**WIAS PhD Proposal**

**General information**

Chair Group (s): Animal Breeding and Genomics Centre (ABGC) and Biometris

Project title (English): Towards precision breeding using genomic prediction

Start date – End date:

01-05-2016 to 30-04-2020

**Composition of the project group and scheduled time for project**

Name Role hours per week

Biaty Raymond, MSc. PhD candidate 40

Dr. AC Bouwman Daily supervisor

Prof. Dr. Roel Veerkamp Daily supervisor/promotor

Prof. Dr. Jeanine Houwing-Duistermaat Supervisor/promotor

Prof. Dr. Fred van Eeuwijk Supervisor

**Cooperation with organisations outside WIAS**

Wageningen UR Other Graduate Schools:

Research Institutes:

The Netherlands Universities: LUMC groups Medical Statistics and Bioinformatics

Research Institutes:

Industry and organisations:

International Universities:

Research Institutes:

Industry and organisations:

**Where will the project be carried out:**

Wageningen University and Research, the Netherlands.

Will vertebrate animals be used: NO

Does the project involve biotechnological research: NO

*If one or both answers are ‘yes’, please, take care yourself of appropriate submission to the relevant committee and other legal aspects.*

**Summary**

**Summary of objectives and hypotheses**

The main objective of this project is to develop statistical methodologies that can optimally utilise the sources of variation in sequence data to describe the total genetic variation in animal populations. The efficient use of these variants is expected to increase the accuracy and persistency of genomic prediction (GP) both within and across breeds. The main research question is: How can the accuracy of GP be increased using whole genome sequence (WGS) data? This question is hence divided into five objectives:

1. Is there an added advantage of prior selection of variants based on biological information or statistical significance for GP, as against the use of all available variants in sequence data? Also, what variants weighing scheme is optimum for GP? – All predictions will be carried out using the genomic best linear unbiased prediction (GBLUP) model.
2. Is there an added advantage of prior selection of variants based on biological information or statistical significance for GP, as against the use of all available variants in sequence data? Also, what variants weighing scheme is optimum for GP? – All predictions will be carried out using Bayesian models.
3. How can we improve the power for genome-wide association studies (GWAS) in populations with strong family structure and long range linkage disequilibrium (LD) using WGS data?
4. What is the difference in power/accuracy when single trait GWAS/GP is compared with multi-trait GWAS/GP?
5. What is the best GP strategy for new and expensive traits where small training populations will be common?

**Relevance for the WIAS mission**

The aim of this project is to develop new methods for genomic prediction that efficiently utilise sources of variation in sequence data. These methods are expected to increase the efficiency of breeding programs and contribute to sustainable animal production systems.

**Data management**

The data for this project will be provided by CRV and other partner organizations in the 1,000 bull genomes project. Data will be on traits that are not economically important. Thus, publications using the data will not be commercially sensitive. Data will be handled carefully and in compliance with the data management rules of the chair group.

**Feasibility**

**How is adequate supervision guaranteed?**

For the period of the project, regular meetings between the PhD candidate and supervisors will be held. The supervisors of the project come from a wide area of expertise. Prof. Dr. Roel Veerkamp is an expert in quantitative genetics and animal breeding. Prof. Dr. Fred van Eeuwijk is an expert in statistics, population genetics and plant breeding. Prof. Dr. Jeanine Houwing-Duistermaat is an expert in statistical genetics in the field of human disease studies. Dr. AC Bouwman is an expert in quantitative genetics and animal breeding. Dr. AC Bouwman and Prof. Dr. Roel Veerkamp are the daily supervisors. Bi-weekly, a meeting will be held with the daily supervisors and skype meetings will be held regularly with Prof. Dr. Jeanine Houwing-Duistermaat. Every two months, there will be a meeting with the whole supervising group. Moreover, there is an industry user group meeting every six months. In addition to the normal supervisory group, the user group includes industry representatives from the breed4food partner companies (Cobb, CRV, Hendrix Genetics and TopigsNorsvin). The supervisory team can always be contacted face to face, by email, skype or telephone in case of emergency questions.

