385 SNPs on a 50 Mb genome) and their accuracies were already very high, i.e.

around 0.9, which left little room for improvement. On a general note, Druet et al. [24] used a smaller genome (50 Mb with five 10-Mb chromosomes), an effective population size of ~100 and a large reference population of 1021 individuals. The small(er) effective population size leads to extensive LD and a substantially smaller number of effective chromosome segment effects to be estimated, hence better predictions and higher accuracies [25].

In dairy cattle, Hayes et al. [8] used imputed WGS data from the 1000-bull genomes project and observed a 2 % increase in prediction accuracy compared to HD data. Our results differ from those of Hayes et al. [8]. This difference is probably explained by the fact that Hayes et al. [8] used an imputed WGS data, in which case the accuracy of prediction depends in part on how accurately the common and rare variants are imputed. Variants with a MAF higher than 5 % were imputed with an accuracy of about 0.7 to 0.9, while with a MAF lower than 5 %, imputation accuracy rapidly declined [8]. With accurate imputation of common variants, an extra 2 % increase in accuracy of genomic prediction was observed [8], which suggests that if all variants (common and rare) are accurately imputed, a higher accuracy of genomic prediction would be expected with WGS data. As reported by Druet et al. [24], accuracy of genomic prediction could be improved by 2 to 30 % depending on the trait. In inbred *Drosophila melanogaster*, Ober et al. [26] observed no advantage of using WGS data over dense SNP data for genomic prediction. They also reported no difference in prediction performance of SNP-BLUP and BayesB. Their results could be due to: (1) a very large effective population size (~8700), which resulted in a large number of effective chromosome segments (~2000) effects to be estimated; and (2) a small reference population size of about 120. According to [27, 24, 8, 9] the availability of large datasets is important to improve accuracy of genomic prediction even when using sequence data.

In general, the observed increase in accuracy obtained with WGS data can be attributed to the fact that it is not necessary to completely rely on LD between flanking markers and the QTL. However, at a density of 132 QTL/Morgan, MixP and SNP-BLUP performed similarly, which suggests that, in this case, MixP partly relied on LD even when WGS data was used. This means that at high QTL densities (which might be realistic), the ability of MixP to pinpoint the individual QTL with small effects decreases and then relies on LD between SNPs and QTL. According to Meuwissen et al. [18], accuracy of genomic prediction depends on the SNP density and on the LD between SNPs and QTL in order to maximize the proportion of genetic variance explained by the SNPs. However, with WGS data, predictions no longer depend (or at least to a large extent do not) on associations between SNPs and QTL because causative mutations are included in the data and are possibly captured and used in the analysis [27]. Meuwissen and Goddard [9] demonstrated that even with higher SNP densities, an extra gain in accuracy is obtained when including the causative mutations.

Across-population genomic predictions

Using a reference population (A) to predict GEBV of another population (B) resulted in a significant decrease in accuracies compared to within-population predictions, especially at T = 50 generations of divergence between populations (Table 4). Nonetheless, the across-population accuracies that were obtained using WGS data were substantially higher than with SNP densities, for which accuracies were close or equal to zero. Our results are consistent with the literature [6, 28, 29]. Possible reason(s) for the poor results obtained when SNP densities were used could be due to non-persistent associations between SNPs and QTL across populations or inconsistent LD between SNPs and QTL across populations [4-6]. Furthermore, it has been shown that, as the genetic distance between individuals of the reference and validation populations increases, the accuracy of prediction decreases [30, 4, 31, 32]. Differences in allele substitution effects between populations result in differences in genetic

variance and this could impact predictions across populations [33]. Also, a QTL that segregates in one population may not segregate in the other population, thereby resulting in differences in the genetic variance explained by that QTL between populations (see Fig. 2 and 3). The observed differences in QTL variance between populations result directly from differences in allele frequencies, since non-additive effects were not simulated. The use of WGS data suffers much less from changes in LD since it does not need to completely rely on LD between SNPs and QTL. Furthermore, the presence of QTL in the WGS data increases the probability of picking up similar QTL that segregate between populations and that have comparable effects [8]. This may explain why substantially better results were obtained when WGS data was used for across-population prediction, although the accuracies were lower compared to within-population prediction. In summary, the WGS and SNP data differ in the sense that all variants (causative mutations included) are included in the WGS data, which makes it less dependent on LD, while SNP data fully depends on LD.

Multi-breed genomic predictions

One of the key factors that affect accuracy of genomic prediction is the number of reference animals [25, 6]. Accuracy increases as the number of reference animals increases because the amount of phenotypic data becomes sufficient to detect causative mutations and to distinguish their effects from random noise [27]. Numerically small dairy populations are faced with the problem of a small reference population. Therefore, using a multi-breed reference population could be an option. Our study showed that adding individuals from a second population (population B) to the reference population yielded substantially higher accuracies of GEBV for population A when using WGS data (Table 5). The observed increase in accuracy was greater when the populations had diverged for 10 generations compared to 50 generations, which indicates that relatedness between populations plays a role and should be taken into account when considering a multi-breed reference population for genomic prediction. With the SNP

densities that we used, the use of a multi-breed reference population resulted in similar accuracies as those obtained with a single-breed reference population (Table 6). When using equal numbers of reference individuals (400) for multi-breed and single-breed genomic prediction, the single-breed reference population resulted in higher accuracies than the multi-breed reference population (Table 6). Hence, the higher accuracies obtained when using a multi-breed reference population with WGS data can be attributed to: (1) a greater number of reference animals; and (2) the inclusion of causative mutations, which enhances the possibility of picking up similar QTL that segregate between populations and that have comparable effects [8]. The authors of [8] also pointed out that multi-breed prediction using WGS data leads to more accurate predictions because causative mutations that segregate among populations are captured and used in predictions. According to De Roos [27], the maximum benefit of WGS data can be obtained if the number of reference individuals is increased accordingly. Meuwissen [11] also reported that large reference datasets are needed in order to take full advantage of high-density markers. So far, at least to the best of our knowledge, no study (real or simulation) has evaluated the use of WGS data for multi-breed genomic prediction. However in an imputation study, Bouwman and Veerkamp [34] reported greater imputation accuracy (0.83) when using a multi-breed reference population to impute genotypes from a high-density SNP panel (777k) to WGS, compared to an imputation accuracy of 0.70 when using a single-breed reference population. This shows the benefit of using a multi-breed reference population when reference populations are small. Our study shows that with WGS data and a sufficient number of reference animals, higher genomic prediction accuracies are reached for low to moderately heritable traits.

Impact of QTL density

In practice, the number of causative mutations that underlie a trait is not known [35]. Thus, we studied two different QTL densities (45 and 132 QTL/Morgan) to investigate the impact of

QTL density. We observed that, as the QTL density increased from 45 to 132 (i.e. 1350 or 3960 QTL for a 30-Morgan genome), accuracy decreased markedly. This decrease in accuracy is consistent with results from other simulation studies [35, 25, 9, 10]. As explained by Meuwissen and Goddard, [9], QTL effects and opportunities for their detection become smaller with increasing QTL density, resulting in less accurate GEBV. It should be noted that in spite of the decrease in accuracy when going from 45 to 132 QTL/Morgan, WGS data still resulted in higher accuracies than the SNP densities.

Evaluation method

The availability of WGS data for a large number of individuals would provide a large amount of information for genomic prediction. These data would contain millions of variants and it would be necessary to estimate their effects accurately. Therefore, we examined the relative performance of SNP-BLUP and a variable selection method with a mixture of two normal distributions, MixP. The results showed that MixP outperformed SNP-BLUP at a density of 45 QTL/Morgan and also resulted in higher accuracies with WGS data. However, as QTL density increased, accuracy decreased for both methods but MixP still yielded higher accuracies. At a density of 132 QTL/Morgan and for a trait with a low heritability (0.07), both methods gave very similar accuracies, which means that for lowly heritable traits that are controlled by a large number of QTL, SNP-BLUP is as good as MixP. Studies by [35, 25, 9, 10] also demonstrated that when the number of QTL became large, the advantage of allowing for large SNP effects decreased. In such a situation, SNP-BLUP, which assumes a normal distribution with equal variance for all SNP effects, performs as well as variable selection methods. With real data, the performance of these two methods have been reported to be quite similar [36, 3] for most traits. However for traits known to be controlled by a small number of major genes (e.g. *diacylglycerol O-acyltransferase 1* (*DGAT1*), which is involved in the control of fat percentage in dairy cattle), Cole et al. [37] and VanRaden et al. [38] reported that variable selection methods outperformed

SNP-BLUP. Thus, the method used for genomic prediction is important and the superiority or relative performance of the methods depends on the genetic architecture that underlies the trait [25].

Assumptions and implications

Using the scaling argument of [11, 9], the results presented here were obtained for scenarios with only one chromosome, and 45 (132) QTL and 4648 variants on average, whereas WGS data in cattle would cover the 30 bovine chromosomes and contain millions of variants and thousands of QTL. The number of SNPs that was simulated (4648 SNPs/Morgan) was much lower than that in real cattle WGS data (~10 to 20 million SNPs/30 Morgan = 0.3 to 0.6 million SNPs/Morgan). This is probably due to the relatively small historical effective population size of 200 used in this study, whereas historical population sizes in cattle were much larger (although current effective sizes are small). This results in blocks with high LD being much shorter in real WGS data than in the data used for our study. On the one hand, this is beneficial for the real WGS data since it will be easier to pinpoint the QTL (fewer SNPs are in real high LD with the QTL), but on the other hand, the number of SNPs to choose from increases dramatically, making it more difficult for variable selection methods to select the right set of SNPs. The small number of variants included in our study made the use of WGS data less challenging for genomic prediction; dealing with millions of variants, as in the case of real WGS data, would be a challenge [8], coupled with an increased number of uninformative variants which might impact accuracy of prediction. Therefore, it would be very beneficial to reduce the amount of uninformative variants as much as possible. Biological information (e.g. coding/regulatory regions and gene sets in that are most likely to harbour mutations affecting traits of interest) that is obtained via (1) the analysis of genome annotation and (2) atlases of bovine gene expression, can be used to prioritize and identify a subset of variants that can then be used to impute densely genotyped animals up to sequence and or genomic prediction [8]. In

essence, maximizing the accuracy of genomic prediction by using WGS data would very much rely on how well informative variants are exploited and not necessarily on the number of variants.

In our study, we assumed that WGS data contain all causative mutations but that may not be the case in practice, because (rare) SNPs may be missed during (stringent) data filtering or if relatively few individuals are sequenced with limited coverage and the remaining individuals are imputed using SNP chip data. However, for across-breed predictions, the results in Table 4 suggest that it is essential to include the causative mutations in the data and thus (over-) stringent filtering of WGS should be avoided. Finally, when dealing with WGS data, methods that are either able to pinpoint the causative mutations or allow a few variants that are in real high LD to capture the effects of causative mutations and not smear there effects across multiple variants that are in moderate LD with the QTL would be very instrumental to achieving sustained accuracy of genomic prediction across generations [8].

In both the simulation and analysis of data in this study, we assumed only additive genetic effects because it is the most important source of genetic variance and they reflect the actual breeding value of an animal [14, 17]. Furthermore, according to [9] the effect of dominance deviations (the simplest non-additive effects) on the accuracy of genomic prediction using WGS data is: (1) if there are a few QTL with large effects (three per Morgan), these are poorly modelled by the additive genomic prediction models; (2) if there are many QTL with small effects (more than 30 per Morgan), the non-additive effects are much smaller and blend with the residual effects, which results in virtually the same accuracy of total genetic value as when gene effects were purely additive (at equal narrow sense heritability).

In our case, epistatic interactions may result in differences in additive effects between the breeds. If the correlation of the additive effects of QTL between breeds is 0.9 (instead of 1 as

assumed here), the accuracy of across-breed genomic prediction would be reduced by 10 %. In the case of multi-breed genomic prediction, the reduction in accuracy of prediction would be less than 10 % (depending on the breed contributions).

We chose to only simulate QTL with a MAF higher than 0.02, which eliminates very rare QTL that may have occurred in only one of the populations, and thus the accuracies of across-breed predictions were favoured. However, very rare QTL would probably not contribute much to the accuracy of prediction, because the genomic prediction models would not detect them.

Conclusions

This study shows that use of WGS data can increase accuracy of genomic prediction for low to moderately heritable traits in small populations. This increase in accuracy with WGS data depended on QTL density, the size of the reference population and the evaluation method used. In the absence of a sufficiently large reference population, aggregation of breeds that share close ancestral ties is an option to increase the reference population size and improve accuracy of genomic prediction. The use of WGS data was especially beneficial for multi-breed predictions and when a variable selection method was used. Thus, to take full advantage of a multi-breed reference population, WGS data, large reference sets and variable selection methods are required.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

OOMI performed the study and drafted the manuscript. JAW contributed to draft and revise the manuscript critically. XY wrote the MixP program. RW contributed to writing the simulation computer program. THEM planned and coordinated the whole study, contributed to writing the

simulation computer program and manuscript. All the authors read and approved the final manuscript.

Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 289592 - Gene2Farm. We also appreciate the anonymous reviewers and editors for their useful comments. Neither the European Commission nor the partners of the project can be held responsible for views expressed in this manuscript.

References

1. Meuwissen THE, Hayes BJ, Goddard ME. Accelerating improvement of livestock with genomic selection. Annu Rev Anim Biosci. 2013;1:221-37.

2. Solberg TR, Heringstad B, Svendsen M, Grove H, Meuwissen THE. Genomic predictions for production and functional traits in Norwegian Red from BLUP analyses of imputed 54K and 777K SNP data. Interbull Bull. 2011;44:240-43.

3. Su G, Brondum RF, Ma P, Guldbrandtsen B, Aamand GP, Lund MS. Comparison of genomic predictions using medium-density (approximately 54,000) and high-density (approximately 777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy cattle populations. J Dairy Sci. 2012;95:4657-65.

4. de Roos APW, Hayes BJ, Goddard ME. Reliability of genomic predictions across multiple populations. Genetics. 2009;183:1545-53.

5. Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet. 2009;10:381-91.

6. Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol. 2009;41:51.

7. Illumina. An introduction to Illumina next-generation sequencing technology for agriculture. 2013. http://res.illumina.com/documents/products/appspotlights/app_spotlight_ngs_ag.pdf. Accessed 01/02/2014.

8. Hayes BJ, MacLeod IM, Daetwyler HD, Bowman PJ, Chamberlain AJ, Vander Jagt CJ, et al. Genomic prediction from whole genome sequence in livestock - the 1000 Bull Genomes Project. In: Proceedings of 10th World Congress of Genetics Applied to Livestock Production: 17-22 August; Vancouver. 2014. https://asas.org/docs/default-source/wcgalp-proceedings-oral/183_paper_10441_manuscript_1644_0.pdf?sfvrsn=2.

9. Meuwissen THE, Goddard ME. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics. 2010;185:623-31.

10. Yu X, Meuwissen THE. Using the Pareto principle in genome-wide breeding value estimation. Genet Sel Evol. 2011;43:35.

11. Meuwissen THE. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet Sel Evol. 2009;41:35.

12. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One. 2008;3:e3395.

13. Goddard ME. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009;136:245-57.

14. Falconer DS, Mackay TFC. Introduction to quantitative genetics. 4th ed. London: Longman Group Ltd; 1996.

15. Corbin LJ, Liu AY, Bishop SC, Woolliams JA. Estimation of historical effective population size using linkage disequilibria with marker data. J Anim Breed Genet. 2012;129:257-70.

16. Sonesson AK, Meuwissen THE. Testing strategies for genomic selection in aquaculture breeding programs. Genet Sel Evol. 2009;41:37.

17. Hill WG, Goddard ME, Visscher PM. Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet. 2008;4:e1000008.

18. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819-29.

19. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414-23.

20. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics. 2011;12:186.

21. Meuwissen THE, Solberg TR, Shepherd R, Woolliams JA. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genet Sel Evol. 2009;41:2.

22. Hayes BJ, Fries R, Lund MS, Boichard DA, Stothard P, Veerkamp RF, et al. 1000 Bull Genomes Consortium Project [Abstract]. In: Proceedings of Plant and Animal Genome XX Conference: 14-18 January 2012; San Diego. 2012.

23. VanRaden PM, O'Connell JR, Wiggans GR, Weigel KA. Genomic evaluations with many more genotypes. Genet Sel Evol. 2011;43:10.

24. Druet T, Macleod IM, Hayes BJ. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity. 2014;112:39-47.

25. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 2010;185:1021-31.

26. Ober U, Ayroles JF, Stone EA, Richards S, Zhu D, Gibbs RA, et al. Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. PLoS Genet. 2012;8:e1002685.

27. De Roos APW. Genomic selection in dairy cattle. Ph.D. thesis. Wageningen: Wageningen University; 2011.

28. Olson KM, VanRaden PM, Tooker ME. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. J Dairy Sci. 2012;95:5378-83.

29. Pryce JE, Gredler B, Bolormaa S, Bowman PJ, Egger-Danner C, Fuerst C, et al. Short communication: Genomic selection using a multi-breed, across-country reference population. J Dairy Sci. 2011;94:2625-30.

30. Clark SA, Hickey JM, Daetwyler HD, van der Werf JH. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet Sel Evol. 2012;44:4.

31. Habier D, Tetens J, Seefried FR, Lichtner P, Thaller G. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol. 2010;42:5.

32. Wientjes YC, Veerkamp RF, Calus MP. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics. 2013;193:621-31.

33. Wientjes YCJ, Veerkamp RF, Bijma P, Bovenhuis H, Schrooten C, Calus MPL. Empirical and deterministic accuracies of across-population genomic prediction. Genet Sel Evol. 2015;47:5.

34. Bouwman AC, Veerkamp RF. Consequences of splitting sequencing effort over multiple breeds on imputation accuracy. BMC Genet. 2014;15:105.

35. Coster A, Bastiaansen JW, Calus MP, van Arendonk JA, Bovenhuis H. Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. Genet Sel Evol. 2010;42:9.

36. 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;183:1119-26.

37. Cole JB, VanRaden PM, O'Connell JR, Van Tassell CP, Sonstegard TS, Schnabel RD, et al. Distribution and location of genetic effects for dairy traits. J Dairy Sci. 2009;92:2931-46.

38. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, et al. Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci. 2009;92:16-24.

39. Hill WG, Robertson A. Linkage disequilibrium in finite populations. Theor Appl Genet. 1968;38:226-31.

40. De Roos AP, Hayes BJ, Spelman RJ, Goddard ME. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics. 2008;179:1503-12.

41. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet. 2007;81:559-75.

Figures

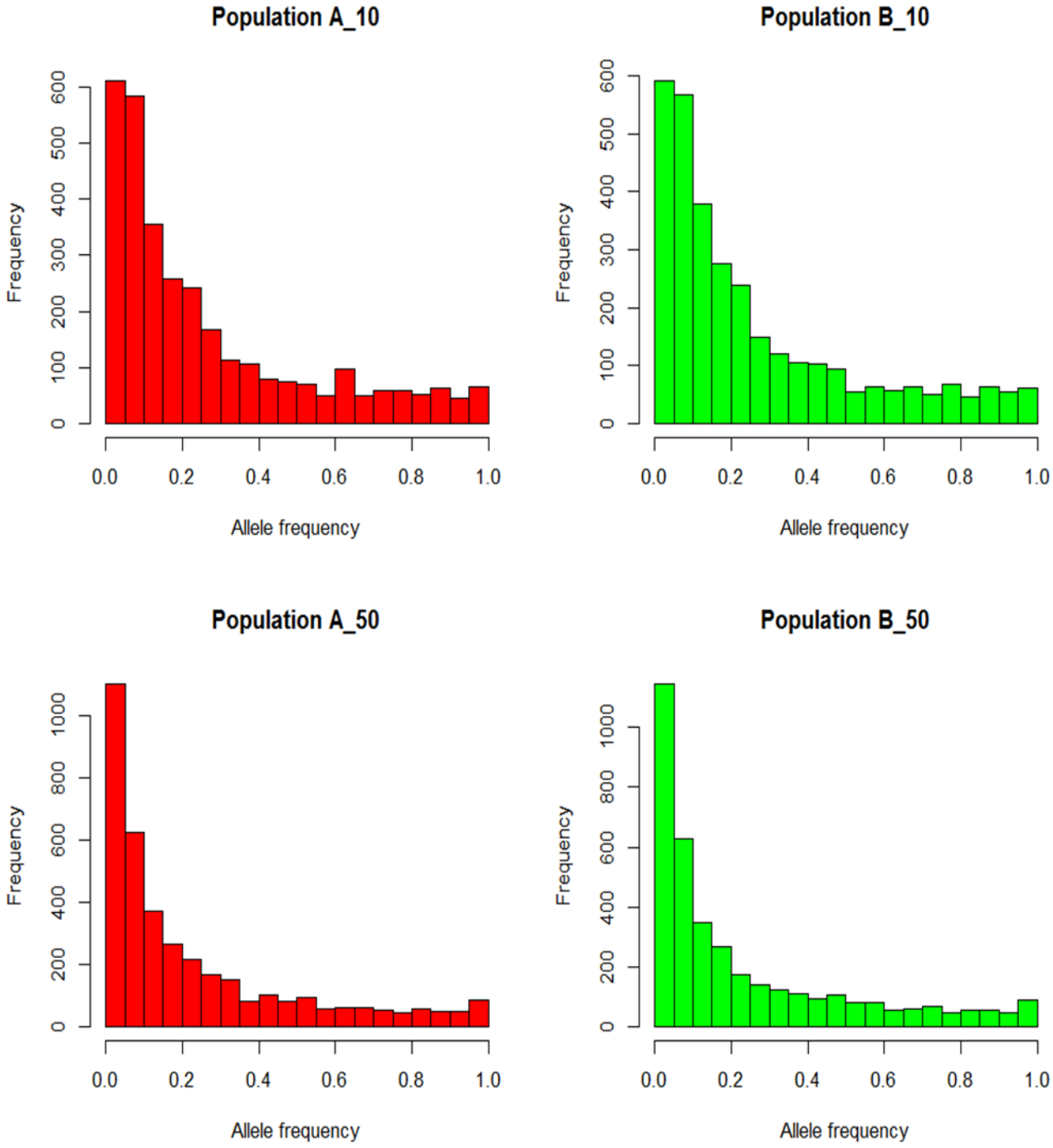

Figure 1 Distribution of allele frequencies in populations A and B at 10 and 50 generations of divergence.

A_10 (B_10) and A_50 (B_50) refers to different times of divergence (T= 10 or 50) between both the two populations. The plots are the result of one replicate. SNP alleles that were fixed in both populations were excluded.

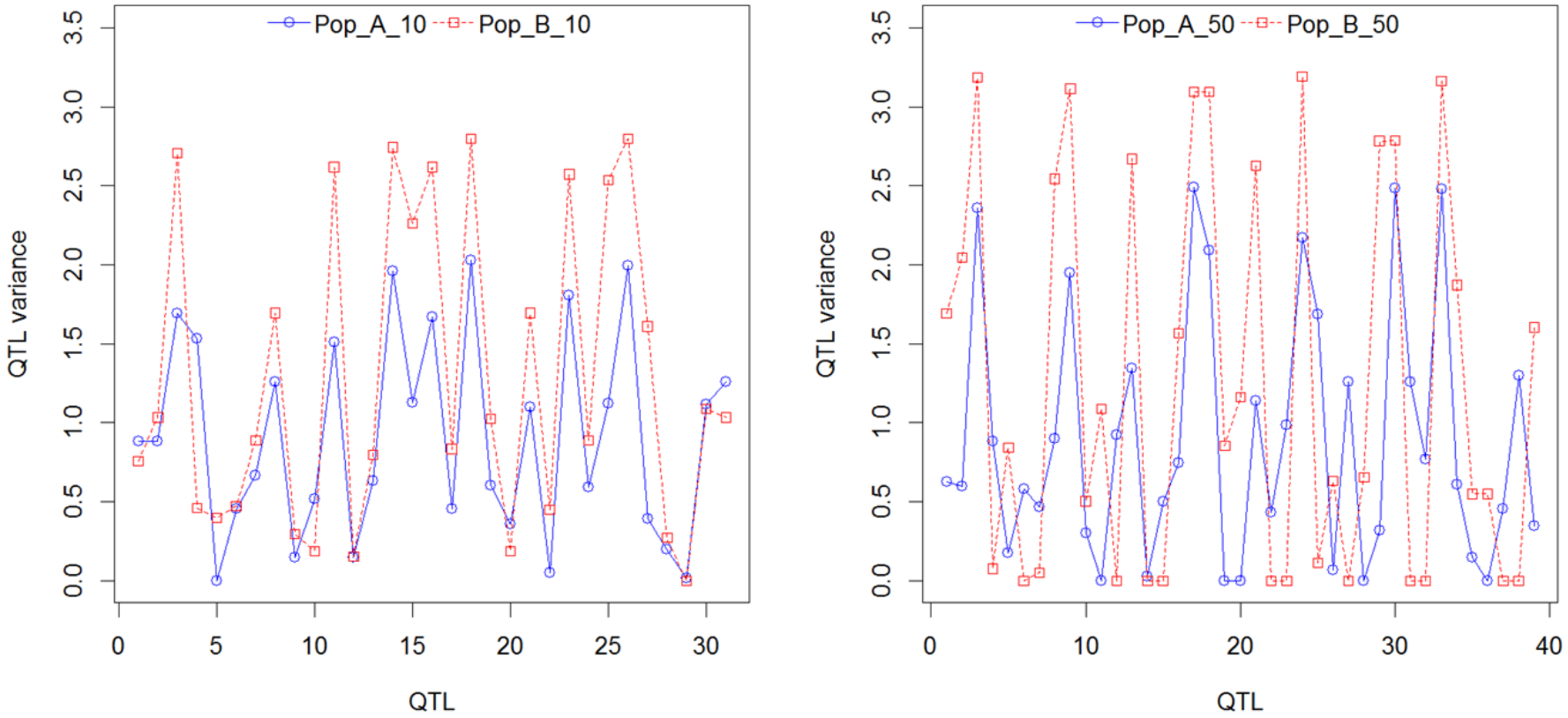

Figure 2 QTL variance for one of the replicates of populations A and B at 10 and 50 generations of divergence.

QTL variance was calculated as $2pqa^2$. QTL that were fixed in both populations were excluded. Pop_A and Pop_B refers to populations A and B at T= 10 or 50 generations of divergence, respectively.

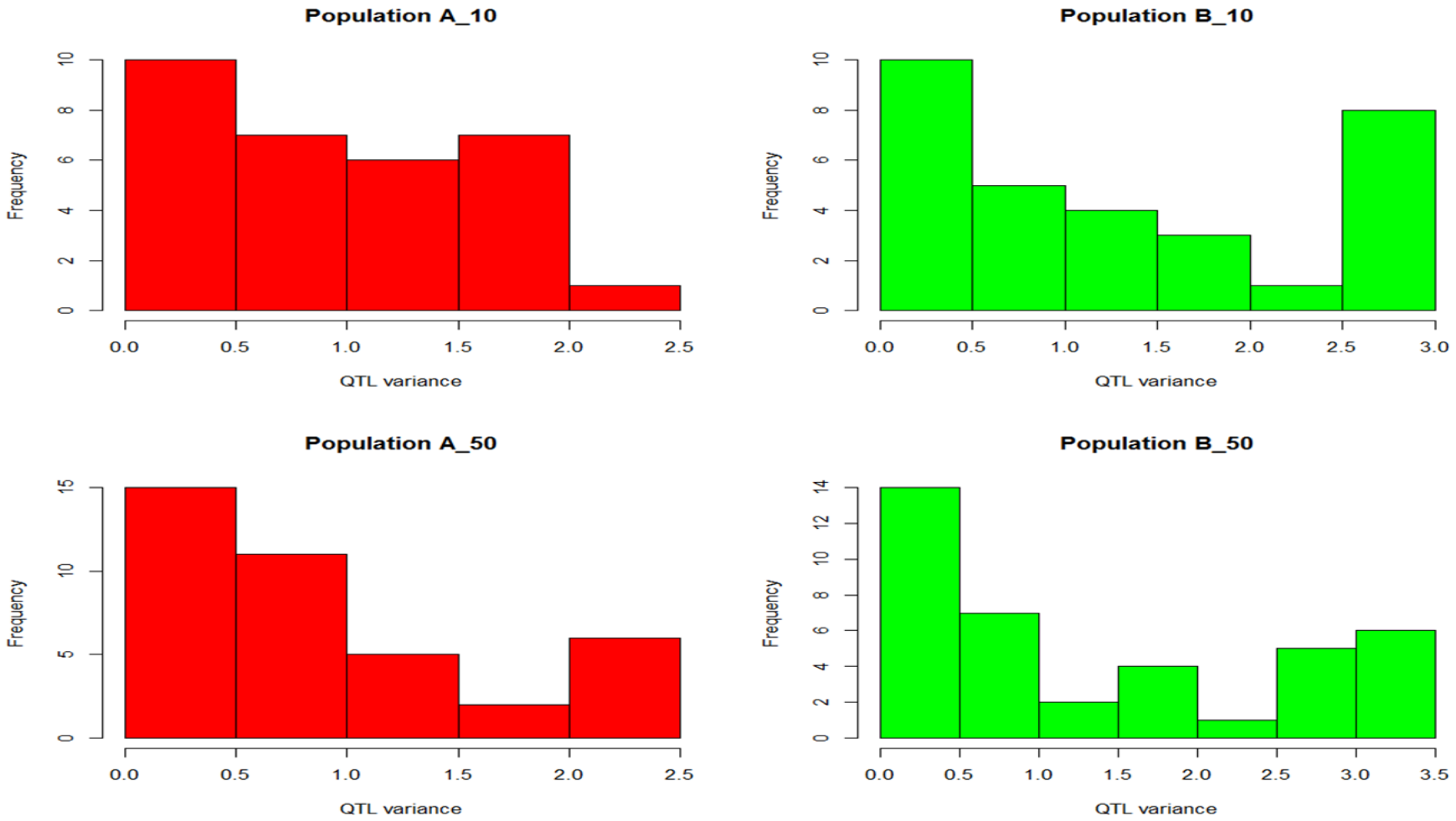

Figure 3 Distribution of QTL variance for populations A and B at 10 and 50 generations of divergence.

QTL variance was calculated as $2pqa^2$. A_10 (B_10) and A_50 (B_50) refer to different times of divergence (T= 10 or 50) between both populations. The plots are the result of one replicate. QTL that were fixed in both populations were excluded.

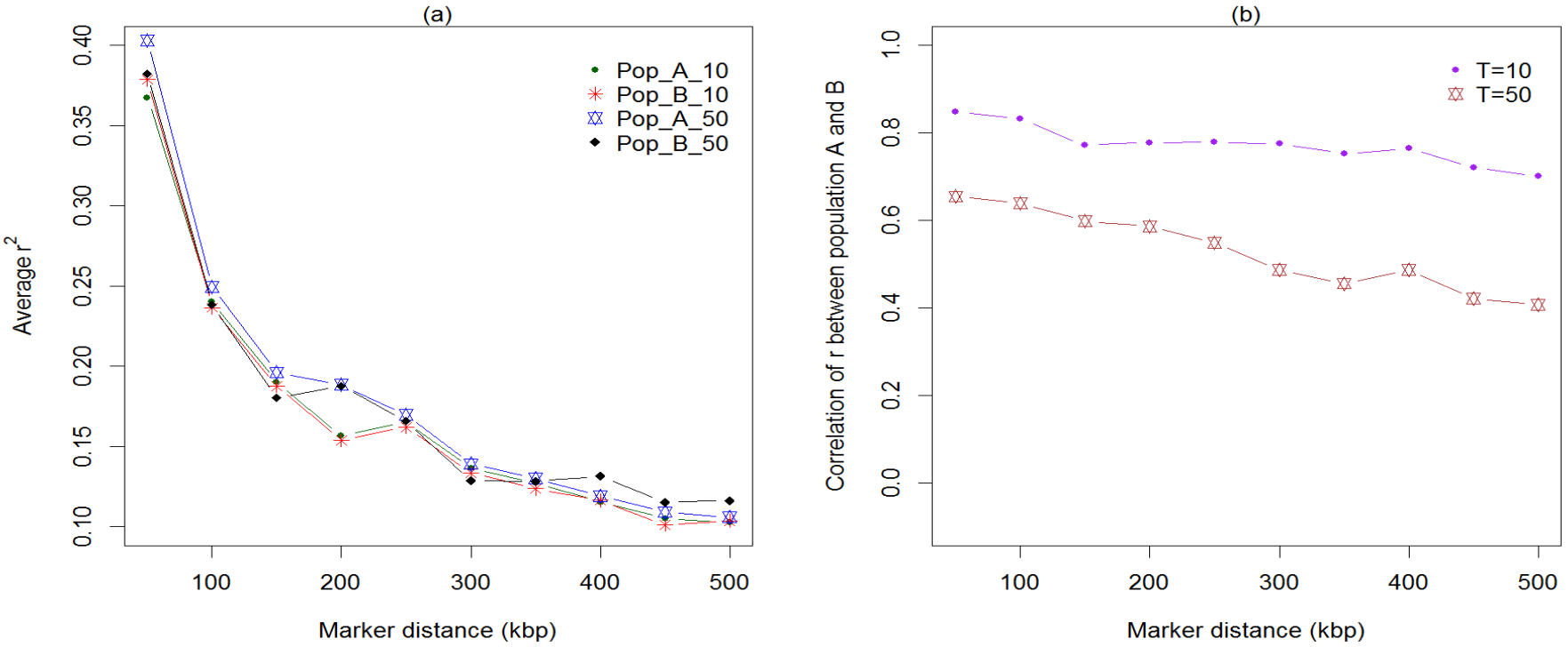

Figure 4 Linkage disequilibrium (r^2) and persistency of phase (r) as a function of genomic distance.

(a) Average linkage disequilibrium (LD) between SNPs estimated according to [39]. Pop_A_10 and Pop_B_10 refer to divergence of populations A and B by 10 generations while _50 refers to divergence by 50 generations.

(b) Persistency of LD phase (i.e. the correlation of LD between populations A and B, [40]). Calculations are within populations A and B at different times of divergence (T= 10 or 50). Values are binned at an interval of 50 kb. The plots are the result of one replicate of simulated WGS data. Calculations were done with PLINK [41].

Tables

Table 1 Population structure and parameters used in the simulation

Number of chromosomes	1
Genome length	1 Morgan
Mutation rate	10^{-8}/bp/meiosis
Effective population size (Ne)	200
Recombination	Haldane map function
QTL density	45 or 132 / Morgan
QTL effects	Normal distribution
Number of generations	1,950 + 50 or 1,990 + 10
Heritability	0.30 and 0.07

Table 2 Accuracy of genomic prediction (±se) for a trait with heritability 0.30 for population A based on SNP-BLUP or MixP using the different datasets

	45 QTL/Morgan		132 QTL/Morgan	
Dataset	**Accuracy**	**% decrease**	**Accuracy**	**% decrease**
SNP-BLUP				
WGS data	0.596 (± 0.015)		0.582 (± 0.014)	
data3000	0.578 (± 0.016)	3.0	0.575 (± 0.014)	1.2
data2000	0.576 (± 0.014)	3.4	0.568 (± 0.014)	2.4
data1000	0.564 (± 0.014)	5.4	0.555 (± 0.015)	4.6
data200	0.473 (± 0.015)	20.0	0.468 (± 0.014)	19.6
MixP				
WGS data	0.632 (± 0.018)		0.587 (± 0.014)	
data3000	0.598 (± 0.018)	5.4	0.579 (± 0.014)	1.4
data2000	0.591 (± 0.015)	6.5	0.573 (± 0.014)	2.4
data1000	0.579 (± 0.015)	8.4	0.562 (± 0.015)	4.3
data200	0.484 (± 0.016)	23.4	0.465 (± 0.014)	20.8

Accuracy of prediction was measured as the correlation between simulated true and predicted genetic values in the validation dataset

% decrease in accuracy of prediction relative to that obtained with WGS data

Table 3 Accuracy of genomic prediction (±se) for a trait with heritability 0.07 for population A based on SNP-BLUP or MixP using the different datasets

	45 QTL/Morgan		132 QTL/Morgan	
Dataset	**Accuracy**	**% decrease**	**Accuracy**	**% decrease**
SNP-BLUP				
WGS data	0.413 (± 0.024)		0.347 (± 0.019)	
data3000	0.400 (± 0.023)	3.1	0.332 (± 0.019)	4.3
data2000	0.394 (± 0.021)	4.6	0.326 (± 0.019)	6.1
data1000	0.377 (± 0.023)	8.7	0.326 (± 0.019)	6.1
data200	0.326 (± 0.021)	20.6	0.273 (± 0.016)	21.1
MixP				
WGS data	0.431 (± 0.028)		0.348 (± 0.019)	
data3000	0.407 (± 0.025)	5.6	0.333 (± 0.019)	4.3
data2000	0.404 (± 0.023)	6.3	0.327 (± 0.019)	6.0
data1000	0.382 (± 0.024)	11.4	0.328 (± 0.019)	5.7
data200	0.338 (± 0.023)	21.6	0.276 (± 0.016)	20.7

Accuracy of prediction was measured as the correlation between simulated true and predicted genetic value in the validation dataset