How is the execution of the research guaranteed? (facilities, technical assistance)

The groups ABGC and Biometris, both at Wageningen UR and the LUMC groups Medical Statistics and Bioinformatics are the main groups involved in the project. These groups together have the required expertise and knowledge on statistical genetics, theory, software and computer facilities. State of the art computing facilities are available via access to High Performance Cluster (HPC), a versatile platform that can handle jobs having heavy memory demands, raw computing power or enormous storage demands. Cattle data will be made available by CRV and other partners in the 1,000 bull genomes project.

Which agreements have been made regarding cooperation with other groups/universities/institutes?

At Wageningen UR, the PhD student will be physically working at the Animal Breeding and Genomics Centre (ABGC) four days a week and one day a week at Biometris. At the moment, there are no agreements made with other institutes or universities.

**Content**

**Review of literature**

**Genomic prediction**

Genetic improvement in livestock through artificial selection has traditionally been based on performance and pedigree recordings on the selection candidates and their relatives. These records are combined in the so called best linear unbiased prediction (BLUP) [1] model to estimate the breeding values of selection candidates. Estimated breeding values (EBVs) are then used as a criteria to select individuals that will serve as parents for the next generation. Although a rather simple model, it has led to huge gains in most traits of economic importance [2]. Nowadays however, genomic prediction (GP) [3] is widely accepted and implemented by breeding organizations as the preferred method for genetic selection. This has been made possible by the discovery that large scale information at DNA level can be utilised for the prediction of the breeding value of individual animals [3, 4], and the development of appropriate genotyping technology that made such information available.

In GP, individual animals in a group called the reference population are genotyped for DNA markers such as single nucleotide polymorphisms (SNPs) and also phenotyped for all the traits of interest. The combined effect of all loci on each trait are estimated (SNP effects). These SNP effects are used to predict the genomic breeding values (GEBV) of selection candidates which are young individuals that do not have phenotypic records yet [3]. Current methods do not depend on identifying the true functional mutations, but the accuracy of the prediction equation relies on the association between SNPs and the causal variants, caused by existence of linkage disequilibrium (LD) [5].

The use of GP has been highly successful in both plant and animal breeding [6-8], leading to a doubling in the rate of genetic gain [9] compared to the traditional progeny testing scheme. Despite the success however, the accuracy of GP has remained rather constant and does not seem to increase further even with increased sample size and genotype information [10]. This accuracy is even lower for across breeds prediction when breeds have a large genetic distance. Across breed prediction is needed in situations where the reference populations are small (unique traits, numerically smaller breeds). According to the results of Saatchi & Garrick [11], within-breed genomic predictions has no utility for most traits when applied to other breeds (accuracies close to zero), except for some traits due to the segregation of common large-effect QTLs with conserved linkage phase among different breeds. A second reason for improving GP further is to be able to select solely on GEBVs for several generations. Currently it has been observed that there is a decay in the accuracy of GP when the relationship between animals in the reference population and those in the test population is decreasing, even within the same breed [12]. To overcome this, is very important because animals will increasingly be selected solely on genomic information, with no phenotypes of their own. Other factors that influence the accuracy and utility of GP include the size of the reference population [10, 13, 14], the heritability of the trait [14, 15], the extent of LD in both the reference and the test population [16], the number and minor allele frequency of the QTLs underlying the trait [17].

To overcome the current limitations of GP and to increase its accuracy, the use of whole genome sequence (WGS) data instead of only SNP chips has been proposed [18, 19]. This is based on the assumption that all causal mutations are included in sequence data, therefore the accuracy and persistency of GP will no longer have to depend on the LD between SNPs and causal mutations. With this approach, it is expected that GP across breeds will be as accurate as within breeds prediction, and the persistency of GP will be high irrespective of the generation gap and additive genetic relationship between the reference and training population.