% decrease in accuracy of prediction relative to that obtained with WGS data

Table 4 Accuracy of across-population genomic prediction (se) for heritability 0.30 or 0.07 when populations have diverged for T (10 or 50) generations: population A (reference) and population B (validation) based on SNP-BLUP or MixP using the different datasets

	$h^2 = 0.30$				$h^2 = 0.07$			
	T = 10		T = 50		T = 10		T = 50	
Dataset	**SNP-BLUP**	**MixP**	**SNP-BLUP**	**MixP**	**SNP-BLUP**	**MixP**	**SNP-BLUP**	**MixP**
45 QTL/Morgan								
WGS data	0.396 (0.017)	0.482 (0.023)	0.276 (0.021)	0.360 (0.030)	0.270 (0.018)	0.286 (0.022)	0.171 (0.023)	0.197 (0.029)
data3000	-0.004 (0.015)	0.008 (0.016)	0.000 (0.022)	0.026 (0.023)	0.015 (0.016)	0.015 (0.017)	-0.003 (0.019)	-0.003 (0.018)
data2000	0.001 (0.018)	0.001 (0.027)	0.001 (0.017)	-0.002 (0.017)	0.030 (0.018)	0.032 (0.018)	-0.003 (0.017)	-0.017 (0.023)
data1000	0.020 (0.019)	0.009 (0.019)	-0.023 (0.019)	-0.005 (0.020)	-0.013 (0.015)	-0.017 (0.015)	-0.009 (0.014)	-0.005 (0.015)
data200	0.001 (0.013)	-0.010 (0.015)	-0.030 (0.021)	-0.012 (0.017)	0.002 (0.021)	-0.001 (0.022)	-0.034 (0.021)	-0.044 (0.026)
132 QTL/Morgan								
WGS data	0.392 (0.021)	0.403 (0.021)	0.238 (0.019)	0.246 (0.019)	0.209 (0.016)	0.211 (0.015)	0.186 (0.018)	0.186 (0.019)
data3000	0.022 (0.017)	0.025 (0.017)	0.049 (0.023)	0.052 (0.015)	0.005 (0.016)	0.005 (0.017)	-0.018 (0.019)	-0.018 (0.019)
data2000	0.013 (0.017)	0.021 (0.019)	-0.003 (0.017)	-0.004 (0.017)	-0.026 (0.014)	-0.018 (0.012)	0.001 (0.015)	0.001 (0.015)
data1000	0.022 (0.017)	0.028 (0.018)	-0.001 (0.017)	-0.002 (0.017)	0.009 (0.017)	0.009 (0.017)	0.017 (0.018)	0.019 (0.017)
data200	0.005 (0.017)	0.003 (0.015)	-0.022 (0.014)	-0.019 (0.015)	0.002 (0.014)	0.005 (0.017)	-0.005 (0.017)	-0.005 (0.017)

Accuracy of prediction was measured as the correlation between simulated true and predicted genetic value in the validation dataset

Table 5 Accuracy of genomic prediction (se) for heritability 0.30 or 0.07 using a multi-breed reference population when populations have diverged for T (10 or 50) generations, based on SNP-BLUP or MixP using the different datasets

	h^2 = 0.30				h^2 = 0.07			
	T = 10		T = 50		T = 10		T = 50	
Dataset	**SNP-BLUP**	**MixP**	**SNP-BLUP**	**MixP**	**SNP-BLUP**	**MixP**	**SNP-BLUP**	**MixP**
45 QTL/Morgan								
WGS data	0.654 (0.013)	0.710 (0.015)	0.627 (0.013)	0.675 (0.019)	0.475 (0.021)	0.525 (0.025)	0.448 (0.023)	0.480 (0.028)
data3000	0.578 (0.016)	0.602 (0.018)	0.571 (0.016)	0.592 (0.019)	0.404 (0.023)	0.414 (0.024)	0.398 (0.024)	0.409 (0.026)
data2000	0.571 (0.015)	0.574 (0.016)	0.564 (0.016)	0.558 (0.017)	0.400 (0.022)	0.414 (0.025)	0.387 (0.022)	0.390 (0.024)
data1000	0.552 (0.017)	0.551 (0.019)	0.546 (0.017)	0.549 (0.017)	0.376 (0.022)	0.377 (0.023)	0.368 (0.022)	0.367 (0.022)
data200	0.424 (0.017)	0.406 (0.017)	0.428 (0.019)	0.423 (0.021)	0.281 (0.019)	0.273 (0.019)	0.295 (0.020)	0.289 (0.021)
132 QTL/Morgan								
WGS data	0.647 (0.011)	0.662 (0.011)	0.612 (0.014)	0.624 (0.014)	0.410 (0.016)	0.412 (0.017)	0.370 (0.018)	0.374 (0.019)
data3000	0.569 (0.012)	0.573 (0.013)	0.573 (0.015)	0.576 (0.015)	0.343 (0.016)	0.345 (0.016)	0.332 (0.019)	0.332 (0.020)
data2000	0.568 (0.012)	0.570 (0.013)	0.565 (0.015)	0.567 (0.015)	0.341 (0.018)	0.341 (0.018)	0.331 (0.019)	0.330 (0.019)
data1000	0.537 (0.012)	0.536 (0.012)	0.542 (0.016)	0.521 (0.015)	0.313 (0.013)	0.313 (0.013)	0.322 (0.019)	0.321 (0.020)
data200	0.427 (0.017)	0.416 (0.017)	0.430 (0.014)	0.411 (0.014)	0.233 (0.017)	0.224 (0.017)	0.246 (0.016)	0.237 (0.017)

Accuracy of prediction was measured as the correlation between simulated true and predicted genetic value in the validation dataset

Table 6 Accuracy of genomic prediction (se) for a trait with heritability 0.30 based on single- and multi-breed reference populations (RP) obtained with SNP-BLUP or MixP using the different datasets

	Single-breed				Multi-breed			
	RP = 200		RP = 400		RP = 400			
					T = 10		T = 50	
Dataset	**SNP-BLUP**	**MixP**	**SNP-BLUP**	**MixP**	**SNP-BLUP**	**MixP**	**SNP-BLUP**	**MixP**
45 QTL/Morgan								
WGS data	0.596 (0.015)	0.632 (0.018)	0.696 (0.014)	0.736 (0.017)	0.654 (0.013)	0.710 (0.015)	0.627 (0.013)	0.675 (0.019)
data3000	0.578 (0.016)	0.598 (0.018)	0.681 (0.013)	0.719 (0.015)	0.578 (0.016)	0.602 (0.018)	0.571 (0.016)	0.592 (0.019)
data2000	0.576 (0.014)	0.591 (0.015)	0.679 (0.013)	0.713 (0.014)	0.571 (0.015)	0.574 (0.016)	0.564 (0.016)	0.558 (0.017)
data1000	0.564 (0.014)	0.579 (0.015)	0.668 (0.013)	0.697 (0.014)	0.552 (0.017)	0.551 (0.019)	0.546 (0.017)	0.549 (0.017)
data200	0.473 (0.015)	0.484 (0.016)	0.566 (0.014)	0.576 (0.016)	0.424 (0.017)	0.406 (0.017)	0.428 (0.019)	0.423 (0.021)
132 QTL/Morgan								
WGS data	0.582 (0.014)	0.587 (0.014)	0.691 (0.011)	0.702 (0.012)	0.647 (0.011)	0.662 (0.011)	0.612 (0.014)	0.624 (0.014)
data3000	0.575 (0.014)	0.579 (0.014)	0.681 (0.010)	0.688 (0.011)	0.569 (0.012)	0.573 (0.013)	0.573 (0.015)	0.576 (0.015)
data2000	0.568 (0.014)	0.573 (0.014)	0.670 (0.012)	0.677 (0.012)	0.568 (0.012)	0.570 (0.013)	0.565 (0.015)	0.567 (0.015)
data1000	0.555 (0.015)	0.562 (0.014)	0.654 (0.012)	0.665 (0.012)	0.537 (0.012)	0.536 (0.012)	0.542 (0.016)	0.521 (0.015)
data200	0.468 (0.014)	0.465 (0.014)	0.568 (0.013)	0.563 (0.013)	0.427 (0.017)	0.416 (0.017)	0.430 (0.014)	0.411 (0.014)

Accuracy of prediction was measured as the correlation between simulated true and predicted genetic value in the validation dataset

Paper II

Comparison of genomic prediction methods using medium- and high-density single-nucleotide polymorphism datasets in Norwegian Red Cattle

O. O. M. Iheshiulor, J. A. Woolliams, M. Svendsen, T. Solberg, and T. H. E. Meuwissen

Manuscript

Comparison of genomic prediction methods using medium- and high-density single-nucleotide polymorphism datasets in Norwegian Red Cattle

O. O. M. Iheshiulor,*[1] J. A. Woolliams,*† M. Svendsen,‡ T. Solberg,‡ T. H. E. Meuwissen*

*Department of Animal and Aquacultural Sciences, Norwegian University of life Sciences, PO Box 5003, NO-1432 Ås, Norway

†The Roslin Institute (Edinburgh), Royal (DICK) School of Veterinary Studies, University of Edinburgh, Midlothian, EH25 9RG, Scotland, UK

‡GENO SA, Holsegata 22, 2317 Hamar, Norway

[1]Corresponding author

Oscar O.M. Iheshiulor

Department of Animal and Aquacultural Sciences, Norwegian University of life Sciences, PO Box 5003, NO-1432 Ås, Norway

+4767232658

oscar.iheshiulor@nmbu.no

INTERPRETIVE SUMMARY

Accuracy of genomic breeding values is of key importance for successful application of genomic selection. Presently, the variable and non-variable selection methods for genomic prediction are treated as distinct approaches. The aim of this study was to evaluate a method that incorporates aspects of both approaches. Comparisons are made with the commonly used genomic-BLUP, which does not select variables, and a variable selection method that fits a mixture of two normal distributions for single-nucleotide polymorphism effects using the Pareto principle. The results show that simultaneously incorporating aspects of both methods can improve accuracy over genomic-BLUP whilst being comparable to other variable selection methods.

ABSTRACT

The aim of this study was to evaluate the implementation of an iterative method (called GBC) that incorporates aspects of genomic-BLUP (G-BLUP) and BayesC for genomic predictions. Its relative performance was compared to pedigree-BLUP, G-BLUP, and MixP (a method that fits a mixture of two normal distributions for SNP (single-nucleotide polymorphism) effects using the Pareto principle). Two datasets were available for the analysis: 1) imputed medium-density (50K; MD) SNP dataset based on Illumina Bovine50K BeadChip, containing 48,249 SNPs and 3,244 records; and 2) imputed high-density (777K; HD) SNP dataset originating from the Illumina BovineHD chip and containing 539,665 SNPs and 3,164 records. Daughter yield deviations (DYD) were used as the response variables and the study was performed on somatic cell count, fat yield, milk yield, and protein yield. Accuracy of prediction was measured as correlations between DYD and predicted values, divided by the square root of the average reliability of the DYD. With the MD SNP dataset, GBC showed an advantage over G-BLUP for all traits, while in comparison to MixP, accuracy was slightly lower. With the HD

SNP dataset, GBC also performed better than G-BLUP and slightly below that of MixP except for fat yield where it gave higher prediction accuracy than both methods. The results show that incorporating aspect of G-BLUP and BayesC in a single model can improve accuracy of genomic prediction over the commonly used method: G-BLUP. On the other hand, MixP showed higher accuracies than G-BLUP for all traits studied and in most cases slightly higher than GBC. Thus, GBC is quite a flexible tool in the sense that it simultaneously incorporates aspects of G-BLUP and BayesC for genomic prediction, thereby exploiting family relationship while simultaneously accounting for genes of large effects. MixP on the hand seems to strike a good balance between genes of large and small effects using the Pareto principle. The application of both methods in genomic prediction merits further exploration.

Key words: genomic prediction, single-nucleotide polymorphism, Pareto principle, Norwegian Red

INTRODUCTION

Genomic selection (GS) is increasingly implemented as a means of estimating the genetic values of animals and plants. It is useful for predicting an individual's genetic risk of developing a complex disease or breeding values for parents to breed the next generations (Meuwissen et al., 2001; De Roos et al., 2009; Goddard et al., 2011). Hence, accuracy of breeding values is of key importance for successful application of GS (Goddard and Hayes, 2009; Luan et al., 2009; Lund et al., 2014). GS adoption has been propelled by two key factors: 1) the availability of high-density genotyping that is presently tending toward whole-genome sequencing and 2) the availability of statistical methods for estimation of breeding values. The latter is still ongoing and so far, no consensus on which is the best approach has been established. Presently, many methods are available for genomic prediction (GP), and they can be broadly classified into two groups: variable and non-variable selection methods (Daetwyler

et al., 2010). The G-BLUP is a typical example of a non-variable selection method while the Bayesian methods (e.g. BayesB) and others are variable selection methods, which are usually implemented by Markov Chain Monte Carlo (MCMC) algorithms. A major difference between methods lies in their prior assumptions about the marker effects, and the methods have been reviewed in detail by Neves et al. (2012) and De Los Campos et al. (2013). Currently, these methods are independently implemented and results from most empirical studies have shown both methods to yield similar prediction accuracies. In contrast, simulation studies reported significant differences between methods (e.g. Meuwissen et al. (2001); Meuwissen (2009)). This has been resolved by Daetwyler et al. (2010) where it was demonstrated that the numbers of QTL (quantitative trait loci) in relation to the structure of the genome was a major factor in this discrepancy.

Genomic prediction have been demonstrated to utilize two sources of information, i.e. genetic relationship among individuals and LD between SNPs and QTL (Habier et al., 2007). Both information sources are utilized differently by current GP methods (i.e. G-BLUP and variable selection methods). G-BLUP through the genomic relationship matrix (**G)** exploits all relationship information in a given population more comprehensively than the pedigree-based relationship (**A** matrix) (Habier et al., 2007; Odegard and Meuwissen, 2014; Odegard et al., 2014). On the other hand, variable selection methods are much more able to utilize LD information than G-BLUP (Habier et al., 2007) and LD enables mapping of QTL. Thus, methods that would exploit genetic relationship and LD might help increase prediction accuracy.

Here we develop an iterative method (referred to as GBC) that combines relationship information using the G-BLUP approach and LD between QTL and neighbouring SNPs using the BayesC (Habier et al., 2011) approach for genomic prediction. Predictions from GBC were compared to A-BLUP (i.e. traditional BLUP which is pedigree based), G-BLUP, and MixP

(Yu and Meuwissen, 2011), using real data from a dairy population of genotyped sires. MixP fits a mixture of two normal distributions for SNP effects using the Pareto principle. In simulation studies, MixP has been shown to yield higher accuracies than G-BLUP (Yu and Meuwissen, 2011; Iheshiulor et al., 2014; Yu et al., 2014). Its performance in real data was also evaluated here.

MATERIALS AND METHODS

Phenotypes. Daughter yield deviations (DYD; VanRaden and Wiggans (1991); Liu et al. (2004)) from 2013 routine genetic evaluations on 3,244 Norwegian Red bulls were used for genomic prediction of three production traits (Fat yield, KgF; Milk yield, KgM; and Protein yield, KgP) and one health indicator trait (somatic cell count, SCC). The DYD are an estimate of the average performance of each bull's daughters, corrected for all fixed and non-genetic random effects of the daughters and genetic effects of their mates, and were calculated for each trait. Calculation of reliabilities of DYD followed Fikse and Banos (2001), i.e. ($r^2_{DYD} = d_e/(d_e + K)$, where d_e is the effective number of daughters and $K = (4 - h^2)/h^2$). The average effective number of daughters was 173 with a standard deviation of ~34. The heritabilities (0.136 - 0.277) and reliabilities of DYD (0.858 - 0.927) that are used by GENO SA for routine genetic evaluations (www.geno.no) are presented in Table 1.

SNP Data. Two sets of imputed SNP data on 3,244 (3,164) progeny-tested Norwegian Red bulls were kindly provided by GENO SA (www.geno.no) for the analyses. A total of 2,450 bulls were genotyped with the 25K Affymetrix chip (Affymetrix Inc., Santa Clara, CA), 1,650 were genotyped with the Illumina Bovine50K BeadChip (Illumina Inc., San Diego, CA), while 856 of the bulls were genotyped with both chips. Also genotyped were 384 bulls with the 777K Illumina BovineHD chip (Illumina Inc., San Diego, CA). Bulls genotyped with the 25K Affymetrix chip were imputed to 50K (i.e. 25K/50K) and afterwards to 777K (i.e.

25K/50K/777K). Genotype imputation was performed by CIGENE (www.cigene.no). Imputation procedures applying Beagle package (Browning and Browning, 2009) and other in-house developed software are described by Solberg et al. (2011) . Quality control involved: removal of animals with an individual call rate <97%; deletion of SNP with a call rate <25%; pedigree relationships between parent and offspring were set to missing if they exceeded Mendelian error threshold of >1%; SNPs with an overall Mendelian error rate >2.5% were deleted; then, for parent-offspring relationships with Mendelian errors <1%, genotypes for that loci were set to missing; finally, SNPs with minor allele frequency <0.05 were discarded. Following this, the 50K (777K) dataset contained 48,249 (539,665) SNP for a total of 3,244 (3,164) Norwegian Red bulls. The 50K SNP dataset is hereafter referred to as medium-density (MD) SNP dataset and the 777K referred to as high-density (HD) SNP dataset. For both datasets, SNPs on the X chromosome were not included.

Reference and Validation Dataset. The reference set comprised of 3,091 animals for the MD SNP dataset and 3,040 animals for the HD SNP dataset. Reference animals were born between 1965 and 2005. The validation scheme was based on forward predictions following a standard animal breeding selection scheme. Hence, validation dataset consisted of the youngest sires born between 2007 and 2008. The 124 youngest sires having more than 100 daughters with lactation records were used for validation. As a check of relationship, based on the genomic relationship according to VanRaden (2008), four measures of genomic relatedness were calculated between each animal in the validation set and the animals in the reference population (Table 2) following Clark et al. (2012) and Daetwyler et al. (2013). The four measures of genomic relatedness calculated includes: 1) an animal's mean relationship with the reference population (meanRel); 2) its maximum relationship (Relmax); 3) its average top 5 relationships (Rel5), and 4) its average top 10 relationships (Rel10).

Statistical Analyses

Analysis of Data. Three prediction methods, G-BLUP, MixP, and GBC, were implemented for genomic prediction.

G-BLUP. The G-BLUP model (Meuwissen et al., 2001; VanRaden, 2008) used to predict GEBV was as follows:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Zg} + \boldsymbol{e}, \qquad [1]$$

where **y** is a vector of records; **1** is a vector of ones; μ is the overall mean; $\boldsymbol{Z}$ is a design matrix that maps the records to genomic values; $\boldsymbol{g}$ is a vector of genomic values assumed to follow a multivariate normal distribution $MVN(0,\boldsymbol{G}\sigma_g^2)$, where $\boldsymbol{G}$ is the genomic relationship matrix and σ_g^2 is the genetic variance; and $\boldsymbol{e}$ is the vector of residuals, assumed to follow a multivariate normal distribution $MVN(0,\sigma_e^2)$. $\boldsymbol{G}$ was calculated, following Method 1 of VanRaden (2008), as $\boldsymbol{G} = \boldsymbol{MM'}/2\sum p_j\,(1-p_j)$**,** and $M_{ij} = x_{ij} - 2p_j$, where x_{ij} is the genotype of animal i for SNP j, with x_{ij} = 0**,** 1 or 2 for reference homozygote, heterozygote and alternative homozygote respectively, and p_j is the allele frequency of SNP j.

MixP. MixP assumes that SNP effects come from a mixture of two normal distributions which differ in variances, i.e. one with a large variance (σ_1^2) and the other with small variance (σ_2^2) (Yu and Meuwissen, 2011). The distribution of the total genetic variance (Vg) over the 'large' and the 'small' SNPs is according to the Pareto principle so that x% of the SNPs with the largest effects cause $(100-x)$% of the genetic variance. Given the prior value (π) for the mixing frequency ($\pi = x/100$) and using the Pareto principle, the variances of the large and small SNP effects are respectively:

$$\left.\begin{aligned}\sigma_1^2 &= \frac{(1-\pi)V_g}{\pi N_m}\\ \sigma_2^2 &= \frac{\pi V_g}{(1-\pi)N_m}\end{aligned}\right\}, \qquad [2]$$

where N_m is the number of SNPs, such that $N_m(\pi\sigma_1^2 + (1-\pi)\sigma_2^2) = V_g$. The π values used were estimated from the dataset via a search between 1% and 30% to get the optimal π values. MixP uses the Iterative Conditional Expectation (ICE) algorithm of Meuwissen et al. (2009). A detailed description of the MixP method is provided in Yu and Meuwissen (2011). The model to describe the records, based on SNP effects is as follows:

$$\boldsymbol{y} = 1\mu + \boldsymbol{Z}\mathbf{M}\mathbf{q} + \mathbf{e}\,, \qquad [3]$$

where **y** is a vector of records; **1** is a vector of ones; μ is the overall mean; $\boldsymbol{M}$ is the design matrix of standardized SNP genotypes as in the calculation of $\boldsymbol{G}$ above; $\boldsymbol{q}$ is the vector of SNP effects (q_i), and **e** is the vector of residuals. The model assumed that q_i ~ N(0, σ_1^2) with probability π or q_i ~ N(0, σ_2^2) with a probability (1 - π). Given the estimates of SNP effects, the genetic value for the validation animals was calculated as the sum of all estimated SNP effects multiplied with the standardized SNP genotypes.

GBC. The GBC fits simultaneously a polygenic effect (as in G-BLUP) and a BayesC effect (Habier et al., 2011) using the ICE algorithm (Meuwissen et al., 2009), which includes a correction for the uncertainty of the other SNP effects when deciding whether SNP '*i*' has an effect or not as described by Wang et al. (2015). The polygenic term is expected to capture the part of the breeding value that is explained by relationships between the animals and co-segregation of alleles within the families, whilst the BayesC term is expected to pick-up the LD between SNPs and major genes. The model of analysis used by GBC, based on fitting G-BLUP and SNPs with large effects is thus:

$$\boldsymbol{y} = 1\mu + \boldsymbol{Z}\boldsymbol{g} + \boldsymbol{Z}\mathbf{M}\mathbf{Q}\mathbf{q} + \mathbf{e}\,, \qquad [4]$$

where **y** is a vector of records; **1** is a vector of ones; μ is the overall mean; $\boldsymbol{Z}$ is a design matrix that maps the records to genomic values; $\boldsymbol{g}$ is vector of polygenic effects (i.e. G-BLUP aspect of the model) with distributional assumptions as described above for G-BLUP, $\boldsymbol{M}$ is a design

matrix of standardized SNP genotypes as defined in Equation 1 above, $\boldsymbol{q}$ is the vector of SNP effects (q_i), and $\boldsymbol{e}$ is the vector of residuals; **Q** is a diagonal matrix with 1 on the diagonal if the SNP has a large BayesC effect (with prior probability π) and 0 if it has no such effect (with prior probability (1-π)). In the ICE algorithm the elements of $\boldsymbol{Q}$ are obtained using the posterior probabilities (i.e. prior probability times likelihood) that the SNP has large effect. The starting prior (π) probability of a SNP having large effects was 1%; however the program estimates the optimal π. The G-BLUP term was implemented as described in the G-BLUP section. The BayesC term was implemented as described in the MixP section, except that here the 'small' SNPs had genetic variance 0, whereas the 'large' genes were assumed to have a variance of $0.01\sigma_g^2$. Because fitting many 'small' SNP effects is equivalent to fitting a GBLUP term (e.g. Habier et al., 2007), the core difference between the GBC and MixP model is the variance assumed for the 'small' (i.e. the GBLUP term) and 'large' SNPs, which was in MixP assumed to follow the Pareto principle (Equation [2]).

A-BLUP. The A-BLUP ignores genomic information and relies on pedigree information from ancestors using a numerator relationship matrix (**A**). Using the data set, with the genomic information excluded, and the pedigree, A-BLUP predictions of EBV were obtained. The analysis was carried out using ASReml version 3.0 package (Gilmour et al., 2009).

Predictive Ability. A primary criterion used for the evaluation of the performances of the methods was the predictive ability (also referred to as accuracy of predictions (r)), calculated as the correlation between the predicted values i.e. GEBV or EBV (in the case of traditional BLUP) and DYD, divided by the square root of the average reliability of the DYD $\left(\sqrt{r_{DYD}^2}\right)$. The bias of predictions was measured as the regression of DYD on the predicted values. A regression coefficient of 1 denotes no bias, <1 implies that the spread of the (G)EBVs is too large, and >1 implies a too low spread of GEBVs. Standard errors of the prediction accuracies

and the regression coefficients on the DYD were computed using bootstrapping with the R-package (R Core Team, 2015). The bootstrapping procedure involved sampling with replacement of the GEBVs 10,000 times. For each bootstrap sample, the (G)EBVs were correlated to the DYD. Standard errors were computed from the 10,000 bootstraps of estimated accuracies and regression coefficients. A Hotelling-Williams test (Steiger, 1980) for dependent correlations was used to determine whether differences between the validation correlations using alternative methods were statistically significant.

RESULTS AND DISCUSSION

Genomic Relatedness of Validation to Reference Individuals

Table 2 shows the average genomic relatedness within reference individuals and of validation to reference individuals based on both datasets. Estimated genomic relatedness based on MD SNP dataset was similar to that based on HD SNP dataset. Overall mean relationship (meanRel) was 0.03, while maximum estimated genomic relationship (Relmax) between the validation and reference population was ~0.5, suggesting that some animals in the validation were closely related to the reference population (i.e. their sire is in the reference population). In the case of Rel5 and Rel10, genomic relatedness estimates of 0.29 and 0.24, respectively were obtained. Based on Relmax and Rel5 (10), we can generally say, our validation set shares a certain degree of relationship with the reference population. On one hand, this is beneficial in the sense that accurate predictions could be expected. Several studies have shown this to be true (e.g. Habier et al. (2007); Meuwissen (2009); Habier et al. (2010)). On the other hand, the latter authors and others also demonstrated that accuracies obtained when validation and reference individuals are considerably related tend to decay across generations as relationship decreases.

Medium Density Versus High Density SNP Dataset

Accuracies of predictions for the youngest sires based on the MD and HD SNP datasets are presented in Table 3. In almost all cases, there was no advantage of HD SNP dataset over the MD SNP dataset. On average across all traits, prediction accuracies ranged between 0.679 to 0.704 for MD SNP dataset and 0.678 to 0.692 for HD SNP dataset. Overall, higher prediction accuracies were observed for the production traits (i.e. KgF, KgM, and KgP) than for the health indicator trait, SCC.

The LD between markers and QTL is known to affect accuracy of GP (Meuwissen et al., 2001). LD structure is largely dependent on effective population size (Ne) (Goddard and Hayes, 2009). This means that for populations with large Ne, LD is expected to be large and vice-versa for populations with small Ne. The degree of LD in Norwegian Red cattle is low in comparison to Holstein due to the relative large Ne (>100) (Solberg et al., 2011). So, to better capture LD's in this population, HD SNP panels have been suggested (Solberg et al., 2011). Although, LD was not examined in this study, previous studies (e.g. Calus et al. (2008); Solberg et al. (2008); Meuwissen (2009)) have shown that with increased maker density, much stronger LD between markers and QTL are realized. Such is expected to lead to increased accuracy of prediction when using HD SNP dataset, however in this present study an increase was not observed across all traits studied. This result agrees with previous studies such as Erbe et al. (2012), Su et al. (2012) and VanRaden et al. (2013) who reported little or no gain or even a decrease in accuracy when going from 50K to HD SNP dataset. While increasing marker density leads to increased SNP-QTL associations, it also leads to increased number of uninformative SNPs and linear functions of the uninformative SNPs may predict random errors in the reference phenotypes (Harris and Johnson, 2010; VanRaden et al., 2013). Another possible explanation for the lack of benefit of HD over MD SNP dataset could be due to the size of the reference population. In both datasets, the reference population comprised of ~3,000 individuals, but the number of

SNPs increased about 11 times more (i.e. 50K to HD). Increasing SNP density without proportionally increasing the reference population decreases the chance of accurately detecting causative mutations and to distinguish their effects from random noise (De Roos, 2011). Thus, increasing SNP density without increasing reference population might lead to diminishing return (VanRaden et al., 2011). Therefore, to take full advantage of high-density markers or even the commonly used 50K SNP density, large reference populations are needed (Goddard and Hayes, 2009; Meuwissen, 2009; Daetwyler et al., 2010; VanRaden et al., 2011).

Prediction Methods

Table 3 also shows the accuracies of predictions using alternative prediction methods. A-BLUP, which is pedigree based was less accurate than the genomic methods. The improvement of GS methods over A-BLUP was statistically significant using the Hotelling-Williams test ($P<0.05$). Within the GS methods, G-BLUP had significantly lower accuracy relative to GBC and MixP only for KgF when using the HD dataset. With A-BLUP, accuracies across the 4 traits ranged between 0.437 – 0.586 with the highest accuracy observed for KgF. With G-BLUP, GBC, and MixP, accuracies ranged between 0.602 – 0.750 (MD SNP dataset) and 0.568 – 0.759 (HD dataset), with the highest accuracy again observed for KgF. G-BLUP gave slightly higher accuracy (0.594) than GBC (0.568) for SCC when using the HD dataset. GBC gave the highest accuracy only for KgF (0.759, HD dataset). MixP resulted in the highest accuracy across all traits for both datasets except for KgM when using HD SNP dataset. Apart from the aforementioned, we observed that GBC and MixP were frequently, slightly superior to the G-BLUP regardless of the dataset. Except for KgF when using the HD dataset, none of the observed differences was significant.

The significantly higher prediction accuracies of GS methods over A-BLUP is in part due to improved predictions of the Mendelian sampling component of the breeding values, which accounts for half the additive genetic variance among animals (Daetwyler et al., 2007). The

performance of MixP is in agreement with the studies of Yu and Meuwissen (2011) and Iheshiulor et al. (2014). For GBC, we anticipated that by fitting a genomic polygenic term next to BayesC SNP effects, both models would complement each other: the G-BLUP term mainly picking up effects that could be explained by linkage analysis (Habier et al., 2007), and the BayesC term picking up close LD between SNPs and genes. Consequently, we expected GBC to result in a high GP accuracy, however this was not observed convincingly. GBC performed always slightly better than G-BLUP and slightly inferior to MixP, except for KgF using the HD dataset where it performed better than both methods. A possible explanation could be that the modelled LD blocks surrounding the major genes were large and were also reasonably well captured by G-BLUP (Daetwyler et al., 2010). In addition, a considerable level of relationship exist in the dataset (Table 2). In that case, GBC results in a performance that is only slightly better than G-BLUP. As mentioned in the Methods section, MixP and GBC are somewhat similar models, differing mainly in their prior variances assumptions for large and small SNPs. If the Pareto principle based division of variances across large and small SNPs is appropriate, the MixP would yield best results. The GBC model assumes that there are some genes with large effect with variance ~$0.01\sigma_g^2$, which seems to be mainly true for KgF, as is known from the literature (Grisart et al., 2002).

The regression coefficients in Table 4 are a measure of bias of the EBV predictions. Except for the trait SCC where regression coefficients were below 1 (0.847 – 0.894), it was above 1 for the other 3 traits across the methods. This means that for the production traits, GEBVs were deflated while for SCC, GEBVs were inflated. In NRF, there is strong selection pressure against directly recorded mastitis. Since mastitis is quite highly correlated to SSC, and we did not include mastitis in our GP models, biased genomic predictions might have been expected for SSC. However, considering the fact that the regression coefficients from the A-BLUP

method did not differ from those of other methods, one may attribute the biases to the data sets rather than the (genomic) breeding value estimation methods.

SNP Effects: GBC and MixP

A key difference between GBC and MixP lies in the way both methods estimate SNP effects. GBC fits a BayesC like prior for the SNPs with large effects assuming an estimated fraction π of the SNPs having large effects and uses a G-BLUP component for the SNPs with small effects (i.e. polygenic effects). MixP on the other hand fits a mixture of two normal distributions where a fraction (π x 100%) of the SNPs have large effects and explain a fraction (1 - π) of the genetic variance while (1 – π) x 100% of the SNPs have small effects and explain a fraction π of the genetic variance. As an example, the SNP effects estimated by GBC and MixP using the MD SNP dataset are presented in Figures 1 and 2. Especially for the production traits, both methods behaved differently in the number of SNPs with effects and the magnitude of effects. For KgF, GBC picked up 2 SNPs with large effects on chromosome 5 and 12. The rest of the SNP effects were substantially shrunk towards zero. The same SNPs were observed to have large effects with MixP, however, several other SNPs with small to moderate effects were also observed by this method. For KgM, a similar trend was observed with both methods identifying SNPs on chromosome 6, 12, and 28 as regions with large effects. Chromosome 6 was also identified by both methods as a region with large effects on KgP. We did not try to identify candidate genes underlying these regions, as this was outside the scope of our study. A recent study by Raven et al. (2014) reported QTL affecting KgF on chromosomes 5 and 12 and QTL affecting KgM and KgP on chromosome 6. There was a general tendency that GBC yielded large effects to very few SNPs while MixP identified many more SNPs with moderate to large effects. This implies that, with GBC, only markers in high LD with the QTL pick up the major gene effects while the rest are treated as polygenic effects. MixP captured SNPs with moderate to large effect and in addition allowed neighboring SNPs to be associated with the traits. The observed

differences in prediction accuracies between both methods is a reflection of how genomic regions with large effects are treated. The ability of MixP to assign effects to SNPs with small to moderate effects may explain why MixP outperformed GBC in terms of accuracy of GP for some of the traits studied. It seems that, if the interest is on identifying QTL then GBC yields clearer QTL signals. Although the actual and prior distribution of SNP effects remains unknown, the results of this study indicate that the assumed prior distribution for SNP effects in MixP tends to yield somewhat higher accuracy than the assumptions underlying GBC.

CONCLUSIONS

We introduced GBC, which incorporates aspects of G-BLUP and BayesC approaches for genomic prediction. The method was evaluated using imputed MD and HD SNP datasets and its performance compared to A-BLUP, G-BLUP, and MixP. GBC showed (slight) advantage over G-BLUP for all traits and performed slightly inferior to MixP in terms of prediction accuracy. Only for KgF when using the HD SNP dataset did GBC perform better than both methods which agrees with the fact that KgF is known to be controlled by few genes with large effects. MixP outperformed G-BLUP for all traits studied and showed slight advantage over GBC in most cases. In conclusion, GBC is quite a flexible tool in the sense that it simultaneously incorporates aspects of variable and non-variable models for genomic prediction, thereby exploiting family relationships while also accounting for genes of large effects. MixP on the hand seems to strike a good balance between genes of large and small effects based on the Pareto principle.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 289592 - Gene2Farm. The authors thank Geno SA (Ås, Norway) for

providing the datasets and CIGENE (Ås, Norway) for the quality control and genotype imputation. Neither the European Commission nor the partners of the Gene2Farm project can be held responsible for views expressed in this manuscript.

REFERENCES

Browning, B. L. and S. R. Browning. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84:210-223.

Calus, M. P., T. H. E. Meuwissen, A. P. de Roos, and R. F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553-561.

Clark, S. A., J. M. Hickey, H. D. Daetwyler, and J. H. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet Sel Evol 44:4.

Daetwyler, H. D., M. P. Calus, R. Pong-Wong, G. de Los Campos, and J. M. Hickey. 2013. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193:347-365.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021-1031.

Daetwyler, H. D., B. Villanueva, P. Bijma, and J. A. Woolliams1. 2007. Inbreeding in genome-wide selection. J Anim Breed Genet 124:369-376.

De Los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. Calus. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327-345.

De Roos, A. P. W. 2011. Genomic selection in dairy cattle. in Animal Breeding and Genomic Centre. Vol. Ph.D. Wageningen University, the Netherlands, Wageningen.

De Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of genomic predictions across multiple populations. Genetics 183:1545-1553.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci 95:4114-4129.

Fikse, W. F. and G. Banos. 2001. Weighting factors of sire daughter information in international genetic evaluations. J Dairy Sci 84:1759–1767.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson. 2009. ASReml User Guide Release 3.0., VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Goddard, M. E. and B. J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet 10:381-391.

Goddard, M. E., B. J. Hayes, and T. H. Meuwissen. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet 128:409-421.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional Candidate Cloning of a QTL in Dairy Cattle: Identification of a Missense Mutation in the Bovine DGAT1 Gene with Major Effect on Milk Yield and Composition. Genome Res 12:222-231.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

Habier, D., J. Tetens, F. R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol 42:5.

Harris, B. L. and D. L. Johnson. 2010. The impact of high density SNP chips on genomic evaluation in dairy cattle. Interbull Bull 42:40-43.

Iheshiulor, O. O. M., J. A. Woolliams, X. Yu, R. Wellmann, and T. H. E. Meuwissen. 2014. Genomic Predictions Using Whole Genome Sequence Data and Multi-breed Reference Populations. in Proc. 10th World Congress of Genetics Applied to Livestock Production, Vancouver, Canada.

Liu, Z., F. Reinhardt, A. Bunger, and R. Reents. 2004. Derivation and calculation of approximate reliabilities and daughter yield-deviations of a random regression test-day model for genetic evaluation of dairy cattle. J Dairy Sci 87:1896–1907.