The potential to increase the accuracy of GP by the use of WGS has been demonstrated by a number of simulation studies [19-22]. For example, Meuwissen & Goddard [19] reported an increase in the accuracy of GP by more than 40% when WGS data was used instead of a 30k SNP chip. Using WGS data, they also showed that accuracy of GP when both the reference and test population come from the same generation was the same with that in which the reference and test population were separated by 10 generations. This means that the accuracy of GP no longer depends on the existence of LD between SNPs and causal variants when WGS data is used. MacLeod et al. [20] also reported an increase in the accuracy of GP by 22% when WGS was used compared to a 600k SNP chip. This increase in accuracy however was achieved under the condition that the effective population size was large, therefore, average LD was low in the population. When the effective population size was small, there was no added advantage of using WGS over SNP chip.

Initial results using real data have however contradicted the conclusions of these simulation studies [15, 23-25]. The first reported use of real WGS data for GP was by Ober et al. [24]. In their study, they used WGS data to predict quantitative traits phenotypes in *Drosophila melanogaster*. They found that WGS added little or no benefit to the accuracy of GP when compared with HD SNP chip. It must be noted however that they had a small sample size (157). Binsbergen et al. [23] used imputed WGS data for GP in Holstein cattle with a relatively bigger sample size (3416 bulls in the reference population, and 2087 bulls in the validation set). They also could not find any advantage for using WGS over HD SNP chip. Brøndum et al. [25] reported a 2-5% increase in the accuracy of GP when some significant SNPs from a previous GWAS were added to a custom 54k SNP panel. This increase could be due to the fact that the added markers were derived from a GWAS on WGS data, therefore they are most likely very close to (if they are not themselves) the causal variants. While theoretically, the use of WGS holds the potential to greatly improve the accuracy of GP and possibly account for the hitherto missing heritability for complex traits, evidences from using real data are so far suggesting otherwise.

A number of possible explanations can be given for these disappointing results. First, the current application and success of GP is too much dependent on the existence of long-range LD due to strong family structures [16, 26, 27] and this hinders the accurate pinpointing of the genomic regions underlying complex traits and thus the utility of WGS data for GP. If the causal variants can be accurately identified, then GP can be performed directly on them and this is expected to result in more accuracy since the statistical noise caused by non-causal variants in the model is avoided [28]. Khansefid et al. [29] demonstrated how in a relatively small training population for residual feed intake (RFI) of dairy cattle, prediction accuracy was increased by giving more weight to the SNPs that had a significant effect on RFI in a beef cattle population. This approach was more effective than simply combining the data in an across breed reference population, and similar results were obtained in plants by Bernardo [30].

Secondly, it may be that WGS data is too large compared with the number of phenotypes to be efficiently handled by current statistical models. When using a few thousand markers for prediction, it is reasonable to assume that all the SNPs have an effect on the trait (the infinitesimal model), especially in cases where there are thousands of QTLs underlying the trait. This assumption however is inappropriate when using WGS data because although there are many QTLs underlying most complex traits, these QTLs are finite (thousands), e.g. human height [31]. In a study on human height, Wood et al. [31] showed that a selected 9,500 SNPs from a meta-GWAS analysis (<1% of the total SNPs in the study) at p.value , explained approximately 29% of phenotypic variance, which is more than half of what the total number of SNPs could explain (50% of phenotypic variance). Given the biological expectation that not all variants in WGS are causal, it may thus be more efficient to identify a subset of WGS variants that have some effect on the trait under analysis, and use these for prediction.

There are some studies that have demonstrated the advantage of this approach. Brøndum et al. [25] reported an increase of 2-5% in the accuracy of GP when markers identified as significant from a GWAS on WGS data were added to custom 54k SNP chip. Pérez-Enciso et al. [32] also demonstrated in a simulation study that if all the SNPs within causal genes can be identified and included in the prediction model, the accuracy of prediction could drastically be increased by approximately 40%. Heidaritabar et al. [33] reported a rather negative effect on the accuracy of GP when prior biological information (annotation information of SNPs using variants effects prediction) was incorporated in the prediction model, but as pointed out by Pérez-Enciso et al. [32], the increase in GP using this approach strongly depends on the accuracy of inferring biological relevance to variants. For further research therefore, an important question is: How can we accurately infer statistical significance or biological relevance to the variants in WGS? Moreover, how can we efficiently incorporate this information into GP?

Formulation of the problem

The use of WGS data could potentially increase the accuracy and persistency of GP, given that it contains all variants including the causal mutations that underlie complex traits. With this approach, GP will no longer depend on the existence of long range LD in the population due to strong family structures and across breeds prediction will be as accurate as within breeds prediction. Initial results using real data have however shown little or no benefit of using WGS data for GP over SNP chips. This could be due to a number of reasons, one of which is that the current statistical models cannot efficiently handle such large amounts of data. But given the biological expectation that not all variants in WGS are causal, it may thus be more efficient to identify a subset of WGS that have some effect on the trait under analysis, and use these for prediction. The research question is thus: How can we utilise the variation in WGS to increase the accuracy and persistency of genomic prediction?

This research question is further divided into five objectives as follows:

1. Is there an added advantage of prior selection of variants based on biological information or statistical significance for GP, as against the use of all available variants in sequence data? Also, what variants weighing scheme is optimum for GP? – All predictions will be carried out using the Genomic best linear Unbiased prediction (GBLUP) model.
2. Is there an added advantage of prior selection of variants based on biological information or statistical significance for GP, as against the use of all available variants in sequence data? Also, what variants weighing scheme is optimum for GP? – All predictions will be carried out using Bayesian models.
3. How can we improve the power for genome-wide association studies (GWAS) in populations with strong family structure and long range linkage disequilibrium (LD) using WGS data?
4. What is the difference in power/accuracy when single trait GWAS/GP is compared with multi-trait GWAS/GP?
5. What is the best GP strategy for new and expensive traits where small training populations will be common?

**Methodology**

The five objectives and their corresponding activities are hereby briefly discussed. Methodologies to be applied and the data are discussed in more details.

**Objective 1:**

Is there an added advantage of prior selection of variants based on biological information or statistical significance for GP, as against the use of all available variants in sequence data? Also, what variants weighing scheme is optimum for GP? – All predictions will be carried out using the genomic best linear unbiased prediction (GBLUP) model

**Activity 1:**

We will compare the accuracy of prediction for two approaches that use WGS data differently for GP. The first approach will use all available variants in WGS data while the second approach will use only a subset of WGS variants that are pre-selected based on statistical significance (from meta-GWAS analysis on WGS data) and/or functional annotation (position in the genome) or both. For both approaches, the linear (GBLUP) model will be applied throughout. We will investigate the impact of different weighing schemes for the variants on the accuracy and persistency of GP. For all scenarios, we will use a validation population that is not included in the reference population.

**Objective 2:**

Is there an added advantage of prior selection of variants based on functional annotation or statistical significance for GP, as against the use of all available variants in sequence data? Also, what variants weighing scheme is optimum for GP? – All predictions will be carried out using Bayesian models

**Activity 2**

The activities under the first objective will also be carried out here, the difference being that the non-linear (Bayesian) models will be applied throughout.

**Objective 3**

How can we increase the power of GWAS in populations with strong family structure and long range linkage disequilibrium (LD)?

**Activity 3:**

We will compare QTLs for stature as obtained by meta GWAS analysis [34] with those obtained by using standard GWAS analysis on CRV stature data. Using WGS data, we will develop strategies for increasing power to identify QTLs with small effects as identified in meta GWAS.

**Objective 4:**

What is the difference in power/accuracy when single trait GWAS/GP is compared with multi-trait GWAS/GP?

**Activity 4:**

We will explore the possibility to simultaneously predict for multiple traits in a single model, exploiting the genetic correlation between these traits. We also investigate the possibility of using one trait as a proxy to select animals for other traits. For this, we will use stature as an indicator trait for feed intake, given that feed intake is a very difficult trait to measure.

**Objective 5:**

What is the best GP strategy for new and expensive traits where small training populations will be common?

**Activity 5:**

We will perform research to define optimality criteria for the composition of reference and validation populations, taking into consideration the practical limitations that exist in in breeding programs to genotype and phenotype animal. Here, our model trait will be feed intake.