Luan, T., J. A. Woolliams, S. Lien, M. Kent, M. Svendsen, and T. H. E. Meuwissen. 2009. The accuracy of Genomic Selection in Norwegian red cattle assessed by cross-validation. Genetics 183:1119-1126.

Lund, M. S., G. Su, L. Janss, B. Guldbrandtsen, and R. F. Brøndum. 2014. Genomic evaluation of cattle in a multi-breed context. Livest Sci 166:101-110.

Meuwissen, T. H. E. 2009. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet Sel Evol 41:35.

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

Meuwissen, T. H. E., T. R. Solberg, R. Shepherd, and J. A. Woolliams. 2009. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genet Sel Evol 41:2.

Neves, H. H., R. Carvaheiro, and S. A. Queiroz. 2012. A comparison of statistical methods for genomic selection in a mice population. BMC Genetics 13:100.

Odegard, J. and T. H. Meuwissen. 2014. Identity-by-descent genomic selection using selective and sparse genotyping. Genet Sel Evol 46:3.

Odegard, J., T. Moen, N. Santi, S. A. Korsvoll, S. Kjoglum, and T. H. Meuwissen. 2014. Genomic prediction in an admixed population of Atlantic salmon (Salmo salar). Front Genet 5:402.

R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Raven, L.-A., B. G. Cocks, and B. J. Hayes. 2014. Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. BMC Genomics 15:62.

Solberg, T. R., B. Heringstad, M. Svendsen, H. Grove, and T. H. E. Meuwissen. 2011. Genomic predictions for production and functional traits in Norwegian Red from BLUP analyses of imputed 54K and 777K SNP data. Interbull Bull 44:240-243.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen. 2008. Genomic selection using different marker types and densities. J Anim Sci 86:2447-2454.

Steiger, J. H. 1980. Tests for comparing elements of a correlation matrix. Psychol Bull 87:245-251.

Su, G., R. F. Brondum, P. Ma, B. Guldbrandtsen, G. P. Aamand, and M. S. Lund. 2012. Comparison of genomic predictions using medium-density (approximately 54,000) and high-density (approximately 777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. J Dairy Sci 95:4657-4665.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J Dairy Sci 91:4414-4423.

VanRaden, P. M., D. J. Null, M. Sargolzaei, G. R. Wiggans, M. E. Tooker, J. B. Cole, T. S. Sonstegard, E. E. Connor, M. Winters, J. B. van Kaam, A. Valentini, B. J. Van Doormaal, M. A. Faust, and G. A. Doak. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. J Dairy Sci 96(1):668-678.

VanRaden, P. M., J. R. O'Connell, G. R. Wiggans, and K. A. Weigel. 2011. Genomic evaluations with many more genotypes. Genet Sel Evol 43:10.

VanRaden, P. M. and G. R. Wiggans. 1991. Derivation, Calculation, and Use of National Animal Model Information. J Dairy Sci 74:2737-2746.

Wang, T., Y. P. Chen, M. E. Goddard, T. H. Meuwissen, K. E. Kemper, and B. J. Hayes. 2015. A computationally efficient algorithm for genomic prediction using a Bayesian model. Genet Sel Evol 47:34.

Yu, X. and T. H. Meuwissen. 2011. Using the Pareto principle in genome-wide breeding value estimation. Genet Sel Evol 43:35.

Yu, X., J. A. Woolliams, and T. H. Meuwissen. 2014. Prioritizing animals for dense genotyping in order to impute missing genotypes of sparsely genotyped animals. Genet Sel Evol 46:46.

TABLES

Table 6 Heritability (h^2), reliability (r^2_{DYD})[1] of DYD[2], number of animals with phenotype and genotype per dataset[3]

Trait	h^2	r^2_{DYD}	Number of animals	
			MD SNP dataset	HD SNP dataset
Somatic cell count (SCC)	0.136	0.858	3,244	3,164
Fat yield (KgF)	0.213	0.906	3,244	3,164
Milk yield (KgM)	0.277	0.927	3,244	3,164
Protein yield (KgP)	0.235	0.915	3,244	3,164

[1] $r^2_{DYD} = d_e/(d_e + K)$, where d_e is the effective number of daughters and $K = (4 - h^2)/h^2$

[2] DYD = daughter yield deviation

[3] MD SNP dataset = imputed dataset based on Illumina Bovine50K BeadChip; HD = imputed dataset based on 777K Illumina BovineHD (Illumina, San Diego, CA)

Table 2 Average (SD) genomic relatedness[1] of validation to reference individuals based on both datasets[2]

	MD SNP dataset				HD SNP dataset			
Relatedness	meanRel	Relmax	Rel5	Rel10	meanRel	Relmax	Rel5	Rel10
Within reference	0.03 (0.01)	0.49 (0.04)	0.34 (0.05)	0.30 (0.05)	0.03 (0.01)	0.47 (0.05)	0.33 (0.05)	0.29 (0.05)
Between validation and reference	0.03 (0.00)	0.48 (0.09)	0.29 (0.05)	0.24 (0.05)	0.03 (0.00)	0.47 (0.07)	0.29 (0.05)	0.24 (0.05)

[1]Genomic relatedness: meanRel = average relationships $((1/N_P)\sum_{j=1}^{N_p} rel(i,j))$, where N_P is the number of individuals in the reference population, $rel(i,j)$ is the relationship between validation i and reference individual j; Relmax = maximum relationship between validation individual i and reference individual j; Rel5 = average of top 5 relationships $((1/5)\sum_{j=1}^{5} rel(i,j))$, top 5 refers to the 5 largest (i,j); Rel10 = average of top 10 relationships $((1/10)\sum_{j=1}^{10} rel(i,j))$, top 10 refers to the 10 largest (i,j)

[2]MD SNP dataset = imputed dataset based on 50K Illumina BeadChip; HD = imputed dataset based on 777K Illumina BovineHD (Illumina, San Diego, CA)

Table 3 Accuracy[1] (SE[2]) of the predicted values for the youngest sires based on MD and HD SNP dataset[3] and different prediction methods[4]

Trait[5]	A-BLUP[6]	G-BLUP	GBC[7]	MixP[7]
MD SNP data				
SCC	0.437 (0.082)	0.602 (0.066)	0.607 (0.065)	0.638 (0.058)
KgF	0.586 (0.062)	0.716 (0.049)	0.731 (0.047)	0.750 (0.042)
KgM	0.527 (0.064)	0.705 (0.051)	0.719 (0.048)	0.727 (0.046)
KgP	0.496 (0.073)	0.695 (0.053)	0.696 (0.051)	0.701 (0.049)
Average	0.512	0.679	0.688	0.704
HD SNP data				
SCC	0.437 (0.082)	0.594 (0.063)	0.568 (0.067)	0.595 (0.063)
KgF	0.586 (0.062)	0.709 (0.047)	0.759 (0.042)	0.740 (0.044)
KgM	0.527 (0.065)	0.721 (0.048)	0.725 (0.047)	0.736 (0.044)
KgP	0.496 (0.073)	0.687 (0.050)	0.691 (0.049)	0.694 (0.044)
Average	0.512	0.678	0.684	0.692

[1]Accuracy = $\frac{corr(DYD,(G)EBV)}{\sqrt{r^2_{DYD}}}$

[2]SE = $\sqrt{1 - accuracy^2/n - 2}$, where n is the number of individuals

[3]MD SNP dataset = imputed dataset based on Illumina Bovine50K BeadChip; HD = imputed dataset based on 777K Illumina BovineHD (Illumina, San Diego, CA)

[4]A-BLUP: traditional BLUP using pedigree-based relationship matrix; G-BLUP: genomic BLUP using genomic-based relationship matrix; GBC: an iterative method that fits a G-BLUP next to SNP effects with a BayesC prior; MixP: a variable selection method that fits a mixture of two normal distributions for SNP effects using the Pareto principle

[5]SCC = somatic cell count; KgF = fat yield; KgM = milk yield; KgP = protien yield

[6]A-BLUP results is same under MD and HD SNP dataset

[7]With MD SNP dataset, optimal π values (i.e. proportion of SNP having large effects) when using GBC (MixP) was 20% (10%) for SCC, 1% (15%) for KgP, and 10% (10%) for KgF and KgM. With HD SNP dataset, optimal π values when using GBC (MixP) was 1% (10%) for SCC, 20% (1%) for KgF, and 1% (1%) for KgM and KgP.

Table 4 Bias[1] (SE[2]) of the predicted values for the youngest sires based on MD and HD SNP dataset[3] and different prediction methods[4]

Trait[5]	A-BLUP	G-BLUP	GBC	MixP
MD SNP data				
SCC	0.887 (0.168)	0.882 (0.110)	0.880 (0.109)	0.894 (0.103)
KgF	1.376 (0.164)	1.277 (0.119)	1.221 (0.113)	1.516 (0.127)
KgM	1.646 (0.239)	1.532 (0.148)	1.392 (0.136)	1.875 (0.166)
KgP	1.537 (0.254)	1.509 (0.159)	1.286 (0.100)	1.660 (0.169)
HD SNP data				
SCC	0.887 (0.168)	0.847 (0.108)	0.886 (0.103)	0.852 (0.110)
KgF	1.376 (0.164)	1.265 (0.112)	1.320 (0.110)	1.610 (0.131)
KgM	1.646 (0.239)	1.520 (0.143)	1.554 (0.144)	1.936 (0.173)
KgP	1.537 (0.254)	1.503 (0.163)	1.605 (0.153)	1.825 (0.184)

[1]Bias: measured as the regression of DYD on the predicted values

[2]SE: computed from 10,000 bootstrap samples

[3]MD SNP dataset = imputed dataset based on Illumina Bovine50K BeadChip; HD = imputed dataset based on 777K Illumina BovineHD (Illumina, San Diego, CA)

[4]A-BLUP: traditional BLUP using pedigree-based relationship matrix; G-BLUP: genomic BLUP using genomic-based relationship matrix; GBC: an iterative method that fits a G-BLUP next to SNP effects with a BayesC prior; MixP: a variable selection method that fits a mixture of two normal distributions for SNP effects using the Pareto principle.

[5]SCC = somatic cell count; KgF = fat yield, KgM = milk yield; KgP = protien yield.

FIGURES

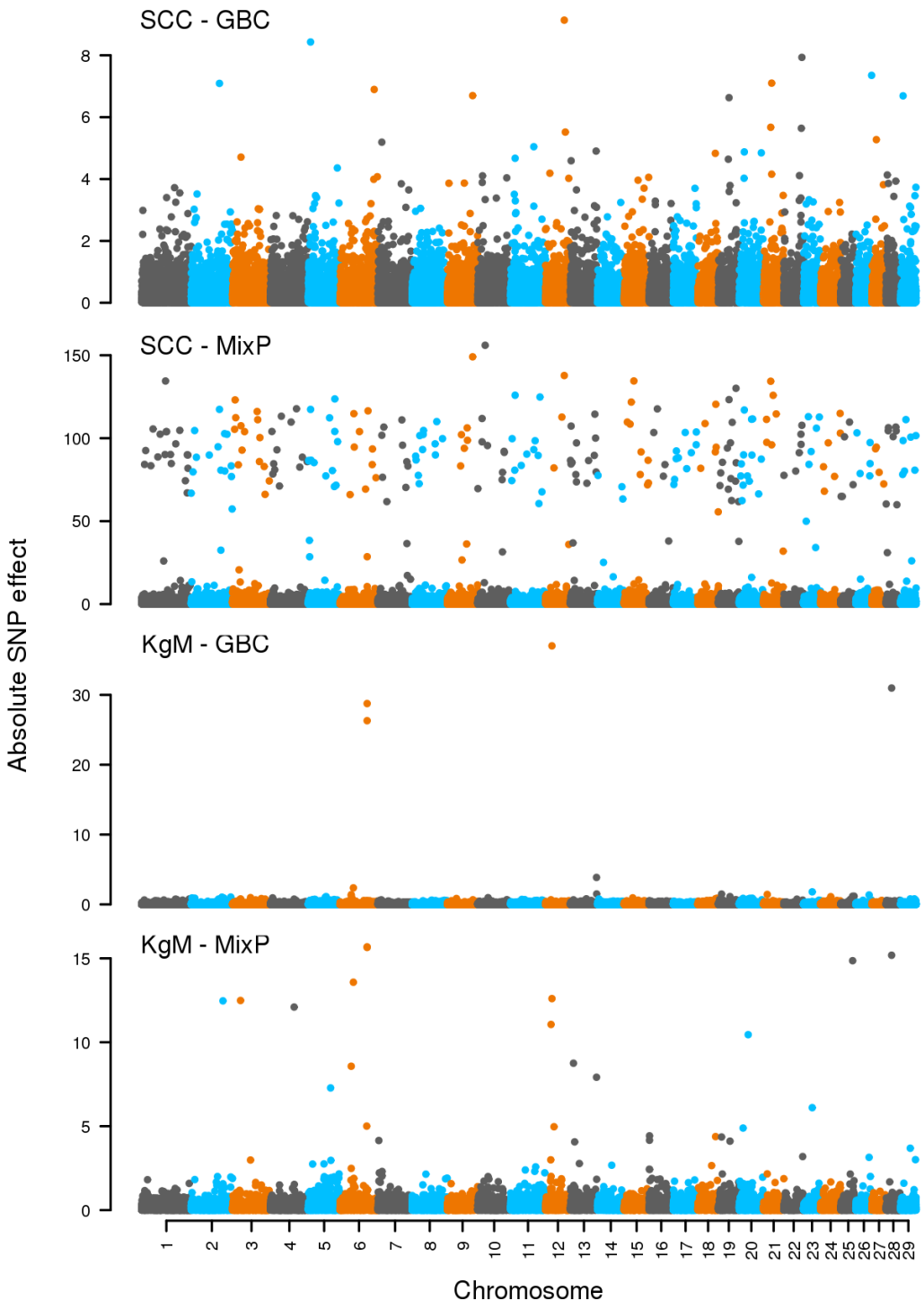

Figure 2 SNP effects estimated using GBC and MixP for SCC, and KgM.

Shown is the absolute value of SNP effect estimates (y-axis) from the MD SNP dataset (based on 54K Illumina BeadChip). X-axis is ordered by chromosomes from 1-29. SCC = somatic cell count, and KgM = milk yield. Note the changed y-axis scale for each graph. Absolute values of SCC was scaled by 10,000, while KgM was scaled by 1000. Scaling was just for plotting purpose.

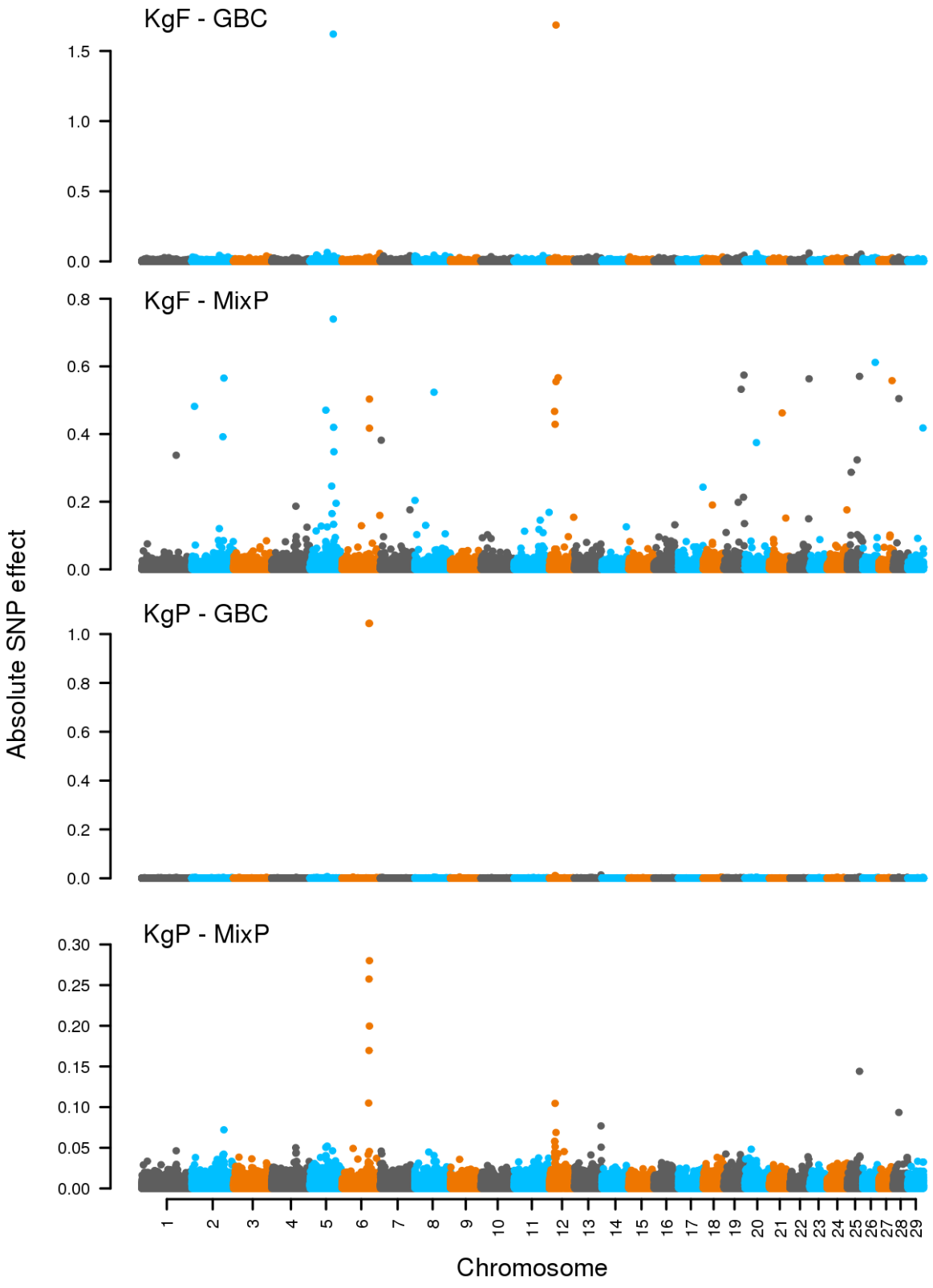

Figure 2 SNP effects estimated using GBC and MixP for KgF, and KgP.

Shown is the absolute value of SNP effect estimates (y-axis) from the MD SNP dataset (based on 54K Illumina BeadChip). X-axis is ordered by chromosomes from 1-29. KgF = fat yield, and KgP = protein yield. Note the changed y-axis scale for each graph.