**Data**

For the entire project, data will be provide by CRV. From partner organisations in the 1,000 bull genomes project, the detailed results of a meta-analysis of GWAS on stature on 8 different breeds of cattle from 17 populations are available . The GWAS included are in the table below. This is probably the most powerful GWAS ever performed in livestock species, that also utilised full sequence data.

***Table 1. Available data for the project***

|  |  |  |  |
| --- | --- | --- | --- |
| Country | Breed | N | Sex |
| USA | Angus (AAN) | 3362 | Bulls |
| GER | Fleckvieh (FLV) | 6838 | Bulls |
| GER | Fleckvieh (FLV) | 1754 | Cows |
| CAN | Hereford (HER) | 610 | Bulls |
| AUS | Holstein (HOL) | 2271 | Bulls |
| AUS | Holstein (HOL) | 3343 | Cows |
| CAN | Holstein (HOL) | 7300 | Bulls |
| DNK | Holstein (HOL) | 5062 | Bulls |
| FRA | Holstein (HOL) | 6378 | Bulls |
| IRL | Holstein (HOL) | 5152 | Bulls |
| NLD | Holstein (HOL) | 5554 | Bulls |
| AUS | Jersey (JER) | 550 | Bulls |
| AUS | Jersey (JER) | 2497 | Cows |
| FRA | Montbeliarde (MON) | 2453 | Bulls |
| FRA | Normande (NMD) | 2095 | Bulls |
| DNK | Nordic Red (RDC) | 924 | Bulls |
| FIN | Nordic Red (RDC) | 2122 | Bulls |

Some individuals were initially genotyped for the BovineHD chip (800K) while others where genotyped with 50K (and imputed up to HD). Subsequently, individuals in each of the 17 populations were imputed to WGS using a multi-breed reference population of 1,147 sequenced animals. The reference population used for imputation was provided by the 1,000 bull genomes project. After imputation, GWAS for the trait stature with the WGS data were performed for each population separately. Bouwman et al. [34] combined the GWAS results in a meta- analysis for the trait stature, including a total of 58,265 individuals with accurate phenotypes. The results of that analysis (i.e. SNP effects and significance levels) are readily available and will be utilised for the first and second objectives.

For the first and second objectives, GP will be performed on the trait stature, building on the results of Bouwman et al. [34]. Stature is an ideal trait for several reasons, which includes the fact that it is a very simple phenotype that can be measured on many individuals. It is also similar to height in humans as was shown by the study of Pryce et al. [35]. In humans, height has been intensively studied as a classical polygenic trait and has given great insight into the genetic architecture of complex traits [31, 36]. Moreover, stature is not an economic trait and thus there is no problem when it comes to sharing data.

Genomic prediction will be performed on two independent cattle populations that were not included in the GWAS: Jersey and Holstein from New Zealand. The data will be provided by CRV (Arnhem, the Netherlands). These population will have genotypes for the BovineHD chip, that will be imputed to WGS using the most recent sequenced reference population from the 1000 bull genomes project. Also, daughter yield deviations for stature will be provided by CRV for these animals. This population will be split in a training and validation population based on the age of the animals, with the youngest animals in the validation population, or training will be performed in one breed and the other breed will be the validation population. Also, the variance explained by different G-matrices will be used to quantify the differences between genomic relationship matrix (GRM) based on select or not selected SNP.

**Methodologies for the first objective**

Different scenarios will be tested for genomic prediction and the estimated h2 using the linear GBLUP model. The scenarios are outlined below:

1. WHOLESEQ\_GBLUP1: Under this scenario, we will work with the assumption of the infinitesimal model, which is that all variants in sequence data have some effect on the trait and that these effects are sampled from the same normal distribution [27]. Thus, equal weights will be given to all variants in the construction of G-Matrix.
2. WHOLESEQ\_GBLUP2: This scenario will be similar to the first scenario, the difference being that a differential weighing scheme will be adopted for the variants in the construction of G-Matrix. SNP weights will be chosen to reflect their effect size in the previously conducted meta-GWAS analysis [34]. We will use different ranges of effect sizes to assign up to 4 different weights to the SNPs and compare this with only two weights (based on significance and non-significance at a certain p.value). Under this scenario, we could also use the 4 classes of SNPs (based on their effect size) to construct four different G-matrices and assign a separate variance component to each of the four classes in the prediction model as was done by Edwards et al. [28]
3. SSNPs\_GBLUP1: Here, only the significant SNPs identified during the previous meta-GWAS analysis [34] will be used for GP. We will select the SNPs with a p.value of , and assign the same weight to all selected SNPs in the construction of G-Matrix. We will compare the results from this procedure to one that sets the p.value at the original threshold of . The aim of this comparison will be to check if by lowering the threshold we can pick up more SNPs with smaller effects that all together significantly contribute to the total genetic variance among animals.
4. SSNPs\_GBLUP2: This scenario will be exactly the same with WHOLESEQ\_GBLUP2, except that it will be applied to only significant SNPs at p.value .
5. REGSNPs\_GBLUP1: SNPs that fall within regulatory regions of the genome will be chosen and used for GP. Here also, we will assign equal weights to all SNPs.
6. REGSNPs\_GBLUP2: this is the same as REGSNPs\_GBLUP1 but with different weights assigned to the SNPs. Weights will be chosen to reflect the biological probability of the SNPs to have some effect on the trait, given their positions in the genome.
7. COMB-GBLUP: Under this scenario, SNPs chosen for SSNPs\_GBLUP and REGSNPs\_GBLUP will be combined in one data set, with duplicates removed. Here, a new criteria for setting weights will be devised, given that the SNPs in the data set were initially selected based on two separate criteria. The new criteria will be an index that sums up the effect of the SNPs in the meta-analysis and the probability of the SNPs to have an effect given their positions in the genome.

**Methodologies for the second objective**

The scenarios here are the same as those under the first objective, but instead of GBLUP, Bayesian models will be used. The scenarios are outlined as follows:

1. WHOLESEQ\_BayesRC: GP will be carried out on WGS data using the procedure of Macleod et al. [37]. In this procedure, priors will be assigned objectively based on biological evidence in the data.
2. SSNPs\_BayesRC: SNPs that are selected for SSNPs\_GBLUP will also be used in this scenario. However, BayesRC will be applied to the data instead of GBLUP. The SNPs will be categorised into four class to fit the requirement of BayesRC and this will be done according to the effect sizes of the SNPs. Thus, setting of priors is done objectively based on statistical evidence of significance.
3. REGSNPs\_BayesRC: Here, SNPs that fall in regions of the genomes that are believed to have no regulatory roles e.g. SNPs within introns will first be removed from the data set. subsequently, BayesRC [37] will be applied on the remaining SNPs for GP.
4. COMB\_BayesRC: SNPs that are used for COMB-GBLUP will be used for this procedure, with BayesRC replacing GBLUP as a method for GP. But because SNPs in this procedure are chosen based on two different criteria (position in the genome and statistical significance), a new criteria for setting priors will be devised. This will be a simple summation of the probabilities of the SNPs to have an effect on the trait (given their positions in the genome), and the effect size of the SNPs (from the meta-GWAS analysis). This new sum will be used as a criteria to divide the SNPs into four classes.

The aim for the third paper is to devise strategies that increases the power of GWAS, utilising WGS data. We will focus especially on populations with strong family relationships and long range LD. We will compare QTLs for stature as obtained by meta GWAS analysis [34] with those obtained by using standard GWAS analysis on CRV stature data. Ultimately, we aim to bring the power of a standard GWAS as close as possible to power obtainable in meta-GWAS analysis.

It may be become interesting to simultaneously predict the GEBVs of animals for multiple traits, especially those that are genetically correlated due to LD between the QTLs affecting them or due to QTLs with pleiotropic effects on them. The aim of the fourth paper will be to establish a framework for multi-trait GP and GWAS using WGS data. We will compare the difference in accuracy/power when single trait GP/GWAS is compared with multi-trait GP/GWAS.