Paper III

The absorption of large numbers of ungenotyped descendants in genomic predictions

O. O. M. Iheshiulor, J. A. Woolliams, and T. H. E. Meuwissen

Manuscript

Running head: ABSORPTION OF UNGENOTYPED DECENDANTS

The absorption of large numbers of ungenotyped descendants in genomic predictions

O. O. M. Iheshiulor,*[1] J. A. Woolliams,*† and T. H. E. Meuwissen*

*Department of Animal and Aquacultural Sciences, Norwegian University of life Sciences, PO Box 5003, NO-1432 Ås, Norway

†The Roslin Institute (Edinburgh), Royal (DICK) School of Veterinary Studies, University of Edinburgh, Midlothian, EH25 9RG, Scotland, UK

[1]Corresponding author

Oscar O.M. Iheshiulor

Department of Animal and Aquacultural Sciences, Norwegian University of life Sciences, PO Box 5003, NO-1432 Ås, Norway

+4767232658

oscar.iheshiulor@nmbu.no

INTERPRETIVE SUMMARY

Single-step genomic-BLUP exploits all available information, however, it extension to the variable selection model is not straightforward. We evaluated an absorption approach that absorbs phenotypic information on large number of ungenotyped animals into genotyped animals in-order to enable the utilization of all information in variable selection methods of genomic prediction. We implemented a variable selection method for genomic prediction on the Absorbed dataset. The results did not show an extra gain of using a variable selection method. Overall, the utilization of all available information led to less biased predictions.

ABSTRACT

The aim of this study was to evaluate an absorption approach that absorbs phenotypic information on large numbers of ungenotyped animals into the mixed model equations of genotyped animals in-order to enable the utilization of all information in variable selection methods of genomic prediction. Two datasets were available for the analysis: 1) DYD dataset, i.e. a combination of DYD (daughter yield deviations) and genotype of 3,244 progeny-tested bulls; and 2) Absorbed dataset, i.e. a combination of absorbed records and genotype probabilities of 20,918 animals. The absorbed records resulted from the absorption of phenotypic information on ungenotyped descendants (4,022,179) into the mixed model equations of genotyped animals. LDMIP was used to generate the genotype probabilities for ungenotyped ancestors. Data on milk yield, fat yield, protein yield, and somatic cell count, were analysed using pedigree-BLUP (A-BLUP), genomic-BLUP (G-BLUP), single-step-GBLUP (SS-GBLUP), and GBC (a method that fits a G-BLUP and a BayesC term in the model). With DYD as response variable, forward prediction accuracies ranged from 0.427 to 0.664 across the traits and evaluation methods. With absorbed records as response variable, accuracies ranged from 0.429 to 0.667 across the traits and evaluation methods. Comparison

of performance between DYD and the Absorbed dataset showed that accuracies were on average slightly higher for the Absorbed dataset for A-BLUP, whilst the opposite was found for G-BLUP, GBC, and SS-GBLUP, although differences were not statistically significant. Predictions based on Absorbed dataset were generally less biased. Although we were able to implement a variable selection method for genomic prediction on the Absorbed dataset, our results did not show an extra advantage of using the variable selection method GBC. Overall, the utilization of absorbed data and thus all available information led to less biased predictions. The use of Absorbed data in combination with alternative genomic prediction methods was discussed, and it was concluded that it might be especially useful in combination with marker based single step genomic predictions.

Key words: genotyped animal, ungenotyped animal, genomic prediction, absorption of phenotype

INTRODUCTION

Presently, most genomic prediction methods focus mainly on genotyped animals, involve multiple-step procedures in estimation of genomic breeding values (GEBV), and often use daughter yield deviations (DYD) as response variable (Legarra et al., 2009; Christensen and Lund, 2010; Meuwissen et al., 2011). However, many animals in livestock populations are ungenotyped. And for these animals, phenotypic information or estimated breeding values might be available. The availability of animals with phenotypes but no genotypes offers an opportunity to enlarge the reference dataset for genomic prediction. In addition, utilization of all available information could lead to more accurate and less biased predictions (Legarra et al., 2014) . The single-step method combines pedigree and genomic information via a special relationship matrix called **H**-matrix (i.e. numerator relationship matrix (**A**-matrix) plus genomic relationship matrix (**G**-matrix)) (Legarra et al., 2009; Misztal et al., 2009; Christensen

and Lund, 2010). The above approach referred to as ssGBLUP accommodates ungenotyped animals thereby ensuring that all available phenotypic information are utilized in genomic prediction. Because ssGBLUP involves the blending of **G** and **A**, it is important that they are both scaled to the same base population in-order to avoid biased predictions (Forni et al., 2011; Meuwissen et al., 2011). To this end, Meuwissen et al. (2011) proposed an approach in which the missing genotypes are imputed by linkage analysis instead of by **A**-matrix based regression coefficients. The modified single-step of Meuwissen et al. (2011) resulted in a higher accuracy and virtually unbiased GEBV in a simulation study and in a national Norwegian Red dataset (Meuwissen et al., 2015) as compared to the original single-step method. Genomic-BLUP based single-step methods have received much attention both in theory and its evaluation in real data (e.g. Aguilar et al. (2010); Forni et al. (2011); Koivula et al. (2012); Li et al. (2014); Lourenco et al. (2015)). Variable selection methods, such as BayesB, have been developed in theory but hardly used in practice (Legarra and Ducrocq (2012); Fernando et al. (2014); Liu et al. (2014)). These and other approaches have been reviewed by Legarra et al. (2014), however, the resulting equations showed convergence problems.

Here we propose and evaluate an approach to absorb phenotypic information on a large number of ungenotyped animals into mixed model equations of genotyped animals. This approach results in a pseudo response variable, which reflect the information from non-genotyped animals more completely than DYD's. These pseudo-records can then be utilized for genomic prediction with variable or non-variable selection methods. Genomic predictions were done using either response variables, DYDs or the resulting pseudo-records from absorption.

MATERIALS AND METHODS

SNP Data. An imputed SNP data on 3,244 progeny-tested Norwegian Red bulls were kindly provided by GENO SA (www.geno.no) for the analyses. A total of 2,450 bulls were genotyped with the 25K Affymetrix chip (Affymetrix Inc., Santa Clara, CA), 1,650 were genotyped with the Illumina Bovine50K BeadChip (Illumina Inc., San Diego, CA), while 856 of the bulls were genotyped with both chips. Bulls genotyped with the 25K Affymetrix chip were imputed to 50K (i.e. 25K/50K). Genotyping and imputation was performed by CIGENE (www.cigene.no). Imputation was by the Beagle package (Browning and Browning, 2009) and other in-house developed software as described by Solberg et al. (2011) . Quality control involved: removal of animals with an individual call rate <97%; deletion of SNP with a call rate <25%; pedigree relationships between parent and offspring were set to missing if they exceeded the Mendelian error threshold of >1%; SNPs with an overall Mendelian error rate >2.5% were deleted; then, for parent-offspring relationships with Mendelian errors <1%, genotypes for that loci were set to missing; finally, SNPs with minor allele frequency <0.05 were discarded. Following this, the 50K dataset contained 48,249 SNP for a total of 3,244 Norwegian Red bulls. SNPs on the X chromosome were not included.

Genotyped Animals and Their Ancestors (GA-set). The GA-set is an enlarged dataset comprising of both animals genotyped with 50K and their ancestors. In total 20,918 animals. The pedigree comprised of 20,918 animals, of which 5 generations of ancestors preceded the genotyped animals and 19 generations in total. Included in the pedigree are 4,401 founders, and 16,517 non-founders. While there were 18,357 animals with progeny, 2,561 animals had no progeny in the GA-set. Genotype probabilities for the ungenotyped animals in the GA-set were generated with the LDMIP program (Meuwissen and Goddard, 2010), without making use of the information on linkage between the markers. In this case the LDMIP algorithm

reduces to a segregation analysis which was found to be more accurate in datasets with relatively few genotyped animals (Meuwissen et al., 2015).

Absorption of Phenotypic Information of Ungenotyped Animals into GA-set Animals. We define a D-set of animals (a total of 4,022,179 animals), which are all none-GA-set animals, and are thus ungenotyped animals that have also no genotyped descendants. Generally, D-set animals are descendants of the GA-set. The information (records) on the D-set animals will be used by absorbing their mixed model equations (MME) into those of the GA-set animals.

We first assume only pedigree relationships ($\boldsymbol{A}$ matrix) for the absorption of the D-set. After absorption of D-set equations and fixed and other random effects equations, the mixed model equations (MME) for the GA-set animals may be written as:

$$(\boldsymbol{M} + \boldsymbol{A}^{-1}\lambda)\boldsymbol{EBV}_{GA} = \boldsymbol{d} \qquad [1a]$$

where $\boldsymbol{EBV}_{GA}$ are the EBV of the GA-set animals; λ is the variance ratio σ_e^2/σ_a^2; $\boldsymbol{M}$ is the information matrix resulting from the absorption; and $\boldsymbol{d}$ is the right-hand-side resulting from the absorption process. We will not write out the exact form of $\boldsymbol{M}$ and $\boldsymbol{d}$ here, since they depend on the fixed and random effects in the model. We will however assume that the information matrix, $\boldsymbol{M}$, can be approximated by a diagonal matrix, such that the same $\boldsymbol{EBV}_{GA}$ and reliabilities result as for the complete data set. The $\boldsymbol{EBV}_{GA}$ and their reliabilities were assumed known (e.g. from a national pedigree based breeding value evaluation). Hence, we will approximate $\boldsymbol{M}$ by a diagonal matrix $\boldsymbol{W}$. In addition, we will write the right hand side $\boldsymbol{d}$ as $\boldsymbol{d} = \boldsymbol{W}\boldsymbol{y}_a$, where $\boldsymbol{y}_a$ are absorbed records resulting in the desired $\boldsymbol{EBV}_{GA}$. Thus the MME for the absorbed records $\boldsymbol{y}_a$ with weights $diag(\boldsymbol{W})$ are:

$$(\boldsymbol{W} + \boldsymbol{A}^{-1}\lambda)\boldsymbol{EBV}_{GA} = \boldsymbol{W}\boldsymbol{y}_a \qquad [1b]$$

The weights $\boldsymbol{W}$, that give the same reliabilities as for the complete data ($GA + D$ set), from (national) evaluations, are derived in the Appendix. The right-hand-side $\boldsymbol{d}$ is obtained by multiplying $(\boldsymbol{W} + \boldsymbol{A}^{-1}\lambda)$ times the known $\boldsymbol{EBV}_{GA}$ (from large-scale evaluations). Finally, the absorbed records $\boldsymbol{y}_a$ are obtained as $\boldsymbol{W}^{-1}\boldsymbol{d} = \boldsymbol{y}_a$, which requires that all weights are positive.

Second, we consider genomic relationships for the GA-set animals, and note that the absorption of D-set animals is not affected by the known marker genotypes since the D-set animals have no marker information, nor have their descendants (Meuwissen and Goddard, 1999), as is also shown in the Appendix. Accounting for the known marker information results thus in the absorbed G-BLUP equations: $(\boldsymbol{W} + \boldsymbol{G}^{-1}\lambda)\boldsymbol{GEBV}_{GA} = \boldsymbol{W}\boldsymbol{y}_a$, where $\boldsymbol{W}$ and $\boldsymbol{y}_a$ are the same as for the $\boldsymbol{A}$ matrix based equations. Due to the equivalence of G-BLUP and SNP-BLUP, the equations (1b) may also be solved by SNP-BLUP or by a Bayesian model for the estimation of SNP effects (e.g. BayesB/C/R). It may be noted that by the absorption of the D-set animals (often many millions of animals) all information included in their phenotypes is utilized.

Reference and Validation Dataset. The reference set comprised of 3,091 animals in the case of only progeny-tested bulls and 20,765 animals in the case of the absorbed dataset. Reference animals were born between 1965 and 2005. The validation scheme was based on forward predictions following a standard animal breeding selection scheme. Hence, the validation dataset consisted of the youngest sires born between 2007 and 2008. The 153 youngest sires having more than 100 daughters with lactation records were used for validation.

Statistical Analyses

For the sake of convenience, a combination of DYD and genotype of the 3,244 progeny-tested bulls is referred to as "DYD dataset"; while a combination of absorbed records and genotype probabilities of the GA-set animals is referred to as "Absorbed dataset". Four

evaluation methods (G-BLUP, SS-GBLUP, GBC, and A-BLUP) in-total were implemented for prediction of breeding values of animals in the validation set. In the case of the DYD dataset, G-BLUP, GBC, and A-BLUP, were implemented using DYD's as the response variable and the variance components estimated from the dataset. In the case of the Absorbed dataset, all four methods were implemented using absorbed records as the response variable and the variance components provided from routine genetic evaluations of Norwegian Red Cattle population. The traits analyzed were milk yield (KgM), fat yield (KgF), protein yield (KgP), and somatic cell count (SCC). The heritabilities (0.136 - 0.277) of the traits and number of records are shown in Table 1, and are as used in NRF national evaluations.

G-BLUP** and **SS-BLUP. The G-BLUP and SS-GBLUP model used to predict GEBV was as follows:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z}g + \boldsymbol{e}, \qquad [2]$$

where **y** is a vector of (pseudo) records; **1** is a vector of ones; μ is the overall mean; $\boldsymbol{Z}$ is a design matrix that maps the records to genomic values; $\boldsymbol{g}$ is a vector of genomic values assumed to follow a multivariate normal distribution $MVN(0,\boldsymbol{G}\sigma_g^2)$, where $\boldsymbol{G}$ is the genomic relationship matrix and σ_g^2 is the genetic variance; and $\boldsymbol{e}$ is the vector of residuals, assumed to follow a multivariate normal distribution $MVN(0,\mathbf{W}^{-1}\sigma_e^2)$, where $\boldsymbol{W}$ equaled the identity matrix for the DYD's or contained the weights of the absorbed records. $\boldsymbol{G}$ was calculated, following Method 1 of VanRaden (2008), as $\boldsymbol{G} = \boldsymbol{XX}'/2\sum p_j\,(1-p_j)$**,** and $X_{ij} = x_{ij} - 2p_j$, where x_{ij} is the genotype of animal i for SNP j, with x_{ij} = 0**,** 1 or 2 for reference homozygote, heterozygote and alternative homozygote respectively, and p_j is the allele frequency of SNP j. The $\boldsymbol{G}$ in the case of the Absorbed dataset is a relationship matrix at the gametic level, referred here as $\boldsymbol{G_{LDLA}}$ matrix because it is based on a combination of relationship information from linkage analysis (LA) and linkage disequilibrium (LD) (Meuwissen et al., 2015). $\boldsymbol{G_{LDLA}}$ was

constructed as detailed in Meuwissen et al. (2015) using genotype probabilities since some of the GA-set animals were ungenotyped. Following Meuwissen et al. (2015): $\boldsymbol{G_{LDLA}} = \boldsymbol{S}(\boldsymbol{D}\widetilde{\boldsymbol{G}}\boldsymbol{D} + \Delta\widetilde{\boldsymbol{A}}\Delta)\boldsymbol{S}'/2$, where $\boldsymbol{D}$ is a diagonal matrix with elements $1/\sqrt{(\tilde{G}_{ii})}$ when $\tilde{G}_{ii} > 1$ or 1 elsewhere; Δ is a diagonal matrix with elements $\sqrt{(1-\tilde{G}_{ii})}$ when $\tilde{G}_{ii} < 1$ or 0 elsewhere; $\widetilde{\boldsymbol{A}}$ is the pedigree based gametic relationship matrix; $\boldsymbol{S}$ is a design matrix indicating which gametes belong to which animals that reduces the gametic relationship matrix $\boldsymbol{DGD} + \Delta\widetilde{\boldsymbol{A}}\Delta$ to an animal relationship matrix of size number of animals squared; $\widetilde{\boldsymbol{G}}$ is a matrix of gametic relationship: $\tilde{G} = \boldsymbol{X_g}\boldsymbol{X_g}'/\sum_j p_j(1-p_j)$, where $\boldsymbol{X_g}$ is a matrix of standardized genotypes, i.e. $X_{g_{ij}}$ is the probability of a '1' allele of gamete i at marker j expressed as a deviation from its mean, which is the frequency (p_j) of the '1' allele ($X_{g_{ij}}$ = [prob("1"allele) - p_j]).

Implementation of SS-GBLUP was as described in Legarra et al. (2009); Misztal et al. (2009); Christensen and Lund (2010). Genetic values were assumed to follow a normal distribution, $N(0,\boldsymbol{H}\sigma_g^2)$ where $\boldsymbol{H}$ is the relationship matrix combining both SNP marker and pedigree information: $\boldsymbol{H} = \begin{bmatrix} \boldsymbol{G}_\omega & \boldsymbol{G}_\omega\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} \\ \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{G}_\omega & \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{G}_\omega\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} + \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} \end{bmatrix}$, where $\boldsymbol{A}$ is pedigree relationship matrix portioned as $\boldsymbol{A} = \begin{bmatrix} A_{11} & \boldsymbol{A}_{12} \\ A_{21} & \boldsymbol{A}_{22} \end{bmatrix}$ with subscript 1 for genotyped animals and 2 for ungenotyped animals and $G_\omega = (1-\omega)G + \omega A_{11}$, where ω is the relative weight on $\boldsymbol{G}$ and $\boldsymbol{A}$ when forming the combined relationship matrix. $\boldsymbol{G}$ was calculated, following Method 1 of VanRaden (2008) as described above. The inverse of $\boldsymbol{H}$ was $\boldsymbol{H}^{-1} = \begin{bmatrix} \boldsymbol{G}_\omega^{-1} - \boldsymbol{A}_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \boldsymbol{A}^{-1}$. Unlike Gao et al. (2012) and Koivula et al. (2012), ω was set to zero in this study since a preliminary study on a range of ω values (0 to 0.30) revealed insignificant effect on the accuracy of genomic prediction in our data. In addition, ω was set to zero in-order to make the $\boldsymbol{H}$ matrix comparable to other relationship matrix in this study that also considered

that the markers could explain all genetic variance. SS-GBLUP was implemented in DMU (Madsen and Jensen, 2013) using the G-ADJUST option which adjust elements in the genomic relationship so that average of diagonal elements and average of off-diagonal elements equal the same averages in the additive relationship for the genotyped animals.