For the fifth paper, the optimal properties of the training and validation populations will be investigated, given that for some expensive traits such as feed intake, only small reference populations exist. Here we will use both real and simulated data.

**Work Plan**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Jan | Feb | Mar | April | May | June | July | Aug | Sept | Oct | Nov | Dec |
| 2016 |  | | | | WIAS proposal / Objective 1 | | | | | Objective 1 | | |
| 2017 | Objective 1 | | Objective 2 | | | | | | | | Objective 3 | |
| 2018 | Objective 3 | | | | | Objective 4 | | | | | | |
| 2019 | Objective 4 | | Objective 5 | | | | | | | | | |
| 2020 | General Discussion | | | |  | | | | | | | |

**Signatures**

Daily supervisor Promotor

Name: Dr. AC Bouwman Name: Prof. Dr. Jeanine Houwing-Duistermaat

Signature: Signature:

Daily supervisor/Promoter Supervisor

Name: Prof. Dr. Roel Veerkamp Name: Prof. Dr. Fred van Eeuwijk

Signature: Signature:

**Literature**

1. Henderson, C.R., *Applications of Linear Models in Animal Breeding*. 1984, University of Guelph, Guelph, ON.

2. Simm, G., *Genetic improvement of cattle and sheep: putting science into practice.* Journal of the Royal Agricultural Society of England (United Kingdom), 1997.

3. Hayes, B. and M. Goddard, *Prediction of total genetic value using genome-wide dense marker maps.* Genetics, 2001. **157**(4): p. 1819-1829.

4. Georges, M., et al., *Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing.* Genetics, 1995. **139**(2): p. 907-920.

5. Wientjes, Y.C.J., R.F. Veerkamp, and M.P.L. Calus, *The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction.* Genetics, 2013. **193**(2): p. 621-631.

6. Dalton, R., *No bull: genes for better milk.* Nature News, 2009. **457**(7228): p. 369-369.

7. Heffner, E.L., J.-L. Jannink, and M.E. Sorrells, *Genomic selection accuracy using multifamily prediction models in a wheat breeding program.* The Plant Genome, 2011. **4**(1): p. 65-75.

8. Hayes, B., et al., *Invited review: Genomic selection in dairy cattle: Progress and challenges.* Journal of dairy science, 2009. **92**(2): p. 433-443.

9. Schaeffer, L., *Strategy for applying genome‐wide selection in dairy cattle.* Journal of Animal Breeding and Genetics, 2006. **123**(4): p. 218-223.

10. VanRaden, P., et al., *Invited review: Reliability of genomic predictions for North American Holstein bulls.* Journal of dairy science, 2009. **92**(1): p. 16-24.

11. Saatchi, M. and D. Garrick. *Across breed genomic predictions in beef cattle*. in *10th World Congress on Genetics Applied to Livestock Production*. 2014. Asas.

12. Muir, W., *Comparison of genomic and traditional BLUP‐estimated breeding value accuracy and selection response under alternative trait and genomic parameters.* Journal of Animal Breeding and Genetics, 2007. **124**(6): p. 342-355.

13. Goddard, M.E. and B.J. Hayes, *Mapping genes for complex traits in domestic animals and their use in breeding programmes.* Nature Reviews Genetics, 2009. **10**(6): p. 381-391.

14. Daetwyler, H.D., et al., *The impact of genetic architecture on genome-wide evaluation methods.* Genetics, 2010. **185**(3): p. 1021-1031.

15. Iheshiulor, O., et al. *Genomic Predictions Using Whole Genome Sequence Data and Multi-breed Reference Populations*. in *10th World Congress on Genetics Applied to Livestock Production*. 2014. Asas.

16. Habier, D., et al., *The impact of genetic relationship information on genomic breeding values in German Holstein cattle.* Genetics Selection Evolution, 2010. **42**(1): p. 5.

17. Wientjes, Y.C., et al., *Impact of QTL properties on the accuracy of multi-breed genomic prediction.* Genet Sel Evol, 2015. **47**: p. 42.