GBC. The GBC fits simultaneously a polygenic effect (as in G-BLUP) and a BayesC effect (Habier et al., 2011) using the ICE algorithm (Meuwissen et al., 2009), which includes a correction for the uncertainty of the other SNP effects when estimating the effect of SNP '*i*' as described by Wang et al. (2015). The polygenic term is expected to catch the part of the breeding value that is explained by relationships between the animals and co-segregation of alleles within the families, whilst the BayesC term is expected to pick-up the LD between SNPs and major genes. The model of analysis used by GBC, based on fitting G-BLUP and SNPs with large effects is thus:

$$\boldsymbol{y} = 1\mu + \boldsymbol{Zg} + \boldsymbol{ZXQq} + \mathbf{e}\,, \qquad [3]$$

where **y** is a vector of records; **1** is a vector of ones; μ is the overall mean; $\boldsymbol{Z}$ is a design matrix that maps the records to genomic values; $\boldsymbol{g}$ is vector of polygenic effects (i.e. G-BLUP aspect of the model) with distributional assumptions as described above for G-BLUP, $\boldsymbol{X}$ is a design matrix of standardized SNP genotypes as used for the calculation of G above, $\boldsymbol{q}$ is the vector of SNP effects (q_i), and $\boldsymbol{e}$ is the vector of residuals; **Q** is a diagonal matrix with 1 on the diagonal if the SNP has a large BayesC effect (with prior probability π) and 0 if it has no such effect (with prior probability (1-π)). In the ICE algorithm the elements of Q are obtained using the posterior probabilities (i.e. prior probability times likelihood) that the SNP has large effect. The starting prior (π) probability of a SNP having large effects was 1%. The G-BLUP term was implemented as described in the G-BLUP section. The BayesC term was implemented as described in Habier et al., 2007, except that here the 'small' SNPs had a non-zero genetic

variance, whereas the 'large' genes were assumed to have a variance of $0.01\sigma_g^2$. Because fitting many 'small' SNP effects is equivalent to fitting a GBLUP term (e.g. Habier et al., 2007), the core difference between the GBC and BayesC model is the variance assumed for the 'small' SNPs (i.e. the GBLUP term) which is in BayesC assumed to have zero effects.

A-BLUP. The A-BLUP ignores genomic information and relies on pedigree information from ancestors using a numerator relationship matrix (**A**). Using the data set, with the genomic information excluded, and the pedigree, A-BLUP predictions of EBV were obtained. The analysis was carried out using ASReml version 3.0 package (Gilmour et al., 2009).

Predictive Ability. Accuracy of prediction (r) was calculated as the correlation between the predicted values i.e. GEBV or EBV (in the case of traditional BLUP) and DYD, divided by the square root of the average reliability of the DYD $\left(\sqrt{r_{DYD}^2}\right)$ of the reference animals . Calculation of reliabilities of DYD was according to Fikse and Banos (2001), i.e. ($r_{DYD}^2 = d_e/(d_e + K)$, where d_e is the effective number of daughters and $K = (4 - h^2)/h^2$). The average effective number of daughters was 173 with a standard deviation of ~34. The bias of predictions from the different evaluation methods was measured as the regression of DYD on their predicted values. A regression coefficient of 1 denotes no bias, <1 implies that extreme high (low) values of the (G)EBV over- (under-) predict the realized phenotypes, and vice versa a regression coefficient of >1. Standard errors of the prediction accuracies and the regression coefficients on the DYD were computed using bootstrapping with the R-package (R Core Team, 2015). The bootstrapping procedure involved sampling with replacement of the (G)EBVs 10,000 times. For each bootstrap sample, the (G)EBVs were correlated to the DYD. Standard errors were computed from the 10,000 bootstraps of estimated accuracies and regression coefficients.

RESULTS AND DISCUSSION

Accuracy of Prediction using DYD as Response Variable

The accuracies of predictions, as well as regression coefficients of DYD on predicted values for the different evaluation methods, using DYD as response variable are presented in Table 2. The accuracies ranged from 0.427 to 0.664 across the traits and evaluation methods, with A-BLUP (0.427 – 0.491), which is pedigree based yielding less accurate predictions in comparison to genomic approaches. While GBC (0.580 – 0.664) yielded more accurate predictions than G-BLUP (0.575 – 0.652). In all of the scenarios, highest accuracy was observed for KgF, while the lowest was observed for SCC. Generally, higher accuracies were obtained for production traits (i.e. KgF, KgM, KgP) as compared to the health indicator trait, SCC. This is not surprising because SCC (like other health traits) has a lower heritability, which impacts the accuracy of prediction. Several studies, e.g. Moser et al. (2009) and Luan et al. (2009), also observed higher accuracy of genomic prediction for highly heritable traits than for lowly heritable traits. According to Goddard and Hayes (2009) more phenotypic information would prove very useful in order to increase prediction accuracy for SCC and other lowly heritable traits.

G-BLUP and GBC yielded similar accuracies. GBC, a method that fits a polygenic term (i.e. G-BLUP) next to BayesC SNP effects, had a small advantage over G-BLUP across all four traits studied. On average, the increase in accuracy using GBC was 1%. A possible explanation for the small advantage of GBC over G-BLUP could be that the modelled LD blocks surrounding the major genes for the traits were large and were also reasonably well captured by G-BLUP (Daetwyler et al., 2010). Another possible explanation is that the few SNPs (Figure 1) with moderate to large effects explain a relatively small amount of the genetic variance relative to that explained by all the SNPs with small effects. In this situation, GBC's result in a performance that is only slightly better than G-BLUP, since it gives extra weight to the SNPs

with large effects. The observed results are in accordance with what has been reported in other studies. For instance, Su et al. (2012) reported 0.5% increase in reliability when using a Bayesian mixture model with 2 normal distributions on Holstein cattle data, but no increase was observed on Nordic Red Dairy cattle.

The regression coefficient of EBV on masked DYD's measures the degree of bias of prediction (Table 2). For an unbiased prediction, this regression coefficient should be 1. As shown in Table 2, the regression coefficients followed a similar trend across the evaluation methods. Although the regression coefficients deviated substantially from 1, only few of these deviations were borderline significant (deviation > 2 standard errors). The latter occurred for KgM, but here the bias was biggest for A-BLUP, which is known to be an unbiased EBV estimator. Since selection is mostly for KgM, the biases are thus most likely due to selections that are not accounted for in the DYD data.

Accuracy of Prediction using Absorbed Records as Response Variable

The accuracies and bias of predictions for the different evaluation methods using the absorbed records as response variable are presented in Table 3. Accuracy of prediction ranged between 0.429 to 0.667 across the traits and evaluation methods, with A-BLUP (0.429 – 0.511) giving lower accuracies as compared to G-BLUP, SS-GBLUP, and GBC. G-BLUP (0.565 – 0.667) using G_{LDLA} yielded slightly more accurate predictions than SS-GBLUP (0.558 – 0.656) and GBC (0.561 – 0.665) across the four traits. However, the accuracies of prediction obtained using GBC were much closer to those of G-BLUP than SS-GBLUP to G-BLUP. Except for KgM, GBC resulted in slightly better predictions than SS-GBLUP. As previously observed, the highest accuracy was obtained for KgF, while the lowest was for SCC. In addition, higher accuracies were obtained for production traits as compared to health indicator trait, SCC.

The regression coefficients for the production traits ranged from 0.982 to 1.076 when using A-BLUP (Table 3), which suggests that A-BLUP is unbiased as is also known from theory.

Thus, the approximate absorption of the D-set animals results in a reduced data set, that seems to account for the selections that took place within the D-set animals. The genomic models also yielded unbiased EBV, although their regression coefficients tended to be smaller than 1, especially for KgP. For SSC, the A-BLUP regression coefficient of 0.828 deviates substantially from 1, but the difference is not statistically significant. This substantial deviation is probably because SSC is strongly correlated to mastitis, and there is strong direct selection for mastitis resistance in NRF cattle, which is not picked-up by the current single trait model for SSC. Similarly the genomic prediction models yielded small regression coefficients for SSC varying between 0.795 and 0.835.

Previous studies (e.g. Gao et al. (2012); Koivula et al. (2012); Li et al. (2014)) based on de-regressed proofs as the response variable and the $\boldsymbol{H}$ matrix as the relationship matrix have reported that SS-GBLUP performed slightly better than G-BLUP. However, a recent study by Meuwissen et al. (2015) showed that G-BLUP with an alternative relationship matrix (G_{LDLA}; based on linkage analysis and linkage disequilibrium information) performed slightly better than SS-GBLUP. In the present study, G-BLUP using the absorbed record and $\boldsymbol{G}_{LDLA}$ yielded on average 0.7% higher accuracy than SS-GBLUP which used $\boldsymbol{H}$ matrix. Suggesting that in a case of large dataset with relatively few genotyped animals (3,244 out of 20,918 in this study), utilization of $\boldsymbol{G}_{LALD}$ matrix could be at least equally accurate than the $\boldsymbol{H}$ matrix. The large number of ungenotyped animals in the Absorbed dataset might have affected the performance of SS-GBLUP.

The key objective of this study was to implement a variable selection method of genomic prediction given a dataset composed of genotyped and ungenotyped animals. This objective was achieved, however, the accuracies of predictions were very similar to that of G-BLUP (Table 3). This was not expected since GBC had a small advantage over G-BLUP when using DYD's as response variable and actual genotypes. A possible reason for the reduced accuracy

of GBC versus GBLUP is that with the Absorbed dataset, GBC was not able to capture SNPs with moderate or large effects, due to the large scale use of genotype probabilities instead of actual genotypes. A check on the estimated SNP effects showed that with the DYD dataset, there were clear indications of SNPs with moderate or large effects for most of the traits studied, while with the Absorbed dataset, no clear indications were seen (Figure 1 and 2). The latter suggests that the LD between SNP genotype probabilities and the QTL is substantially lower than the LD between actual SNP genotypes and the QTL.

Comparison of DYD and Absorbed Dataset

A comparison of Table 2 and 3 shows that differences in accuracy between the DYD and the Absorbed dataset, although not statistically significant, were on average slightly higher for the Absorbed dataset for A-BLUP, whilst the opposite was found for G-BLUP, GBC, and SS-GBLUP (where the latter was compared to GBLUP in Table 2, since in Table 2 there were no missing genotypes). It seems that the large-scale use of genotype probabilities reduced the accuracies of the genomic selection methods, whereas for the analysis of the DYD dataset, actual genotypes were used. Currently, genotyped cows are entering the reference population, which may favor the use of absorbed records instead of DYD's, since genotype cows introduce complications with respect to the weighting of DYD's (information content on cows and bulls are very different), and double counting (a genotyped cow may enter the data directly and as part of a DYD).

The use of Absorbed data was tested here in combination with a limited number of genomic prediction methods. However, it may be applied in combination with any kind of genomic prediction method, that can handle ungenotyped ancestors, and/or in situations where the ungenotyped ancestors may be neglected (e.g. if they are few and/or old). For instance, for the marker based single step method of Fernando et al. (2014), which can fit Gibbs-sampling based variable selection models, the use of Absorbed data may be computationally very

advantageous. In situations where there are K millions of ungenotyped descendants, the Absorbed data would eliminate their equations and thus eliminate the need for a K million times M matrix of genotype probabilities that would enter every cycle of the Gibbs-chain, and thus greatly reduces the computational speed of this method.

CONCLUSIONS

An approach to absorb phenotypic information of large numbers of ungenotyped animals into mixed model equations of genotyped animals was proposed and evaluated. A-BLUP, G-BLUP, SS-BLUP, and GBC (a variable selection method), methods of evaluation was implemented on the resulting dataset. The methods performed as follows A-BLUP<SS-GBLUP<GBC<G-BLUP. We were able to implement a variable selection method for genomic prediction on the Absorbed dataset, however, our results did not show an extra advantage of using a variable selection method. Overall, the utilization of all available information led to less biased predictions.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 289592 - Gene2Farm. The authors thank Geno SA (Ås, Norway) for providing the datasets and CIGENE (Ås, Norway) for the quality control and genotype imputation. Neither the European Commission nor the partners of the project can be held responsible for views expressed in this manuscript.

APPENDIX

Approximation of the Absorption Matrix W

Absorption When Using A Matrix. We will consider here that the absorption of pedigree relationship (***A*** matrix) based mixed model equations (MME), where descendants of genotyped animals (D-set) are absorbed into the equations for genotyped animals and their ancestors (GA-set). The absorbed MME are described by:

$$(\boldsymbol{M} + \boldsymbol{A}^{-1}\lambda)\boldsymbol{EBV}_{GA} = \boldsymbol{d}$$

where $\boldsymbol{EBV}_{GA}$ are the EBV of the GA-set animals; λ is the variance ratio σ_e^2/σ_a^2; $\boldsymbol{M}$ is the information matrix resulting from the absorption; and $\boldsymbol{d}$ is the right-hand-side resulting from the absorption process. Here we will derive an approximation for $\boldsymbol{M}$ called $\boldsymbol{W}$, which is assumed a diagonal matrix. The above equations can thus be seen as MME for GA-set animals with absorbed records $\boldsymbol{y}_a = \boldsymbol{W}^{-1}\boldsymbol{d}$, and weights $diag(\boldsymbol{W})$. We further assume that $\boldsymbol{EBV}_{GA}$ and their reliabilities, r^2, are known, e.g. from a large-scale (national) pedigree-based genetic evaluation.

The algorithm to approximate the weights $diag(\boldsymbol{W}) = \boldsymbol{w}$ of the absorbed records is:

Step 0: Translate reliabilities r^2 into an effective number of records: $n_{ei} = \lambda r_i^2/(1 - r_i^2)$; and set $w_i = n_{ei}$ for all GA animals i.

Step 1: Estimate approximate reliabilities r_{ai}^2 using as weights for own performance records w_i and the algorithm of Tier and Meyer (2004). Using their notation, the Tier and Meyer algorithm splits its estimate of the total effective number of records (F_i) into a term due to parents, $\sum_j E_j^*$ (where summation is over the known parents j); a term due to offspring, $\sum_l E_l$ (where summation is over offspring l within the GA-set); and a term due to own records D_i

(which equals our weight w_i); E_j^* is the information / effective number of records due to parent j; and E_l is the information / effective number of records due to offspring l.

Step 2: The weights of the own records w_i are updated such that the total effective number of records equal n_{ei} which are known from the reliabilities (Step 0). Specifically the weights are updated by:

$$w_i = n_{ei} - (\sum_j E_j^* + \sum_l E_l)$$

If $w_i \leq 0$, set $w_i = 10^{-6}$ (the avoidance of $w_i = 0$ is explained below). If these updated w_i are sufficiently close to the previous weights, the algorithm stops. Otherwise, go to step 1 (i.e. recalculate information from parents and offspring using the new weights).

Given the weights $\boldsymbol{W}$ from the above algorithm, the absorbed right-hand-side $\boldsymbol{d}$ can be calculated as (assuming known $\boldsymbol{EBV_{GA}}$):

$$\boldsymbol{d} = (\boldsymbol{W} + \boldsymbol{A^{-1}}\lambda)\boldsymbol{EBV}_{GA}$$

and absorbed record $\boldsymbol{y}_a$ can be calculated as:

$$\boldsymbol{y_a} = \boldsymbol{W^{-1}d}$$

The latter requires that the weight $diag(\boldsymbol{W})$ are positive, which explains the above mentioned avoidance of $w_i = 0$. The MME of the absorbed records $\boldsymbol{y_a}$, with weights $diag(\boldsymbol{W})$ are thus:

$$(\boldsymbol{W} + \boldsymbol{A}^{-1}\lambda)\boldsymbol{GEBV}_{GA} = \boldsymbol{W}\boldsymbol{y}_a$$

which result in the same EBV and reliabilities as the large scale (national) evaluations. Below we will show that for the current situation, where descendants D-set are absorbed into their ancestors GA-set, the absorption matrix $\boldsymbol{W}$ (or $\boldsymbol{M}$) and the right-hand-side $\boldsymbol{d}$ do not depend on the availability of genomic relationships for the GA-set animals. Thus, in case we have a genomic relationship matrix **G** for the GA set animals, the MME become:

$$(\boldsymbol{W} + \boldsymbol{G}^{-1}\lambda)\boldsymbol{EBV}_{GA} = \boldsymbol{W}\boldsymbol{y}_a$$

where $\boldsymbol{W}$ and $\boldsymbol{d} = \boldsymbol{W}\boldsymbol{y}_{\boldsymbol{a}}$ are the same as for the pedigree based MME.

Absorption When Using G Matrix for GA-set Animals. Following Meuwissen and Goddard (1999), we will show that $\boldsymbol{W}$ and $\boldsymbol{d}$ do not depend on $\boldsymbol{G}$ or $\boldsymbol{A}$ being used as relationship matrix for the GA-set (the D-set animals are assumed ungenotyped). For this, we will consider Henderson (1976) rules to set up the inverse of the relationship matrix of the GA and D set animals. For the genetic value of D-set animal i we write:

$$u_{Di} = {}^{1}/_{2}\, u_s + {}^{1}/_{2}\, u_d + d_{Di}$$

where subscript s (d) denote the sire (dam) of animal i (which may be within the D or GA set); and d_i is the Mendelian sampling deviation with variance $Var(d_{Di}) = {}^{1}/_{2}$ (for simplicity we ignore inbreeding here). The genetic values of the GA-set animals are not splitted in a sire and dam component (since they are explained by marker data), and modelled entirely by the deviation component, i.e. by definition $u_{GAi} = d_{GAi}$, and $Var(\boldsymbol{u}_{GA}) = Var(\boldsymbol{d}_{GA}) = \boldsymbol{G}$. In matrix form, the breeding values of the GA and D set animals can be modelled as:

$$\begin{bmatrix} \boldsymbol{I}\boldsymbol{u}_{GA} \\ \boldsymbol{I}\boldsymbol{u}_{D} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{R} & \boldsymbol{S} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_{GA} \\ \boldsymbol{u}_{D} \end{bmatrix} + \begin{bmatrix} \boldsymbol{d}_{GA} \\ \boldsymbol{d}_{D} \end{bmatrix}$$

where **R** is a matrix with element (i, j) equal ${}^{1}/_{2}$ when j is a (GA-set) parent of i and 0 otherwise; and **S** is a matrix with element (i, j) equal ${}^{1}/_{2}$ when j is a (D-set) parent of i and 0 otherwise. These equations may be rewritten as:

$$\boldsymbol{T}\boldsymbol{u} = \boldsymbol{d}$$

where $\boldsymbol{T} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{R} & \boldsymbol{I}-\boldsymbol{S} \end{bmatrix}$ and $\boldsymbol{d} = \begin{bmatrix} \boldsymbol{d}_{GA} \\ \boldsymbol{d}_{D} \end{bmatrix}$. Thus, $\boldsymbol{u} = \boldsymbol{T^{-1}d}$, and $Var(\boldsymbol{u}) = \boldsymbol{G}_{all} = \boldsymbol{T}^{-1}Var(\boldsymbol{d})(\boldsymbol{T}^{-1})'$, where

$$Var(\boldsymbol{d}) = \boldsymbol{D} = \begin{bmatrix} \boldsymbol{G} & 0 \\ 0 & \frac{1}{2}\boldsymbol{I} \end{bmatrix}$$

We need the inverse of $\boldsymbol{G_{all}}$, i.e. $\boldsymbol{G}_{all}^{-1} = \boldsymbol{T}'\boldsymbol{D}^{-1}\boldsymbol{T}$, where:

$$\boldsymbol{D^{-1}} = \begin{bmatrix} \boldsymbol{G^{-1}} & 0 \\ 0 & 2\boldsymbol{I} \end{bmatrix}$$

And writing out $\boldsymbol{T}'\boldsymbol{D}^{-1}\boldsymbol{T}$ gives:

$$\boldsymbol{G}_{all}^{-1} = \begin{bmatrix} \boldsymbol{G}^{-1} + 2\boldsymbol{R}'\boldsymbol{R} & -2\boldsymbol{R}'(\boldsymbol{I}-\boldsymbol{S}) \\ -2(\boldsymbol{I}-\boldsymbol{S}')\boldsymbol{R} & 2(\boldsymbol{I}-\boldsymbol{S}')(\boldsymbol{I}-\boldsymbol{S}) \end{bmatrix}$$

Now, absorption of the D-set equations into the GA-set involves the terms $-2\boldsymbol{R}'(\boldsymbol{I}-\boldsymbol{S})$,$-2(\boldsymbol{I}-\boldsymbol{S}')\boldsymbol{R}$, and $2(\boldsymbol{I}-\boldsymbol{S}')(\boldsymbol{I}-\boldsymbol{S})$, which all stem from Henderson's rules and do not involve any genomic relationships $\boldsymbol{G}$. Thus, the absorption process is identical (and based on pedigree relationships) whether the GA-set relationships are based on markers or on pedigree.

REFERENCES

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J Dairy Sci 93(2):743-752.

Browning, B. L. and S. R. Browning. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84:210-223.

Christensen, O. F. and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet Sel Evol 42:2.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021-1031.

Fernando, R. L., J. C. Dekkers, and D. J. Garrick. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet Sel Evol 46:50.

Fikse, W. F. and G. Banos. 2001. Weighting factors of sire daughter information in international genetic evaluations. J Dairy Sci 84:1759–1767.

Forni, S., I. Aguilar, and I. Misztal. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet Sel Evol 43:1.

Gao, H., O. F. Christensen, P. Madsen, U. S. Nielsen, Y. Zhang, M. S. Lund, and G. Su. 2012. Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. Genet Sel Evol 44:8.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson. 2009. ASReml User Guide Release 3.0., VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Goddard, M. E. and B. J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet 10:381-391.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

Henderson, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69-83.

Koivula, M., I. Stranden, G. Su, and E. A. Mantysaari. 2012. Different methods to calculate genomic predictions--comparisons of BLUP at the single nucleotide polymorphism level (SNP-BLUP), BLUP at the individual level (G-BLUP), and the one-step approach (H-BLUP). J Dairy Sci 95(7):4065-4073.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J Dairy Sci 92(9):4656-4663.

Legarra, A., O. F. Christensen, I. Aguilar, and I. Misztal. 2014. Single Step, a general approach for genomic selection. Livest Sci 166:54-65.

Legarra, A. and V. Ducrocq. 2012. Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction. J Dairy Sci 95:4629-4645.

Li, X., S. Wang, J. Huang, L. Li, Q. Zhang, and X. Ding. 2014. Improving the accuracy of genomic prediction in Chinese Holstein cattle by using one-step blending. Genet Sel Evol 46:66.

Liu, Z., M. E. Goddard, F. Reinhardt, and R. Reents. 2014. A single-step genomic model with direct estimation of marker effects. J Dairy Sci 97:5833-5850.

Lourenco, D. A. L., S. Tsuruta, B. O. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J. K. Bertrand, T. S. Amen, L. Wang, D. W. Moser, and I. Misztal. 2015. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. J Anim Sci 93:2653–2662.

Luan, T., J. A. Woolliams, S. Lien, M. Kent, M. Svendsen, and T. H. Meuwissen. 2009. The accuracy of Genomic Selection in Norwegian red cattle assessed by cross-validation. Genetics 183(3):1119-1126.

Madsen, P. and J. Jensen. 2013. A User's Guide to DMU. A Package for Analysing Multivariate Mixed Models. Version 6, Release 5.2 ed, Faculty Agric. Sci. (DJF), Dept. Genet. Biotechnol., Univ. Aarhus, Res. Center Foulum, Tjele, Denmark.

Meuwissen, T. H. E. and M. E. Goddard. 1999. Marker assisted estimation of breeding values when marker information is missing on many animals. Genet Sel Evol 31:375-394.

Meuwissen, T. H. E. and M. E. Goddard. 2010. The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole-genome sequence density genotypic data. Genetics 185:1441-1449.

Meuwissen, T. H. E., T. Luan, and J. A. Woolliams. 2011. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. J Anim Breed Genet 128(6):429-439.

Meuwissen, T. H. E., T. R. Solberg, R. Shepherd, and J. A. Woolliams. 2009. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genet Sel Evol 41:2.

Meuwissen, T. H. E., M. Svendsen, T. R. Solberg, and J. Odegard. 2015. Genomic predictions based on animal models using genotype imputation on a national scale in Norwegian Red cattle. Genet Sel Evol 47:79.

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J Dairy Sci 92(9):4648-4655.

Moser, G., B. Tier, R. E. Crump, M. S. Khatkar, and H. W. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet Sel Evol 41:56.

R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Solberg, T. R., B. Heringstad, M. Svendsen, H. Grove, and T. H. E. Meuwissen. 2011. Genomic predictions for production and functional traits in Norwegian Red from BLUP analyses of imputed 54K and 777K SNP data. Interbull Bull 44:240-243.

Su, G., R. F. Brondum, P. Ma, B. Guldbrandtsen, G. P. Aamand, and M. S. Lund. 2012. Comparison of genomic predictions using medium-density (approximately 54,000) and high-density (approximately 777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. J Dairy Sci 95:4657-4665.

Tier, B. and K. Meyer. 2004. Approximating prediction error covariances among additive genetic effects within animals in multiple-trait and random regression models. J Anim Breed Genet 121:77-89.

Wang, T., Y. P. Chen, M. E. Goddard, T. H. Meuwissen, K. E. Kemper, and B. J. Hayes. 2015. A computationally efficient algorithm for genomic prediction using a Bayesian model. Genet Sel Evol 47:34.

TABLES

Table 7 Heritability (h^2), reliability (r^2_{DYD})[1] of DYD[2], number of animals with phenotype and genotype per dataset[3]

Trait	h^2	r^2_{DYD}	Number of animals	
			DYD dataset	Absorbed dataset
Somatic cell count (SCC)	0.136	0.855	3,244	20,918
Fat yield (KgF)	0.213	0.904	3,244	20,918
Milk yield (KgM)	0.277	0.925	3,244	20,918
Protein yield (KgP)	0.235	0.912	3,244	20,918

[1] $r^2_{DYD} = d_e/(d_e + K)$, where d_e is the effective number of daughters and $K = (4 - h^2)/h^2$

[2] DYD = daughter yield deviations

Table 2 Accuracy[1] (±SE[2]) and bias (±SE) of the predicted values for the youngest sires based DYD dataset and different prediction methods[3]

Trait[4]	A-BLUP	G-BLUP	GBC
Accuracy			
SCC	0.444 (±0.076)	0.575 (±0.065)	0.580 (±0.065)
KgF	0.491 (±0.070)	0.652 (±0.054)	0.664 (±0.053)
KgM	0.437 (±0.074)	0.616 (±0.059)	0.637 (±0.057)
KgP	0.427 (±0.078)	0.601 (±0.058)	0.604 (±0.057)
Bias			
SCC	0.922 (±0.160)	0.811 (±0.101)	0.812 (±0.100)
KgF	1.161 (±0.172)	1.158 (±0.115)	1.139 (±0.110)
KgM	1.335 (±0.242)	1.317 (±0.149)	1.279 (±0.133)
KgP	1.240 (±0.243)	1.223 (±0.148)	1.187 (±0.139)

[1]Accuracy = $\frac{corr(DYD,(G)EBV)}{\sqrt{r^2_{DYD}}}$

[2]SE = $\sqrt{1 - accuracy^2/n - 2}$, where n is the number of individuals

[3]A-BLUP: traditional BLUP using pedigree-based relationship matrix; G-BLUP: genomic BLUP using genomic-based relationship matrix; GBC: an iterative method that fits a G-BLUP next to SNP effects with a BayesC prior

[4]SCC = somatic cell count; KgF = fat yield; KgM = milk yield; KgP = protien yield.

Table 3 Accuracy[1] (±SE[2]) and bias (±SE) of the predicted values for the youngest sires based Absorbed dataset and different prediction methods[3]

Trait[4]	A-BLUP	G-BLUP	SS-GBLUP	GBC
Accuracy				
SCC	0.429 (±0.079)	0.581 (±0.058)	0.572 (±0.061)	0.579 (±0.058)
KgF	0.511 (±0.067)	0.667 (±0.051)	0.656 (±0.053)	0.665 (±0.051)
KgM	0.450 (±0.069)	0.577 (±0.062)	0.577 (±0.065)	0.574 (±0.064)
KgP	0.463 (±0.072)	0.565 (±0.065)	0.558 (±0.067)	0.561 (±0.066)
Bias				
SCC	0.828 (±0.150)	0.835 (±0.106)	0.795 (±0.105)	0.802 (±0.101)
KgF	0.982 (±0.136)	0.997 (±0.091)	0.961 (±0.097)	0.966 (±0.089)
KgM	1.076 (±0.181)	0.925 (±0.113)	0.937 (±0.122)	0.901 (±0.112)
KgP	1.033 (±0.175)	0.847 (±0.111)	0.842 (±0.118)	0.826 (±0.109)

[1]Accuracy = $\frac{corr(DYD,(G)EBV)}{\sqrt{r^2_{DYD}}}$

[2]SE = $\sqrt{1 - accuracy^2/n - 2}$, where n is the number of individuals

[3]A-BLUP: traditional BLUP using a pedigree-based relationship matrix; G-BLUP: genomic BLUP using genomic-based relationship matrix; SS-GBLUP: single-step genomic BLUP using a combination of pedigree and genomic relationship matrix; GBC: an iterative method that fits a G-BLUP next to SNP effects with a BayesC prior

[4]SCC = somatic cell count; KgF = fat yield; KgM = milk yield; KgP = protien yield

FIGURES

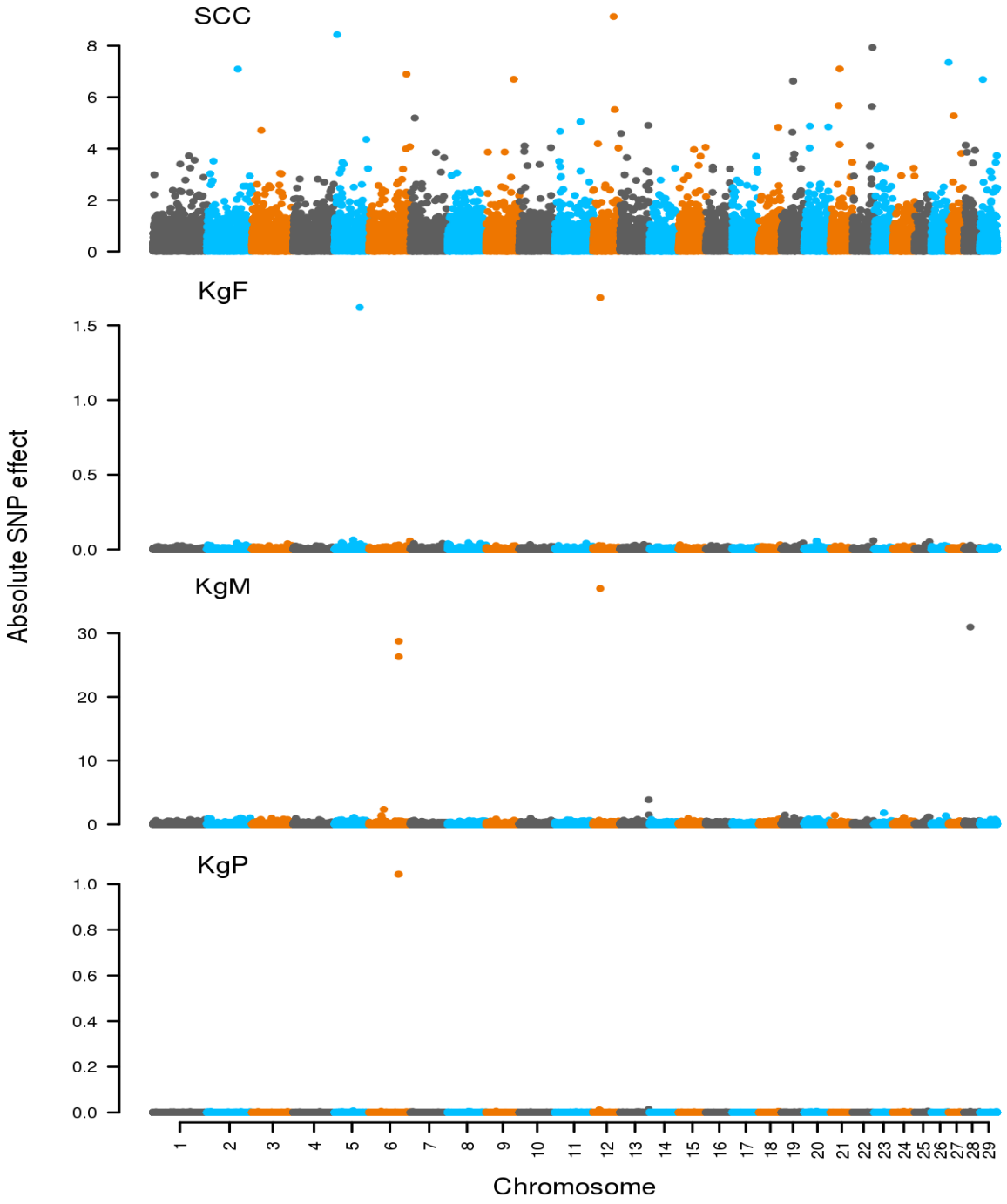

Figure 3 SNP effects based on DYD dataset using GBC method.

Shown is the absolute value of SNP effect estimates (y-axis) from DYD dataset (i.e. a combination of daughter yield deviations and actual genotypes). X-axis is ordered by chromosomes from 1-29. Traits are: SCC = somatic cell count, KgF = fat yield, KgM = milk yield, and KgP = protein yield. Note the changed y-axis scale for each graph. Absolute values of SCC was scaled by 10,000, while KgM was scaled by 1000. Scaling was just for plotting purpose.

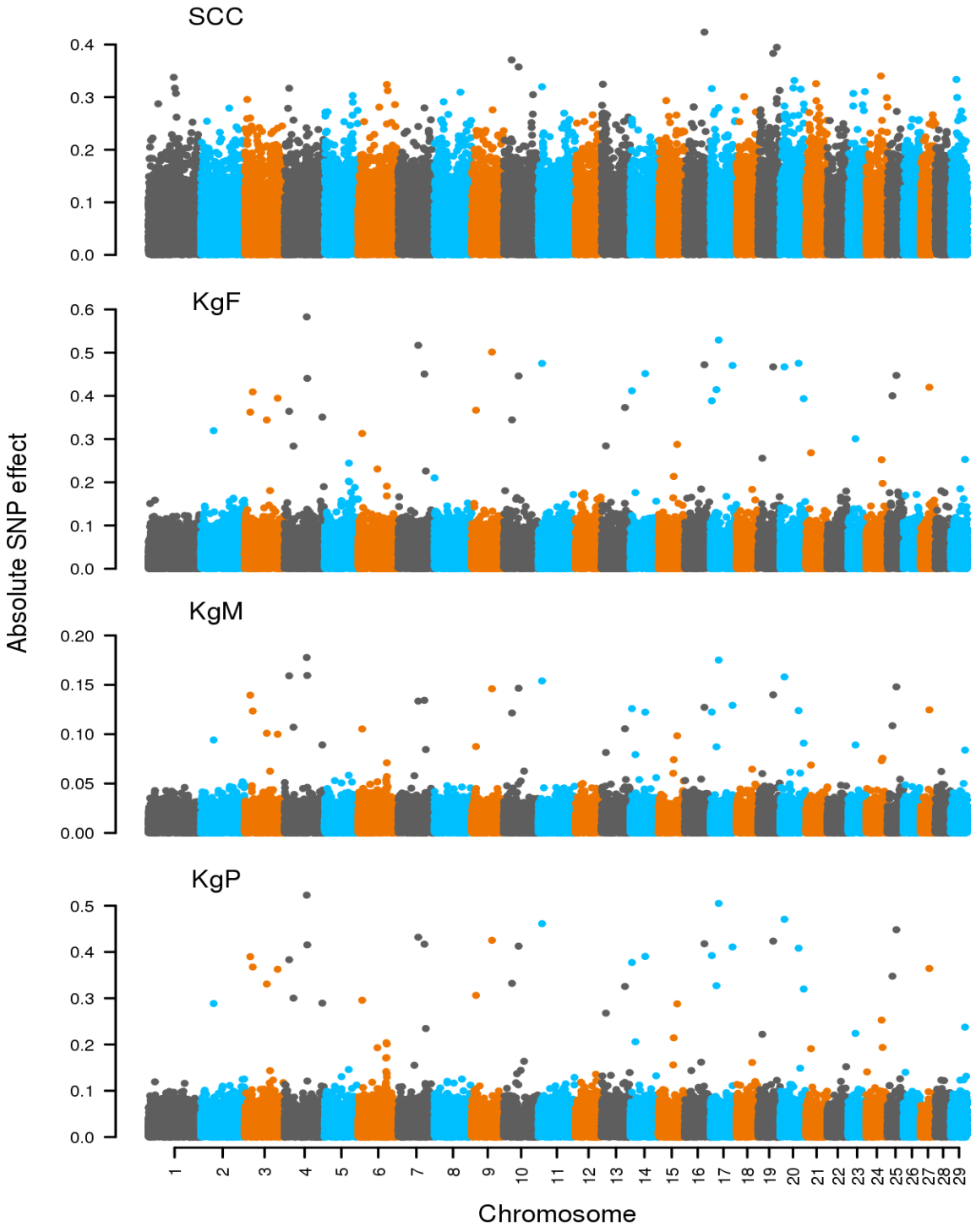

Figure 2 SNP effects based on Absorbed dataset using GBC method.

Shown is the absolute value of SNP effect estimates (y-axis) from Absorbed dataset (i.e. a combination of absorbed record and genotype probabilities). X-axis is ordered by chromosomes from 1-29. Traits are: SCC = somatic cell count, KgF = fat yield, KgM = milk yield, and KgP = protein yield. Note the changed y-axis scale for each graph. Absolute values of SCC was scaled by 10,000, while KgM was scaled by 1000. Scaling was just for plotting purpose.