18. Meuwissen, T., B. Hayes, and M. Goddard, *Genomic selection: A paradigm shift in animal breeding.* Animal Frontiers, 2016. **6**(1): p. 6-14.

19. Meuwissen, T. and M. Goddard, *Accurate prediction of genetic values for complex traits by whole-genome resequencing.* Genetics, 2010. **185**(2): p. 623-631.

20. MacLeod, I.M., B.J. Hayes, and M.E. Goddard, *The effects of demography and long-term selection on the accuracy of genomic prediction with sequence data.* Genetics, 2014. **198**(4): p. 1671-1684.

21. Clark, S.A., J.M. Hickey, and J.H. Van der Werf, *Different models of genetic variation and their effect on genomic evaluation.* Genet Sel Evol, 2011. **43**(18): p. 10.1186.

22. Druet, T., I. Macleod, and B. Hayes, *Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions.* Heredity, 2014. **112**(1): p. 39-47.

23. van Binsbergen, R., et al., *Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle.* Genetics Selection Evolution, 2015. **47**(1): p. 1-13.

24. Ober, U., et al., *Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster.* PLoS Genet, 2012. **8**(5): p. e1002685.

25. Brøndum, R.F., et al., *Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction.* Journal of Dairy Science, 2015. **98**(6): p. 4107-4116.

26. Gianola, D., et al., *Additive genetic variability and the Bayesian alphabet.* Genetics, 2009. **183**(1): p. 347-363.

27. Habier, D., R. Fernando, and J. Dekkers, *The impact of genetic relationship information on genome-assisted breeding values.* Genetics, 2007. **177**(4): p. 2389-2397.

28. Edwards, S.M., et al., *Genomic Prediction for Quantitative Traits Is Improved by Mapping Variants to Gene Ontology Categories in Drosophila melanogaster.* Genetics, 2016.

29. Khansefid, M., et al. *Increasing the Accuracy of Genomic Predictions for RFI in Dairy Cattle through Using Genomic Information from Beef Breeds*. in *10th World Congress on Genetics Applied to Livestock Production*. 2014. Asas.

30. Bernardo, R., *Genomewide selection when major genes are known.* Crop Science, 2014. **54**(1): p. 68-75.

31. Wood, A.R., et al., *Defining the role of common variation in the genomic and biological architecture of adult human height.* Nature genetics, 2014. **46**(11): p. 1173-1186.

32. Pérez-Enciso, M., J.C. Rincón, and A. Legarra, *Sequence-vs. chip-assisted genomic selection: accurate biological information is advised.* Genet Sel Evol, 2015. **47**: p. 43.

33. Heidaritabar, M., et al., *Accuracy of genomic prediction using imputed whole-genome sequence data in white layers.* Journal of Animal Breeding and Genetics, 2016. **133**(3): p. 167-179.

34. Bouwman, A.C., et al. *Meta-analysis of GWAS of bovine stature with> 50,000 animals imputed to whole-genome sequence*. in *Annual Meeting of the European Association for Animal Production*. 2015. Wageningen Academic Publishers.

35. Pryce, J.E., et al., *Polymorphic regions affecting human height also control stature in cattle.* Genetics, 2011. **187**(3): p. 981-984.

36. Yang, J., et al., *Common SNPs explain a large proportion of the heritability for human height.* Nature genetics, 2010. **42**(7): p. 565-569.

37. MacLeod, I., et al., *Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits.* BMC genomics, 2016. **17**(1): p. 1.

**Suggestion for referees**

Please name three referees who are not involved in the project or in the participating groups. These referees should be able to give an independent judgement on the scientific quality and feasibility of the project. At least two of the referees must be from abroad.

Referee 1

name:

affiliation:

area of expertise:

full address:

phone:

fax:

e-mail:

Referee 2

name:

affiliation:

area of expertise:

full address:

phone:

fax:

e-mail:

Referee 3

name:

affiliation:

area of expertise:

full address:

phone:

fax:

e-mail: