Prospects of whole-genome sequence data in animal and plant breeding

Prospects of whole-genome sequence data in animal and plant breeding

Rianne van Binsbergen

Rianne van Binsbergen

PROSPECTS OF WHOLE-GENOME SEQUENCE DATA IN ANIMAL AND PLANT BREEDING

Rianne van Binsbergen

Thesis committee

Promotors

Prof. Dr R.F. Veerkamp
Special professor, Numerical Genetics and Genomics
Wageningen University & Research

Prof. Dr F.A. van Eeuwijk
Professor of Applied Statistics
Wageningen University & Research

Co-promotor

Dr M.P.L. Calus
Senior researcher, Animal Breeding and Genomics
Wageningen University & Research

Other members

Prof. Dr J.M. Hickey, The Roslin Institute, United Kingdom
Prof. Dr H.P. Piepho, University of Hohenheim, Germany
Prof. Dr H. Simianer, Georg-August-University, Germany
Prof. Dr B.J. Zwaan, Wageningen University & Research

This research was conducted under the auspices of the Graduate School of Wageningen Institute of Animal Sciences (WIAS)

PROSPECTS OF WHOLE-GENOME SEQUENCE DATA IN ANIMAL AND PLANT BREEDING

Rianne van Binsbergen

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 5 July 2017
at 1.30 p.m. in the Aula.

Rianne van Binsbergen
Prospects of whole-genome sequence data in animal and plant breeding,
222 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2017)
With references, with summary in English and Dutch

ISBN 978-94-6343-190-3
DOI http://dx.doi.org/ 10.18174/413524

ABSTRACT

Van Binsbergen, R. (2017). Prospects of whole-genome sequence data in animal and plant breeding. PhD thesis, Wageningen University, the Netherlands

The rapid decrease in costs of DNA sequencing implies that whole-genome sequence data will be widely available in the coming few years. Whole-genome sequence data includes all base-pairs on the genome that show variation in the sequenced population. Consequently, it is assumed that the causal mutations (e.g. quantitative trait loci; QTL) are included, which allows testing a given trait directly for association with a QTL, and might lead to discovery of new QTL or higher accuracies in genomic predictions compared to currently available marker panels. The main aim of this thesis was to investigate the benefits of using whole-genome sequence data in breeding of animals and plants compared to currently available marker panels. First the potential and benefits of using whole-genome sequence data were studied in (dairy) cattle. Accuracy of genotype imputation to whole-genome sequence data was generally high, depending on the used marker panel. In contrast to the expectations, genomic prediction showed no advantage of using whole-genome sequence data compared to a high density marker panel. Thereafter, the use of whole-genome sequence data for QTL detection in tomato (*S. Lycopersicum*) was studied. In a recombinant inbred line (RIL) population, more QTL were found when using sequence data compared to a marker panel, while increasing marker density was not expected to provide additional power to detect QTL. Next to the RIL population, also in an association panel it was shown that, even with limited imputation accuracy, the power of a genome-wide association study can be improved by using whole-genome sequence data. For successful application of whole-genome sequence data in animals or plants, genotype imputation will remain important to obtain accurate sequence data for all individuals in a cost effective way. Sequence data will increase the power of QTL detection in RIL populations, association panels or outbred populations. Added value of whole-genome sequence data in genomic prediction will be limited, unless more information is known about the biological background of traits and functional annotations of DNA. Also statistical models that incorporate this information and that can efficiently handle large datasets have to be developed.

CONTENTS

CHAPTER 1

GENERAL INTRODUCTION

1.1 Classical breeding

Traditionally, selection of animals and plants has been based on own phenotypic records and pedigree information. Depending on characteristics of a species and the breeding goal, the breeding programs differ. For example, in dairy cattle the goal is to select the best parents to improve the next generation. In general, dairy farmers own the cows and buy semen from breeding companies to produce this next generation, so the farmers breed the next generation in their own herd. Breeding companies aim to breed the most attractive bull for farmers and provide information about the genetic merit of their bulls (breeding values), for example for milk production, conformation traits, and fertility traits. However, bulls don't produce milk or express the traits that are important for dairy farmers, so to obtain a reliable estimate of a bulls' breeding value breeders traditionally made use of information of many daughters. As a consequence it took several years before the breeding value of a bull is known.

In pig and poultry breeding the breeding company makes the decision about mating of the animals and sells the end product (piglets, eggs, or chickens) to production farms. These breeding companies make use of multiple purebred lines, which are selected on for example product quality traits or fertility traits. The end product is a cross between those purebred lines. By applying this crossbred scheme, companies make use of heterosis, i.e. an increase in trait value of the cross compared to the parents (Falconer and Mackay, 1996).

In many plant breeding programs inbreeding is common practice. Inbreeding allows obtaining uniform populations or lines with all the same genotype. It should be noted that the definition of 'line' for plant breeders refers often to this kind of populations which are purebred, homozygous and genetically homogeneous (all genotypes are the same and remain the same upon selfing), whereas in animal breeding 'line' refers to multiple individuals with similar genotypes, but still considerable genetic variation is present. A consequence of a cultivar being purebred is that farmers can maintain the genotypes by collection of seeds in case of species that do not out-cross often (e.g. tomato). To avoid this, many breeding companies developed F1 hybrids, which forces farmers to buy every season new seeds (Bai and Lindhout, 2007).

1.2 Molecular breeding

These classical breeding approaches have been successful, but also time-consuming and costly. Since a few decades molecular markers (known DNA sequences) can be used for selection. Marker location and effects on traits of interest can be estimated using phenotyped individuals which are also genotyped for these markers. In most cases these markers do not have a direct effect on the trait, but are linked to loci on the DNA that do have an effect (e.g. due to short distance). These loci that do have an effect are often called quantitative trait loci (QTL) or quantitative trait nucleotide (QTN). Position and effect of a QTL can be estimated based on the markers, this approach is also known as QTL mapping. The found QTLs can provide more knowledge on animal or plant physiology , can be used to select the best individuals for further breeding, or can be used to improve cultivars (e.g. by marker assisted backcrossing; Dekkers and Hospital, 2002). The main advantage using these DNA markers in contrast to classical breeding approaches is the decrease in generation interval, i.e. the genetic merit of an individual can be predicted without waiting until the phenotype can be measured or without performing extensive phenotypic trials.

1.2.1 QTL mapping

Use of QTL information in breeding practice, requires that these QTL at first should be located. A powerful way to find QTL is by making use of meiotic recombination events, for example by using mapping populations. In case of self-pollinating species like tomato, mapping populations originate from a cross between two parents that are highly inbred and differ for one or more traits of interest (Collard *et al.*, 2005). Two types of mapping populations are shown in Figure 1.1: backcross population and recombinant inbred lines. Due to recombination, each individual will carry a slightly different combination of alleles. A backcross population is relatively simple to obtain by crossing a F_1 to one of the parents. Recombinant inbred lines (RIL) are obtained by repeated selfing F_2 individuals for a number of generations. Compared to the backcross population, the time to produce a RIL population is long. The advantage of a RIL population, however, is that after six to eight generations they are almost fully homozygous and can be multiplied without genetic change, which makes them very useful to

explore the relationship between genotype and phenotype under different conditions.

QTL analyses in these mapping populations are based on the contrast between the two parental alleles for the traits of interest. For every marker it can be tested if there is a difference in phenotypic mean for the individuals with marker genotype coming from P_1 and the phenotypic mean for the individuals with marker genotype coming from P_2. However, this can only be done for traits that segregate between the parents and only the variation that exists between the two parents can be detected.

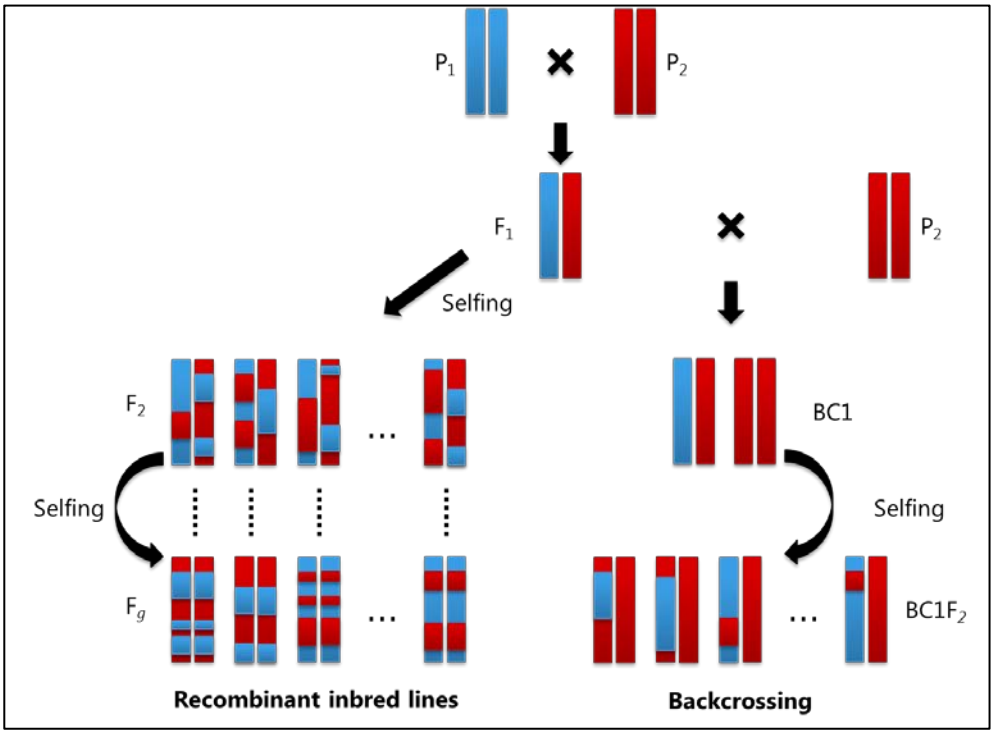

Figure 1.1 Diagram of two types of mapping populations for self-pollinating species

The drawback of these mapping population studies is the relatively large interval in which a QTL might occur. All studies make use of the correlation between a marker and QTL, i.e. linkage disequilibrium (LD), which on average will decrease if the distance between a marker and QTL increases. In case of double haploid (Mayor and Bernardo, 2009) populations this LD stretches over a long distance, so the LD decay is long. This LD decay shortens when moving to

backcross populations, F_2-populations, recombinant inbred lines, and multi-parent advanced generation inter-cross (MAGIC) populations (Huang *et al.*, 2015). In an association panel or (natural) outbred populations the shortest LD decay can be found. The LD decay is dependent on the breeds and family relationships in such a population. For example, considering only one breed can cause long range LD, while by combining multiple populations shorter LD decay can be reached (e.g. dairy cattle; de Roos *et al.*, 2008). Short LD decay decreases the power of QTL detection, but also decreases the QTL interval. Next to the size of QTL effect and minor allele frequency of the QTL, a large population of phenotyped and genotyped individuals is needed to have enough power to detect a causal variant (Spencer *et al.*, 2009).

1.2.2 Genomic prediction

The found QTL provide more knowledge on animal or plant physiology, but can also be used to select the best individuals for further breeding, which is using so-called marker-assist selection. Marker-assist selection is especially useful for traits controlled by a few QTL with large effect. Unfortunately, the number of QTN identified in animals is rather limited (Dekkers, 2004), and for a lot of traits a more polygenic nature is expected, i.e. traits are controlled by many loci with a small effect. In this situation it is difficult to locate all QTL, and as a consequence the proportion of genetic variance that is capitalized by marker-assisted selection is limited (Goddard and Hayes, 2009). For these traits with polygenic inheritance, another strategy should be used to predict genomic potential of an individual.

Meuwissen *et al.* (2001) first described a genome wide prediction method that uses all available markers on a population to predict their additive genetic merit. In this situation it is assumed that all markers together capture the genetic variance of all QTL due to linkage disequilibrium between the markers and QTL. Nowadays, this genomic prediction approach is applied in many breeding programs, for example in cattle (Hayes *et al.*, 2009), pigs (Knol *et al.*, 2016), and maize (Crossa *et al.*, 2013)

In recent years, massive amounts of marker data have been generated in animals and plants, both for research purposes and for application in commercial breeding and selection programs. Often the number of markers exceeds the number of records, which requires methods that can handle this *large-p* with *small-*

n problem, for example by variable selection or shrinkage (de los Campos *et al.*, 2013). At the moment, a lot of linear regression and prediction models for genomic selection in animal and plants are available, which differ in assumptions about the genetic architecture of the trait and the type of shrinkage estimation procedure (de los Campos *et al.*, 2013). As well as the prediction model, other factors, such as marker density, trait heritability, genetic architecture, reference group size, and relationship to the reference group (de Roos *et al.*, 2009; Daetwyler *et al.*, 2010; Pszczola *et al.*, 2012) can influence accuracy of genomic prediction.

1.3 Whole-genome sequence data

With higher marker density the chance that a causal mutation, or a marker in high LD with this causal mutation, is included in the genotyping data increases. In case of currently used marker panels the causal mutations are probably not included in the data and therefore genomic prediction and QTL detection approaches rely on LD between a marker and the causal mutation. The chance that a marker is in high LD with a causal mutation will increase when the number of markers increases. An extreme case of increasing marker density is the use of whole-genome sequence data, i.e. using all base-pairs on the genome that show variation in the population. In case of whole-genome sequence data it is assumed that the causal mutations are included in the data (Meuwissen and Goddard, 2010), at least when enough phenotyped individuals are sequenced to establish these polymorphisms (Meuwissen and Goddard, 2010; Macleod *et al.*, 2013; Druet *et al.*, 2014). In that scenario, there will be less dependency on linkage disequilibrium (LD) compared to SNP array panels. This allows testing a given trait directly for association with the QTL, which might lead to discovery of new QTL or higher accuracy in genomic predictions compared to SNP panel genotypes.

Currently, is genomic prediction across multiple generations or multiple populations difficult, for example because LD between a QTN and a SNP differs across populations (de Roos *et al.*, 2008; Wientjes *et al.*, 2015). By directly testing for a QTL and remove the dependency between marker and QTL, precision of QTL detection and the persistency of genomic prediction across generations or multiple populations is expected to increased (Meuwissen and Goddard, 2010; Clark *et al.*, 2011; Macleod *et al.*, 2013). However, to find the QTL in the data and reach this higher persistency of accuracy of genomic predictions over generations, probably a

large training set of thousands of sequenced individuals is needed. Without a large number of training individuals QTL effects might be estimated with too much error, resulting in little advantage (Druet *et al.*, 2014). Therefore, the objective of this thesis was to investigate these potential benefits of using whole-genome sequence data in breeding of animals and plants compared to currently available marker panels.

1.4 Genotype imputation

To investigate these potential benefits, whole-genome sequence of many individuals is needed. The National Human Genome Research Institute has tracked the costs of DNA sequencing since 2001 (https://www.genome.gov/sequencingcostsdata/). In Figure 1.2 the costs of sequencing a human sized genome are plotted from September 2001 until October 2015. In this period of time the costs decreased rapidly from close to $100,000,000 in 2001 to approximately $1,000 in 2015. This is a much faster decrease than expected using Moore's law (Moore, 1965), which is often used in technology (e.g. computer hardware industry) and predicts a doubling in computing power every two years. Technologies following Moore's law are considered to do extremely well.

Despite the fact that costs of sequencing are decreasing, it is still expensive to sequence large numbers of individuals. A solution might be to sequence a core set of individuals and use this information as reference to predict whole-genome sequence genotypes of other individuals genotyped at lower density, i.e. genotype imputation. At the moment, lots of different methods for genotype imputation are available. Roughly, those methods can be divided into four main categories: 1) naïve approaches; 2) general statistical approaches; 3) family-based approaches; and 4) population-based approaches.

Naïve approaches include filling in the marker mean value or, in case of inbred lines the heterozygous value. Naïve approaches are mainly used if no missing genotype data is allowed in subsequent analyses and the data contains sporadic missing genotypes. Non-parametric methods like random forest or k-nearest neighbor are examples of general statistical approaches. With random forest imputation a combination of tree predictors is established based on a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001; Stekhoven and Bühlmann, 2012). The K-nearest

neighbor relies on filling missing data points with the weighted mean of the k most similar genotypes based on Euclidean distance between standardized observations (Troyanskaya *et al.*, 2001). Both naïve approaches and general statistical approaches assume data is missing at random. However, in many cases the same SNP genotypes are missing for a number of individuals. For example, when using genotype imputation of sequence data into lower density marker panels. In this situation family-based approaches and population-based approaches could be more suitable. However, these methods need a reference genome, whereas approaches like random forest and k-nearest neighbor are map-independent and do not require information on marker order of phase.

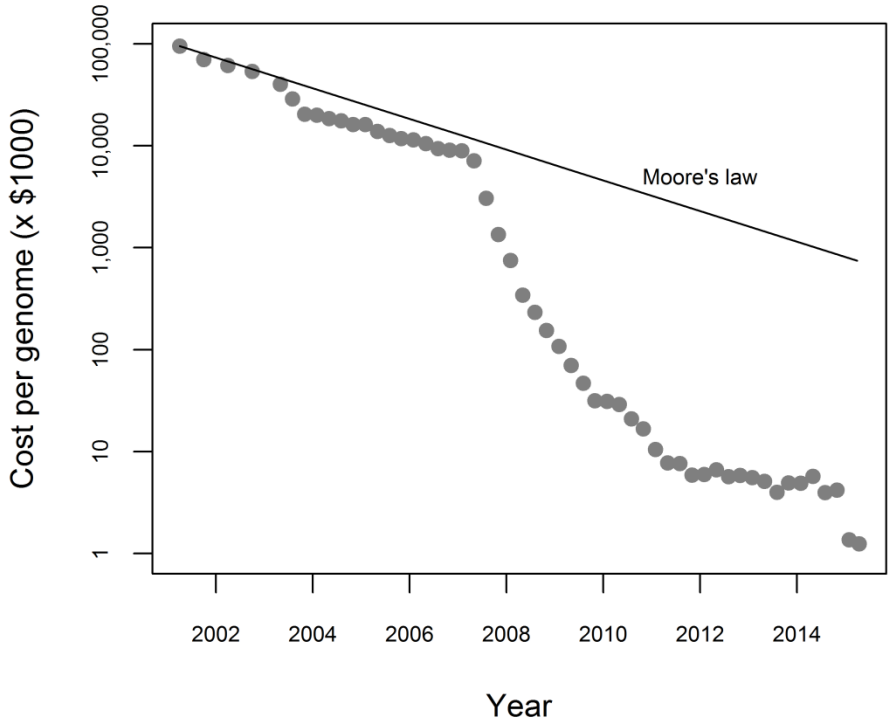

Figure 1.2 Costs of DNA sequencing of a human sized genome from 2001 to 2015 (source: https://www.genome.gov/sequencingcostsdata/)

Family-based methods make use of pedigree and linkage information and are most accurate if the reference set include close relatives. Most of these algorithms are designed for livestock populations, for example FImpute (Sargolzaei *et al.*, 2011), findhap (VanRaden *et al.*, 2011), and AlphaImpute (Hickey *et al.*,

2012b). However, if no (reliable) pedigree information is available, population-based approaches can be used. Many of these methods were designed for human studies, where reliable pedigree information is often missing. Those methods utilize linkage disequilibrium information between SNPs in a population. Examples of population-based approaches are Beagle (Browning and Browning, 2009), IMPUTE (Howie *et al.*, 2009), MaCH (Li *et al.*, 2010), fastPHASE (Scheet and Stephens, 2006), and SHAPEIT (Delaneau *et al.*, 2013).

Performance of the different imputation methods depend on several factors, including the structure of the data used. Studies in barley (Iwata and Jannink, 2010), maize (Hickey *et al.*, 2012a), sheep (Hayes *et al.*, 2012) and cattle (e.g. Druet *et al.*, 2010; VanRaden *et al.*, 2013), among others showed that accuracy of imputation increased with increased number of SNPs on lower density marker panel, decreased distance between imputed SNP to nearest SNP on lower density marker panel, increased minor allele frequency of imputed SNP, increased level of LD, and increased number of close relatives between imputed and reference individuals. This would suggests that in bi-parental populations, like RILs, with high levels of LD and allele frequencies around 0.5, genotype imputation should be straightforward (e.g. Jiang and Zeng, 1997).

1.5 Aim and outline of this thesis

The rapid decrease in costs of DNA sequencing implies that whole-genome sequence data will be widely available in the coming few years. The hypothesis is that the use of whole-genome sequence data could increase accuracy of genomic prediction and increase the power of QTL detection. Therefore, the main aim of this thesis was to investigate the benefits of using whole-genome sequence data in breeding of animals and plants compared to currently available marker panels. The thesis focuses on genotype imputation, genomic prediction, and QTL detection.

In the first part the benefits of using whole-genome sequence data in cattle were studied in cattle. **Chapter 2** describes the work of a large consortium to collect whole-genome sequence data for 234 individuals. One goal of this collaboration was to build a dataset of sequenced individuals that can be used as reference for imputation of sequence data in cattle populations. Accuracy of imputation of the sequence variants into larger data sets genotyped with SNP arrays was assessed using fivefold cross-validation, and potential benefits of whole-

genome sequence data with respect to genome wide association studies were demonstrated. Next, in **Chapter 3** the accuracy of imputation of genotypes from two commercially used SNP panels to whole-genome sequence data in Holstein Friesian dairy cattle was studied in more detail, including factors that could influence accuracy of imputation to sequence data. Accordingly, in **Chapter 4** whole-genome sequence data was imputed into BovineHD genotypes of 5503 Holstein Friesian bulls. Reliability of genomic prediction based on this imputed whole-genome sequence data was compared to genomic prediction using BovineHD genotypes. For genomic prediction two different methods were tested: with the first method it was assumed a priori that the variance of SNP effects is equal; while with the second method it was assumed a priori that the effects of many SNPs is very small or zero, and for only a few SNPs it is large.

The second part of this thesis focuses on the use of whole-genome sequence data in tomato (*S. Lycopersicum*). In **Chapter 5** it was investigated if the use of (imputed) sequence data would reveal additional QTL, relative to a SNP array panel, when performing QTL analyses in a tomato RIL population. Recombinant inbred lines represent a very small part of the variation present in tomato, i.e. only variation of two inbred accessions is captured. Therefore, in the next chapter (**Chapter 6**) we investigated the added value of (imputed) sequence data relative to a SNP array panel when performing a genome-wide association study in a panel of 145 accessions. With the long LD decay in a RIL population (**Chapter 5**) it was expected that increasing marker density to whole-genome sequence data gave no additional power with respect to QTL mapping. Whereas, it was expected that the genome-wide association study in **Chapter 6** would benefit of this increased marker density.

Finally, in **Chapter 7** the results from the previous chapters are discussed within the context of the current status for dairy cattle breeding and tomato breeding with respect to the use of whole-genome sequence data. Thereafter some future perspectives with respect to genotype imputation, genomic prediction, and QTL detection using whole-genome sequence data are discussed.

CHAPTER 2

WHOLE-GENOME SEQUENCING OF 234 BULLS FACILITATES MAPPING OF MONOGENIC AND COMPLEX TRAITS IN CATTLE

HANS D. DAETWYLER[1,2,3]
AURÉLIEN CAPITAN[4,5]
HUBERT PAUSCH[6]
PAUL STOTHARD[7]
RIANNE VAN BINSBERGEN[8]
RASMUS F. BRØNDUM[9]
XIAOPING LIAO[7]
ANIS DJARI[10]
SABRINA C. RODRIGUEZ[4]
CÉCILE GROHS[4]
DIANE ESQUERRÉ[11]
OLIVIER BOUCHEZ[11]
MARIE-NOËLLE ROSSIGNOL[12]
CHRISTOPHE KLOPP[10]
DOMINIQUE ROCHA[4]
SÉBASTIEN FRITZ[5]
ANDRÉ EGGEN[4]
PHIL J. BOWMAN[1,3]
DAVID COOTE[1,3]
AMANDA J. CHAMBERLAIN[1,3]
CHARLOTTE ANDERSON[1]
CURT P. VANTASSELL[13]
INA HULSEGGE[8]
MIKE E. GODDARD[1,3,14]
BERNT GULDBRANDTSEN[9]
MOGENS S. LUND[9]
ROEL F. VEERKAMP[8]
DIDIER A. BOICHARD[4]
RUEDI FRIES[6]
BEN J. HAYES[1,2,3]

Nature Genetics (2014) 46:858–865

Abstract

The 1000 bull genomes project supports the goal of accelerating the rates of genetic gain in domestic cattle while at the same time considering animal health and welfare by providing the annotated sequence variants and genotypes of key ancestor bulls. In the first phase of the 1000 bull genomes project, we sequenced the whole genomes of 234 cattle to an average of 8.3-fold coverage. This sequencing includes data for 129 individuals from the global Holstein-Friesian population, 43 individuals from the Fleckvieh breed and 15 individuals from the Jersey breed. We identified a total of 28.3 million variants, with an average of 1.44 heterozygous sites per kilobase for each individual. We demonstrate the use of this database in identifying a recessive mutation underlying embryonic death and a dominant mutation underlying lethal chrondrodysplasia. We also performed genome-wide association studies for milk production and curly coat, using imputed sequence variants, and identified variants associated with these traits in cattle.

---

[1]Biosciences Research Division, Department of Environment and Primary Industries, Bundoora, Victoria, Australia. [2]School of Applied Systems Biology, La Trobe University, Bundoora, Victoria, Australia. [3]Dairy Futures Cooperative Research Centre, Bundoora, Victoria, Australia. [4]Institut National de la Recherche Agronomique (INRA), UMR 1313 Génétique Animale et Biologie Intégrative, Jouy-en-Josas, France. [5]Union Nationale des Coopératives d'Elevage et d'Insémination Animale, Paris, France. [6]Chair of Animal Breeding, Technische Universitaet Muenchen, Freising-Weihenstephan, Germany. [7]Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, Alberta, Canada. [8]Animal Breeding and Genomics Centre, Wageningen University and Research Centre, Livestock Research, Wageningen, the Netherlands. [9]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark. [10]Institut National de la Recherche Agronomique (INRA), Sigenae Bioinformatics Group, UR875, Castanet, France. [11]Institut National de la Recherche Agronomique (INRA), Get-Plage, UMR 444 Laboratoire de Génétique Cellulaire, Castanet, France. [12]LABOGENA, Jouy-en-Josas, France. [13]US Department of Agriculture, Agricultural Research Service (USDA-ARS), Animal and Natural Resources Institute, Bovine Functional Genomics Laboratory, BARC-East, Beltsville, Maryland, USA. [14]Faculty of Land and Food Resources, University of Melbourne, Parkville, Victoria, Australia.

2.1 Introduction

Cattle were domesticated approximately 10,000 years ago to secure supplies of meat, milk and power. Recently, genomic selection has been adopted globally by cattle industries to accelerate genetic gains (Dalton, 2009). To meet projected global demands for milk and meat, rates of genetic gain must be further accelerated. At the same time, animal health and welfare must also be considered. Improved accuracy of genomic predictions (Meuwissen and Goddard, 2010a; Druet *et al.*, 2014) and rapid identification and management of genetic defects could be achieved if genome sequence data were available for large numbers of cattle phenotyped for traits of interest. However, given the genetic architecture of production traits in cattle (Cole *et al.*, 2009), in which large numbers of loci individually explain relatively little genetic variation, the number of individuals required with both phenotype and genomic sequence would be cost prohibitive.

One cost-effective approach to attain a large number of individuals with sequence data is to make use of the relatively small effective population size (N_e) of many cattle breeds—a result of the intense selection of sires and artificial insemination—with heavy use of limited numbers of key ancestor bulls (The Bovine HapMap Consortium, 2009). Once these ancestors are sequenced, descendants need only be genotyped with a dense SNP array to accurately infer their genome sequence (Druet *et al.*, 2014) by tracing the large segments of chromosomes inherited from the ancestor bulls.

The aim of the 1000 bull genomes project is twofold: (i) to build a database of sequence variant genotypes of individuals, ideally key ancestors, from modern cattle breeds that enables sequence-based genome-wide association studies (GWAS) and genomic prediction and (ii) to enable the use of these same data to rapidly identify mutations that compromise animal health, welfare and productivity.

We identified 129 key ancestors from the global Holstein-Friesian (Holstein) population, 43 key ancestors from the Fleckvieh breed and 15 key ancestors from the Jersey breed for whole-genome sequencing. We included 54 parent-offspring pairs to assess the quality of genotype calls in the sequence data. A large proportion of the chromosomes in modern dairy cattle populations can be traced back to these ancestors—72.3%, 43.3% and 55.3% of the chromosomes in the modern Australian, French and Danish Holstein populations, respectively—and 69%

of the chromosomes in the modern German Fleckvieh population and 59% of chromosomes in the Australian Jersey population can be traced back to the Jersey ancestors (Boichard *et al.*, 1997). Additionally, 47 Angus cattle from high– and low–feed conversion selection groups had been sequenced, and these data were also available for the current project. Although these individuals are not key ancestors, they are still expected to capture a proportion of the genetic diversity within their breed.

2.2 Results

All cattle were sequenced using Illumina sequencing-by-synthesis technology (Bentley *et al.*, 2008; Zimin *et al.*, 2009). Sequence reads were filtered for quality and then aligned to the UMD3.1 reference sequence using the Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009). Individual sequence coverage was 8.3-fold on average and ranged from 2- to 45-fold. Variants (SNPs and indels) were identified and genotyped considering sequence from all cattle using SAMtools 0.0.18 (Li *et al.*, 2009), and these variants were custom filtered (Supplementary Note). BEAGLE (Browning and Browning, 2009) was used to improve genotype calls using genotype likelihoods from SAMtools and inferred haplotypes in the sample.

After filtering, 1.6 million indels and 26.7 million SNPs were identified (dbSNP handle '1000_BULL_GENOMES'). Concordance of sequence genotypes with BovineHD SNP array (800K) genotypes ranged from 96.2% to 98.9% (Supplementary Table 1). BEAGLE correction improved genotype concordance by up to 7.0%, with the greatest improvements for individuals with low-coverage sequencing data (Supplementary Fig. 1). The mean rate of opposing homozygosity for parent-offspring pairs was very low for most of the genome (0.6% of genotypes on average), although this rate was high for a small number of regions (Supplementary Fig. 2).

Holstein and Fleckvieh genomes shared the greatest number of polymorphic variants (Supplementary Fig. 3), likely reflecting a recent introgression of Holstein cattle into the Fleckvieh breed (Pausch *et al.*, 2011). Angus and Jersey genomes shared the lowest number of polymorphic variants, consistent with F_{ST} (fixation index) estimates from SNP array data (The Bovine HapMap Consortium, 2009). The numbers of variants per breed broadly reflected effective population sizes (the Fleckvieh breed had the largest N_e), as well as the selection of bulls from the breed

that were sequenced. For example, the sequenced Holstein bulls included individuals that predate the reduction in N_e resulting from intensive selection and the widespread use of artificial insemination, and, in fact, these bulls were primarily responsible for the reduction in N_e, as they were used very heavily in the industry. Of all the variants identified, 86,152 were homozygous for the alternate allele relative to the reference in all sequenced bulls.

2.2.1 Variant annotation and allele frequency spectra

Comparison with entries in dbSNP identified 5.8 million SNPs (21.6%) as known, and the remaining SNPs were classified as novel. For indels, 140,929 (9.1%) had been described previously. Variants were annotated using NGS-SNP (Grant *et al.*, 2011) (Supplementary Table 2). We observed that indels in coding regions were enriched for 3-, 6- and 9-bp indels, as has been observed in human data (Fujimoto *et al.*, 2010) (Supplementary Fig. 4). Such polymorphisms are likely to be more tolerated and less selected against than those inducing a frameshift.

Variants predicted by SIFT (Kumar *et al.*, 2009) to be deleterious had much lower minor allele frequency (MAF) than those predicted to be tolerated by SIFT (Supplementary Fig. 5). Variants that introduced premature stop codons and those predicted by SIFT to be deleterious occurred almost exclusively in single breeds, indicating that these variants are more likely to have arisen recently. A variant predicted by SIFT to be deleterious was located in *ITGB2*. This mutation was previously reported to cause the disease bovine leukocyte adhesion deficiency (BLAD) in homozygote individuals, resulting in a phenotype of extreme susceptibility to infection and early mortality (Shuster *et al.*, 1992). Our genotype calls correctly indicated that the mutation only segregated in Holstein bulls (it has not been reported in other breeds), and only four of these bulls were carriers, all of which had previously been identified, substantiating the accuracy of the genotype calling in the 1000 bull genomes project data set, in this case for a variant of very low frequency.

Identification of a SMC2 mutation causing embryonic loss. Fertility is a major component of efficiency in both the beef and dairy production systems (Hegarty *et al.*, 2010; Hayes *et al.*, 2013). We investigated whether the 1000 bull genomes project data could be used to discover the causative mutation underlying the association of a chromosome region with embryonic lethality (VanRaden *et al.*,

2011). The Holstein Haplotype 3 (HH3) region was identified by the absence of homozygous individuals for a particular haplotype in US Holstein cattle in the interval from 94.0–96.5 Mb on chromosome 8 (BTA08). In a new data set of 47,878 French Holstein individuals, the frequency of the derived HH3 region was 2.5%; thus, 30 homozygous progeny were expected, whereas none were observed (χ^2 test P value of 4.6×10^{-6}). The effect on fertility (measured by calving rate) of matings between carrier bulls and daughters of carrier sires was estimated to be −5.40% in heifers and −5.54% in cows on the basis of the results of 5,281 at-risk inseminations of heifers and 7,315 at-risk inseminations of cows, respectively (where at-risk inseminations are those in which both the sire for the insemination and the maternal grandsire are carriers). These estimates are close to their expected values of 6.25% and 5%, assuming a recessive lethal mutation and given conception rates for Holstein heifers and cows of 50% and 40%, respectively. There was a single bull identified as a carrier of the derived HH3 region in the 1000 bull genomes project data set on the basis of the inferred haplotype. After filtering for mutations that were (i) carried in the heterozygous state by the HH3 carrier bull, (ii) absent in the 63 predicted non-carrier Holstein bulls, (iii) absent in the homozygous state in the Holstein bulls with unknown status and (iv) absent in the other breeds (assuming that the deleterious mutation is recent), only 1 candidate mutation was retained in the HH3 region: a thymine-to-cytosine transition at position 95,410,507 (g.95410507T>C). As this mutation was observed in only 1 bull among the 234 individuals included in the 1000 bull genomes project, it was genotyped by PCR and Sanger sequencing in a panel of 10 known HH3 carriers; all were heterozygous for the mutation, supporting the association between the HH3 region and the g.95410507C allele. As an additional test, 5,606 Holstein individuals were each genotyped in duplicate for the g.95410507T>C mutation using a custom Illumina BeadChip (all duplicate pairs were concordant). In agreement with the hypothesis that this mutation causes embryonic lethality, no individual with a CC genotype was detected (P < 0.06), whereas 2,476 individuals with the TT genotype and 171 individuals with the TC genotype were identified. In addition, 2,344 Montbeliarde individuals and 615 Normandy individuals were homozygous for the wild-type allele.

This g.95410507T>C polymorphism causes a substitution of a phenylalanine by a serine in the terminal part of structural maintenance of chromosome protein 2 (SMC2; p.Phe1135Ser). The SMC2 protein is a central component of the condensin I and II complexes that regulate chromosome condensation and stability during mitosis (Strunnikov *et al.*, 1995; Freeman *et al.*, 2000; Hudson *et al.*, 2003; Stray *et al.*, 2005; Vagnarelli *et al.*, 2006; Hudson *et al.*, 2009). Interestingly, deficiency in SMC2 homologs has been reported to produce unviable spores in *Saccharomyces cerevisiae* (Strunnikov *et al.*, 1995) and to cause embryonic lethality in *Arabidopsis thaliana* (Siddiqui *et al.*, 2003), *Caenorhabditis elegans* and *Drosophila melanogaster*.

In addition, the p.Phe1135Ser alteration was predicted to be highly damaging to SMC2 function by SIFT and PolyPhen-2. Finally, comparison of SMC2 protein sequences from different species showed a complete conservation of the Phe1135 residue among eukaryotes (all species undergoing mitosis), confirming that this amino acid is essential for the normal function of SMC2 (Supplementary Fig. 6). Taken together, these arguments support a causative role for the g.95410507T>C polymorphism in the HH3 haplotype.

Identification of a mutation causing lethal chondrodysplasia. We can also demonstrate that the 1000 bull genomes project data aid in the identification of dominant mutations, sequencing as few as one or two cases and then using the 1000 bull genomes project data set as controls. A small proportion (1%) of calves from a very widely used sire, Igale (Agerholm *et al.*, 2001), display a 'bulldog' phenotype—short neck and swollen cranium, depressed face with protruding tongue and cleft palate. The disease is a lethal form of chondrodysplasia. Recessive determinism of the disease was excluded for two reasons: (i) the 75 reported affected calves that were subsequently genotyped all carried heterogeneous paternal and maternal haplotypes (no runs of homozygosity) in the interval to which the defect has previously been mapped on BTA05 and (ii) 32 of the dams that produced an affected calf were again mated with Igale. All progeny were unaffected, which is extremely unlikely with recessive inheritance ($P < 1 \times 10^{-4}$). The hypothesis that the disease results from a dominant mutation but that the sire was mosaic for this mutation was pursued. To identify candidate causative mutations, we first filtered for heterozygous polymorphisms in the two affected calves that were (i) absent in the 1000 bull genomes project database (except for

Igale), as none of the other bulls sequenced had been reported to carry the syndrome and most had been extensively progeny tested, and (ii) predicted modified the amino acid sequence of a protein (Supplementary Table 3). This analysis yielded two candidate mutations shared by the affected calves, including one (g.32475732G>A on BTA05) affecting a functional candidate gene, *COL2A1*. Mutations in *COL2A1*, encoding the α1 chain of type II collagen, have been reported to cause a wide spectrum of skeletal disorders in humans, including achondrogenesis type II (ACG2; MIM 200610) (Godfrey *et al.*, 1988; Vissing *et al.*, 1989; Bonaventure *et al.*, 1995; Mortler *et al.*, 1995; Körkkö *et al.*, 2000), which shares strong similarities with the disease features observed in the affected progeny of Igale (Agerholm *et al.*, 2004). Both diseases are characterized by short limbs, short ribs, and absent or abnormal ossification of some skeletal parts. The g.32475732G>A mutation in *COL2A1* is predicted to cause substitution of a glycine residue with arginine (COL2A1 p.Gly960Arg). Glycine at this position is conserved across vertebrates (Supplementary Fig. 7). This substitution, like most of the alterations responsible for ACG2, disrupts the invariant GXY structural motif (with any amino acid at the second and third positions) necessary for perfect triple-helix formation and could thus lead to extensive overmodification, intracellular retention and reduced secretion of type II collagen, as previously established in the human form of the disease (Vissing *et al.*, 1989). Genotyping by PCR and RFLP of the g.32475742G>A mutation in ten additional affected calves showed perfect association between this mutation and the syndrome and suggested mosaicism in the Igale germ line, given the small proportion of affected calves and the fact that affected calves were heterozygous at this position. Finally, Sanger sequencing of the products from conventional PCR and nested PCR performed after PCR and RFLP definitively confirmed mosaicism for Igale at the locus (Supplementary Fig. 7). The identification of this mutation demonstrates the value of the 1000 bull genomes project database in the identification of dominant mutations found in *Bos taurus* cattle breeds, in this case, as a control measure to identify unaffected cattle with only limited additional whole-genome sequencing of two affected individuals.

2.2.2 Accuracy of imputing sequence variants

The accuracy of imputing the sequence variants into larger data sets genotyped with SNP arrays for subsequent GWAS was assessed using fivefold cross-validation in the resequenced Holstein individuals and using BEAGLE (Browning and Browning, 2009). In each fold, 23 sequenced cattle were randomly selected. For chromosome 29, the genotypes for these cattle were reduced to variants on the BovineHD array (13,320 SNPs). Genotypes for the selected cattle were imputed for all sequence variants on chromosome 29 (615,232), with the remaining Holstein individuals used as the reference. The accuracy of imputation (correlation of imputation probability and real genotypes) was reasonably high (Fig. 1a). Our cross-validation procedure did require that SNPs be polymorphic in both the reference and validation sets, which resulted in the exclusion of some rare SNPs and might inflate mean accuracy. Imputation accuracy varied across the chromosome (Figure 2.1a), especially in regions where there were few SNPs on the BovineHD array or errors existed in the genome assembly (Figure 2.1b). The accuracy of imputation decreased rapidly when MAF was below 0.1, suggesting that more sequenced individuals are required to accurately impute rare variants.

The imputation accuracy of using all breeds in the reference set was investigated with the same cross-validation procedure in 118 Holstein, 15 Jersey and 43 Fleckvieh individuals. Higher accuracy of imputation was achieved using a reference population with all three breeds combined, especially for Jersey individuals (Figure 2.1c). The accuracy was improved even further when IMPUTE2 (Howie *et al.*, 2009) was used instead of BEAGLE (Figure 2.1c), perhaps because the haplotype clustering procedure in BEAGLE obscures differences between haplotypes.

We also investigated how well genotypes for a well-described causative mutation in *DGAT1* affecting milk production (Grisart *et al.*, 2004) (included in the 1000 bull genomes project variants) could be imputed into a set of 2,253 German Holstein and 157 Australian Holstein bulls genotyped on dense SNP arrays and specifically for the *DGAT1* mutation using PCR. The imputed genotypes were the same as the real genotypes at the *DGAT1* mutation in 99.8% and 98% of the bulls in the German and Australian populations, respectively.

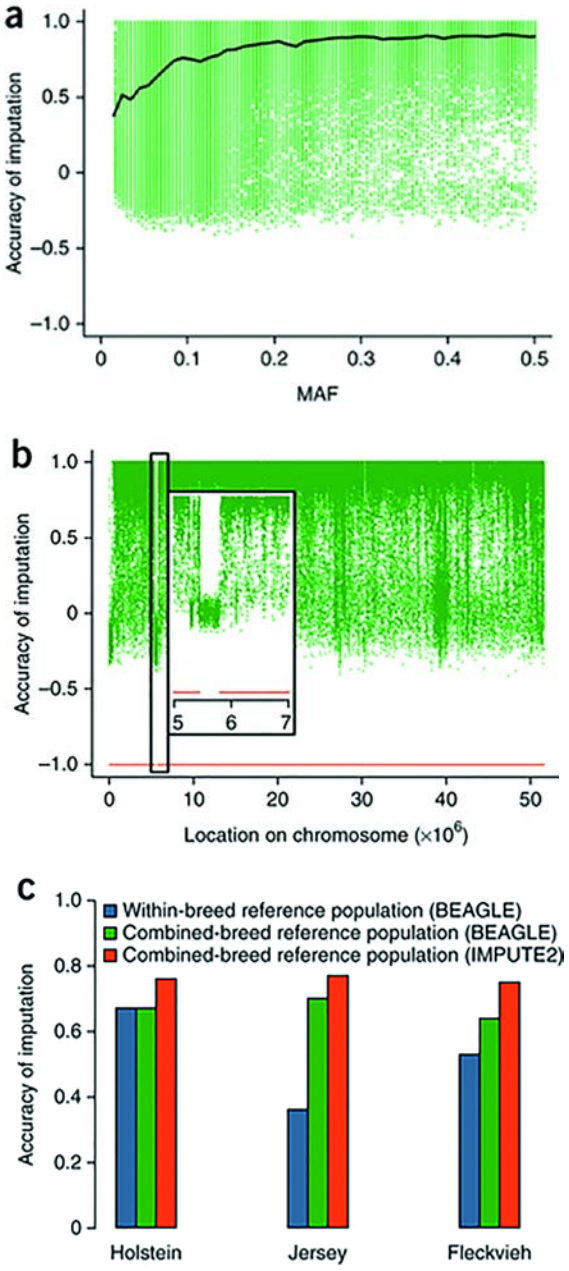

Figure 2.1 Accuracy of imputing sequence variants. **(a)** MAFs of sequenced SNPs versus the accuracy of imputation from the BovineHD panel into whole-genome sequence data for BTA29 in 129 Holstein bulls using BEAGLE. The black line is the average accuracy (defined in the Online Methods). **(b)** Accuracy of imputation relative to position on the chromosome, with BovineHD SNPs shown in red. The inset (from 5–7 Mb) shows the greatly reduced accuracy of imputation using BEAGLE for sequence where there are a limited number of BovineHD SNPs or assembly errors. **(c)** Using all cattle, regardless of breed, gave higher accuracy of imputation than imputing only within a particular breed. There were 15 Jersey and 43 Fleckvieh bulls. The mean accuracy of imputation was slightly higher with IMPUTE2 than with BEAGLE.

A missense mutation in KRT27 is strongly associated with curly coat. Cattle with curly hair (Figure 2.2a,b) experience higher tick and parasite burden than those with short, straight hair, especially under pasture conditions (Martinez *et al.*, 2006; Gasparin *et al.*, 2007). Curly hair is a rare condition in Fleckvieh cattle, with less than 1% of individuals affected, and there is no evidence that this phenotype is present in the Holstein, Jersey and Angus breeds. We used the proportion of a bull's daughters with curly hair as a quantitative trait and performed a sequence-based association study with sequence variants from the 1000 bull genomes project imputed into a population of 3,222 Fleckvieh bulls. This analysis yielded two regions with statistically significant ($P < 1 \times 10^{-8}$) results within clusters of type I and type II keratin genes on BTA19 and BTA05, respectively (Supplementary Fig. 8). No strong candidate mutation was found for the BTA05 region. On BTA19, five variants had P values markedly lower than the others (P values for these five variants were between 7.0×10^{-72} and 4.5×10^{-74} in comparison to the sixth most significant P value of $>1 \times 10^{-68}$) and were not segregating in the Holstein, Jersey and Angus breeds (Supplementary Table 4). Functional annotation of these sequence variants identified a missense mutation in KRT27 (c.276C>G; p.Asn92Lys; g.41636961C>G on BTA19; ss699911276) with a predicted damaging consequence (SIFT annotation and PolyPhen-2; Adzhubei *et al.*, 2010). The four other polymorphisms were located in regions not conserved among mammals (Ensembl). Keratin 27 is part of the type I and type II keratin protein complexes that are involved in the formation of keratin intermediate filaments in the inner root sheath (IRS) of the hair (Porter *et al.*, 2004; Tanakaa *et al.*, 2007). A deletion in the coil 2 domain (or helix termination motif) of the corresponding protein causes a wavy coat in mice (Tanakaa *et al.*, 2007), a phenotype that resembles the curly coat phenotype in Fleckvieh cattle. Other hereditary hair disorders in mice, humans and dogs have been associated with mutation affecting IRS-specific keratins and, more precisely, with mutations affecting the coil domain (Runkel *et al.*, 2006; Cadieu *et al.*, 2009; Fujimoto *et al.*, 2010). Indeed, coil domains are essential for the coiled-coil assembly of the different keratins into filaments. Asn92 is precisely located within the coil1A domain (or helix initiation motif) of *KRT27* (Figure 2.2). Moreover, this residue is perfectly conserved in the keratin 27 proteins of mammals (haired animals) (Supplementary Fig. 9a) and is conserved even across the proteins

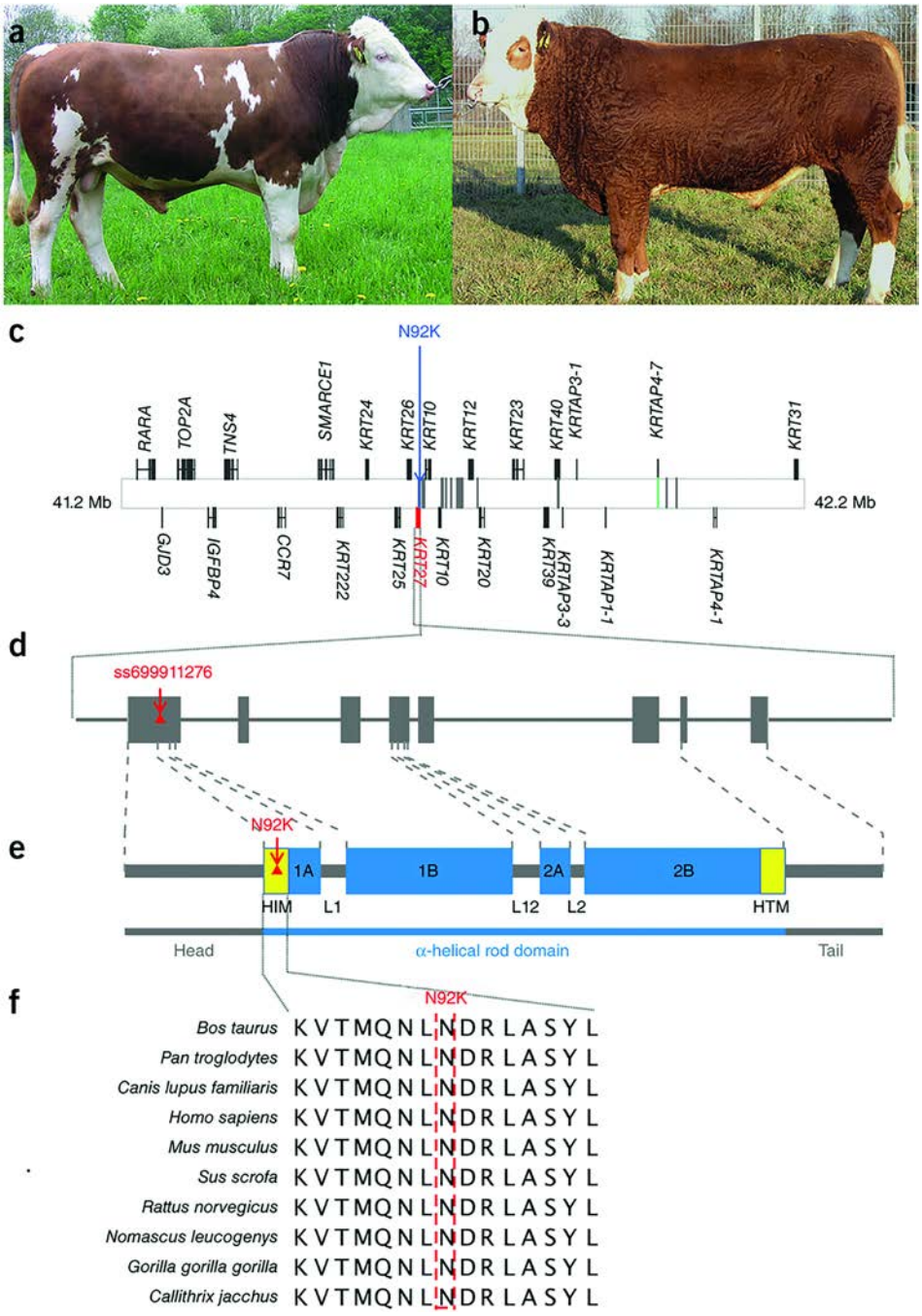
a
b
c
N92K
RARA
TOP2A
TNS4
SMARCE1
KRT24
KRT26
KRT10
KRT12
KRT23
KRT40
KRTAP3-1
KRTAP4-7
KRT31
41.2 Mb
42.2 Mb
GJD3
IGFBP4
CCR7
KRT222
KRT25
KRT27
KRT10
KRT20
KRT39
KRTAP3-3
KRTAP1-1
KRTAP4-1
d
ss699911276
e
N92K
1A
1B
2A
2B
HIM
L1
L12
L2
HTM
Head
α-helical rod domain
Tail
f
N92K
Bos taurus K V T M Q N L N D R L A S Y L
Pan troglodytes K V T M Q N L N D R L A S Y L
Canis lupus familiaris K V T M Q N L N D R L A S Y L
Homo sapiens K V T M Q N L N D R L A S Y L
Mus musculus K V T M Q N L N D R L A S Y L
Sus scrofa K V T M Q N L N D R L A S Y L
Rattus norvegicus K V T M Q N L N D R L A S Y L
Nomascus leucogenys K V T M Q N L N D R L A S Y L
Gorilla gorilla gorilla K V T M Q N L N D R L A S Y L
Callithrix jacchus K V T M Q N L N D R L A S Y L

Figure 2.2 Sequence-based fine-mapping of a locus underlying curly hair in 2,253 Fleckvieh bulls. **(a,b)** The hair of most Fleckvieh individuals is short and straight **(a)**, whereas only a small proportion of Fleckvieh individuals have curly hair **(b)**. Photographs were kindly supplied by Bayern-Genetik. **(c)** The sequence-based association study yielded strong association near a cluster of type I keratin genes on chromosome 19. Vertical bars represent 25 SNPs with $P < 1 \times 10^{-55}$, and gray, green and blue bars represent noncoding, synonymous and nonsynonymous variants, respectively. **(d)** Genomic structure of *KRT27*. Gray boxes represent exons. The nonsynonymous ss699911276 polymorphism is located in the first exon of *KRT27*. **(e)** Domain information for the IRS-specific keratin encoded by *KRT27* was obtained from the Human Intermediate Filament Database. Blue boxes represent the α-helical rod domain (1A, 1B, 2A, 2B) interrupted by three nonhelical linkers (L1, L12, L2) and flanked by nonhelical head and tail domains. The helix initiation and termination motifs (HIM and HTM) of the α-helical rod domain are located at the N and C termini (yellow boxes). **(f)** The codon for p.Asn92Lys resides within the highly conserved helix initiation motif

encoded by *KRT27* orthologs in cattle and humans (Supplementary Fig. 9b), with all of these data supporting a functional role for this amino acid.

Further, an association analysis conditional on this candidate mutation (ss699911276) did not identify any additional SNP as being associated with curly hair on BTA19 (Supplementary Fig. 10).

The curly coat phenotype also segregates in a unique paternal pedigree in the Montbeliarde breed. The Montbeliarde and Fleckvieh breeds both originate from the Swiss Simmental breed and diverged approximately 150 years ago (~30 generations ago). Five curly-haired Montbeliarde bulls from this pedigree were genotyped and found to be heterozygous for the *KRT27* mutation encoding p.Asn92Lys, but they were heterozygous for the four other top markers on BTA19 as well. In a subsequent analysis, 19 straight-haired Montbeliarde bulls with important individual contribution to the actual genetic pool of the breed (1–14.8%) and 72 individuals from 9 different straight-haired breeds were genotyped for the same 5 polymorphisms (Supplementary Table 4). Among these variants, only the mutation encoding p.Asn92Lys was never found in straight-haired cattle. Interestingly, sequencing of one (straight-haired) Rouge des Prés individual (Maine-Anjou) showed that this animal was even heterozygous for the four other polymorphisms but not for the mutation encoding p.Asn92Lys, suggesting a recent occurrence of this mutation in the Simmental lineage. This individual was

homozygous for one intergenic SNP in the region; however, this SNP was considerably less significant in the association testing with the curly coat phenotype. Finally, genotyping of six tagging SNP also showed that the curly-haired Montbeliarde bulls were homozygous for the wild-type allele at the other genomic region with very strong association with the curly coat phenotype—the BTA05 locus. This finding suggests that the mutant BTA19 allele is sufficient to cause curly hair in the heterozygous state, as observed in young mice heterozygous for a deletion in the coil 2 domain encoded by *KRT27* (Tanakaa *et al.*, 2007), making the mutation encoding p.Asn92Lys the prime candidate for the underlying mutation in the BTA19 quantitative trait locus (QTL).

A candidate mutation affecting early-lactation milk fat content. A total of 3,513 Fleckvieh bulls and 2,327 Holstein bulls had phenotypes for early-lactation milk fat content based on daughter averages. Breed-specific sequence-based association studies with 17,640,970 and 18,993,266 imputed sequence variants (segregating in the Fleckvieh and Holstein breeds, respectively) detected 6 QTL regions (Figure 2.3a,b). Among these, two QTLs on BTA14 and BTA27 near *DGAT1* and *AGPAT6*, respectively, were very highly associated in both breeds.

The causal mutation for the milk fat QTL on BTA14 was a previously reported variation in *DGAT1* encoding a lysine-to-alanine substitution (Grisart *et al.*, 2002) (p.Ala232Lys; chr. 14 :1,802,265–1,802,266; Winter *et al.*, 2002) and could therefore be used as a proof of concept of association of complex traits and imputed sequence data. The polymorphism encoding p.Ala232Lys was among the top association signals in the Holstein breed (Supplementary Fig. 11); however, it was not the variant with the lowest *P* value. This might be the result of both imperfect imputation and sampling error. In the case of the Holstein breed, the accuracy of imputation for the polymorphism encoding p.Ala232Lys was 99.8%, as described above (Supplementary Table 5). Hence, the association study results based on imputed sequence should be interpreted with caution: variants with higher P values than the most significant association should not be excluded as potential causative mutations, particularly if there is strong functional evidence for them. In the Fleckvieh breed, imputation accuracy for the polymorphism encoding p.Ala232Lys was considerably lower than in the Holstein breed, most likely as a result of the low frequency of the allele encoding lysine in the Fleckvieh breed and a lower number of sequenced Fleckvieh individuals (Supplementary Fig. 11 and Supplementary

Table 5). Further, the allele for lysine is present on two different haplotypes in the Fleckvieh breed (Winter *et al.*, 2002), which complicates imputation when the reference population is small—imputation accuracy would be higher with more individuals (Supplementary Fig. 12).

The MAFs of the BTA27 QTL (obtained from the most significantly associated BeadChip SNP) were 0.35 and 0.39 for the Holstein and Fleckvieh breeds, respectively, which should result in high imputation accuracy (for example, see Figure 2.1). The top association signals suggested *AGPAT6* as the underlying gene (Figure 2.3c,d). *AGPAT6* is a good functional candidate gene, as its transcription is highly correlated with the concentration of diacylglycerols and triacylglycerols in milk and varies across lactation stages, with a maximum in early lactation (Bionaz and Loor, 2008). Only two variants in the coding region of *AGPAT6* were polymorphic in both breeds: both were synonymous and neither was significantly associated (Supplementary Table 6), so coding variants in *AGPAT6* are not likely to explain the QTL on BTA27. Although most variants were associated in one breed only, we considered further four variants highly associated ($P < 1 \times 10^{-16}$) in both breeds as candidate causal polymorphisms (Figure 2.3e). These four variants included three SNPs (at 36,209,319 bp, 36,211,258 bp and 36,211,708 bp) and one indel (at 36,211,252 bp), all located in the promoter region of *AGPAT6* (Figure 2.3f,g). Among these variants, the indel polymorphism occurred at a site with a high probability that it is within a transcription factor binding site (Supplementary Fig. 13) and was strongly associated with overall milk fat content in the German Holstein population (Wang *et al.*, 2012). It is therefore a plausible causative mutation. Further, when the milk fat phenotypes were conditioned on the effect of this mutation, the association signal in both breeds was absent (Supplementary Fig. 14).

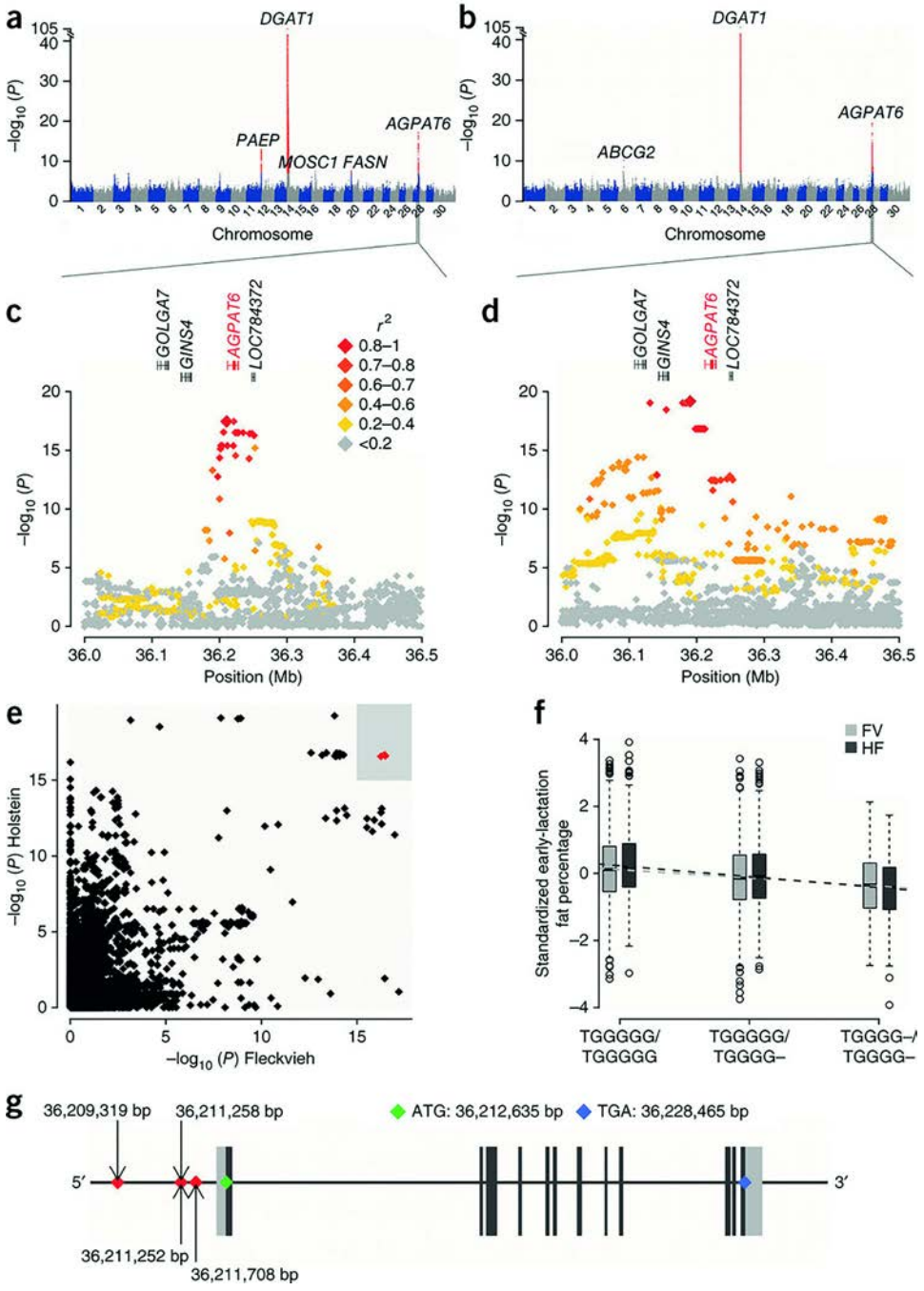
a
b
DGAT1
AGPAT6
PAEP
MOSC1 FASN
ABCG2
−log10 (P)
Chromosome
c
d
GOLGA7
GINS4
AGPAT6
LOC784372
r2
0.8–1
0.7–0.8
0.6–0.7
0.4–0.6
0.2–0.4
<0.2
Position (Mb)
e
−log10 (P) Holstein
−log10 (P) Fleckvieh
f
FV
HF
Standardized early-lactation fat percentage
TGGGGG/ TGGGGG
TGGGGG/ TGGGG–
TGGGG–/ TGGGG–
g
36,209,319 bp
36,211,258 bp
ATG: 36,212,635 bp
TGA: 36,228,465 bp
5′
3′
36,211,252 bp
36,211,708 bp

Figure 2.3 Sequence based association study for early-lactation milk fat percentage. **(a,b)** Manhattan plots show the association of 17,640,970 and 18,993,266 imputed variants with early-lactation milk fat percentage in 3,513 Fleckvieh bulls **(a)** and 2,327 Holstein bulls **(b)**. Red dots represent variants with association $\boldsymbol{P} < 7.7 \times 10^{-8}$. **(c,d)** Detailed overview of the associated QTL region on BTA27. Association of 3,507 and 3,954 imputed variants in 3,513 Fleckvieh **(c)** and 2,327 Holstein **(d)** bulls, respectively. Plots show the association of variants with early-lactation milk fat percentage according to their chromosomal position. Different colors represent the linkage disequilibrium ($\boldsymbol{r^2}$) between the top variant and all remaining variants. Red color highlights the putative functional candidate gene *AGPAT6*. **(e)** Comparison of $\boldsymbol{P}$ values of the imputed variants in the Fleckvieh and Holstein breeds. $\boldsymbol{P}$ values of variants segregating in one breed only are fixed to 1 for the other breed. Red symbols represent four variants with highly significant ($\boldsymbol{P} < 1 \times 10^{-16}$) association in both breeds, among them the candidate causal indel polymorphism (at 36,211,252 bp). **(f)** The allele frequency of the deletion is 0.25 and 0.38 in the Fleckvieh (FV) and Holstein (HF) breeds, respectively. The deletion is associated with lower early-lactation milk fat percentage in both breeds. The median fat percentage values for the wild-type/wild-type, wild-type/deletion and deletion/deletion genotypes were 0.11 (interquartile range, IQR = 1.35, n = 1,949), −0.15 (IQR = 1.33, n = 1,345) and −0.35 (IQR = 1.34, n = 219) in Fleckvieh and 0.18 (IQR = 1.29, n = 893), −0.09 (IQR = 1.31, n = 1,096) and −0.42 (IQR = 1.26, n = 335) in Holstein, respectively. **(g)** Gene structure of *AGPAT6*. Dark and light gray boxes represent the exons and UTRs, respectively. The four variants with highly significant associations in both breeds (red symbols) are located in the promoter of *AGPAT6*. The candidate causal indel polymorphism is located 1,383 bp upstream of the transcription start site of *AGPAT6*.

2.3 Discussion

This first phase of the 1000 bull genomes project identified 28.3 million variants. This large number of variants is perhaps surprising given the recent reduction in effective population size in the B. taurus individuals sequenced here (The Bovine HapMap Consortium, 2009). The average number of heterozygous sites per individual in our data was 1.44 per kilobase, which is higher than has been found in humans (1.03 heterozygous sites/kb for the Yoruba population and 0.68 heterozygous sites/kb for the European population; The 1000 Genomes Project Consortium, 2010). However, the rate of polymorphism is considerably lower than that reported in wild *D. melanogaster*—a recent study detected 4,672,297 SNPs in the comparison of 168 inbred lines sampled from the wild (Mackay *et al.*, 2012), in a

genome one-twentieth the size of the cattle genome. One likely explanation for the high rate of heterozygous sites per individual observed in our analysis is a large past effective population size. This would have allowed the accumulation of large amounts of variation, which has been retained between breeds or even between and within individuals of the same breed (Supplementary Fig. 3). Reconstructions of the *B. taurus* population history from both linkage disequilibrium information and the proportion of polymorphisms still segregating since the common ancestor of bateng and yak do give very large estimates of predomestication effective population sizes, in the order of 1×10^5 (Finlay *et al.*, 2007; MacEachern *et al.*, 2009; MacLeod *et al.*, 2009). Further, many of our sequenced bulls were key ancestors that predate the most recent contraction in effective population size.

The emphasis on health and welfare in cattle breeding programs is increasing (Boichard and Brochard, 2012). Thus, rapid identification and management of genetic defects that compromise health and welfare are necessary. We have demonstrated that the 1000 bull genomes project enables the very rapid identification of such mutations. This information can immediately be deployed in breeding programs to reduce or eliminate such diseases and to improve the efficiency of beef and milk production (for example, through improved fertility, as in the case of the embryonic lethal recessive mutation in *SMC2*). Further, the rapid identification of the mutations underlying sporadic syndromes that also occur in humans increases the attractiveness of cattle as models for such diseases in the postgenomic era.

Genomic selection (Meuwissen and Goddard, 2010b) is already being applied in cattle, but characterization of causal sequence variants will make genomic selection much more robust than the use of markers. In the future, millions of cattle worldwide will be measured for complex traits and genotyped for SNP panels. The pedigree structure of cattle populations makes it possible to impute a full genome sequence for each of these genotyped individuals and to thus provide a discovery population with millions of animals, each with a genome sequence.

2.4 Methods

2.4.1 Samples, resequenced read filtering and alignment.

A total of 234 individuals were resequenced with Illumina technology, where 129 were Holstein (125 Black and White and 4 Red and White) of North American, European and Australian origin, 43 were German Fleckvieh, 47 were Australian Angus and 15 were Jersey of Australian or US origin. All cattle were male, except one Holstein and one Fleckvieh individual. In the Holstein, Fleckvieh and Jersey breeds, key ancestors were resequenced. The Angus cattle were part of an Australian selection study. Many of the Holstein bulls for sequencing were selected with algorithms described by Druet *et al.* (2014), on the basis of their average and marginal contributions to the genetic diversity of the population in a pedigree analysis. Unless otherwise stated, DNA for the bulls in this study were extracted from commercial artificial insemination bull semen straws. Phenotypes were assessed during progeny evaluation by breeding organizations. For the bulldog calf study, samples were collected in 2000. The experiment complied with the French National Institute for Agricultural Research (INRA) ethical guidelines at that time. DNA for the Angus individuals was collected under a protocol approved by the New South Wales Department of Primary Industries Animal Ethics Committee.

Raw resequencing reads were filtered on the basis of chastity score and trimmed on the basis of quality score. Fastq files were then aligned to the reference bovine genome assembly UMD3.1 using BWA (Li and Durbin, 2009).

2.4.2 Variant calling and filtering.

Variants were called using SAMtools 0.1.18 mpileup (Li *et al.*, 2009). BAM files were pooled for variant calling. Variants were then removed if they had two or more alternative alleles, no observations of the alternative allele on either the forward or reverse reads, an overall quality (QUAL) score of <20, a mapping quality (MQ) score of <30, a read depth of <10 or more than the median plus 3 s.d. read depth, >0.1 opposing homozygotes or the same base-pair position (for example, a SNP overlapping an indel). In addition, we applied the following proximity filters: when an indel was within 10 bp of another indel, the indel with the lower QUAL score was removed, and, when any variant was within 3 bp of another variant, the variant with the lower QUAL score was removed. The filters were implemented by extending the Python VCF file parser PyVCF.

Phred score genotype probabilities were converted into true probabilities, and BEAGLE (Browning and Browning, 2009) was used to correct genotype calls. Concordance of resequenced genotypes with BovineHD chip genotypes was calculated as the proportion of identical genotypes before and after using BEAGLE. Opposing homozygotes were calculated as the proportion of non-matching homozygotes in parent-offspring pairs and were collected per pair and per locus.

2.4.3 Variant annotation.

SNPs and indels were annotated with predicted functional consequences using NGS-SNP (Grant *et al.*, 2011). Overlapping genes, transcripts, proteins, protein domains and variants were included among the annotations, along with any known pathways or phenotypes linked to the genes in cattle or to their orthologs in humans. Sequence alignments were constructed between proteins altered by missense SNPs and the proteins encoded by the orthologous genes and were scored using the SIFT algorithm to predict the functional impact of the protein substitutions (Kumar *et al.*, 2009). Variants were classified as 'known' if the non-reference allele was present in the dbSNP database and as 'novel' otherwise. The source databases used by NGS-SNP during annotation included Ensembl release 68 (Flicek *et al.*, 2012), dbSNP Build 133 (NCBI Resource Cooirdinators, 2013), Entrez Gene (NCBI Resource Cooirdinators, 2013) and UniProt release 2012_09 (UniProt Consortium, 2011). Custom scripts were used to further characterize the annotated variants and to compare the affected genes with those listed in the Online Mendelian Inheritance in Animals database.

2.4.4 Identification of a candidate causal mutation for embryonic loss.

This study included 47,878 French Holstein individuals genotyped with the Illumina BovineSNP50 BeadChip. Phasing and imputation of missing genotypes was performed with DualPhase (Druet and Georges, 2010). For each haplotype k with a frequency of >1%, the number of observed homozygous progeny $O(k)$ was compared to its expectation $E(k)$ under neutrality with a χ^2 test, using the following equation for $E(k)$:

$$E(k) = \sum_{i=1}^{n_s} p_{ik} \sum_{j=1}^{n_{mgs}} 0.5\left[q_{jk} + f_k\right] n_{ij}$$

with n_s being the number of sires with at least one copy of the haplotype, n_{mgs} being the number of maternal grandsires, f_k being the frequency of haplotype k,

p_{ik} being the transmission probability of haplotype k by sire i to his progeny (0.5 or 1), q_{jk} being the transmission probability of haplotype k by grandsire j to his daughter (0, 0.5 or 1) and n_{ij} being the number of progeny with sire i and maternal grandsire j.

Each bull was predicted to be a HH3 carrier or non-carrier according to its haplotype in the HH3 region, assuming complete association. No homozygous bull was found. The effect of the HH3 genotype on fertility was assessed by comparing the calving rate after insemination among sire–maternal grandsire genotype combinations. If the assumption of a recessive lethal defect is fulfilled with complete linkage disequilibrium, a reduction of cr/8 (where cr is the mean calving rate) is expected—corresponding to approximately 5% in lactating cows and 6% in heifers. Sixty-four bulls genotyped in France also had whole-genome sequence data available in the 1000 bull genomes project database. One was a heterozygous carrier, whereas the 63 others were predicted to be non-carriers. All variants in the HH3 region were then investigated for causation using the following criteria: (i) concordance with the predicted genotype of each bull, (ii) absence of the mutated allele in individuals from the other breeds and (iii) absence of individuals homozygous for the mutated allele among the Holstein individuals with no predicted status for HH3. The effects of the missense mutation on protein structure were predicted using both SIFT and PolyPhen-2 software (Kumar *et al.*, 2009; Adzhubei *et al.*, 2010). Across-species conservation of the peptide sequence at the affected residue was also taken into account. Protein sequences were retrieved from Ensembl and aligned using ClustalW version 2.0.1 software (Thompson *et al.*, 2002). Finally, the potential effects of the selected mutations were assessed using the literature (animal models and proteins of known function).

Because the g.95410507T>C polymorphism was observed in whole-genome sequence data from only 1 bull among the 234 individuals in the 1000 bull genomes project, this mutation was genotyped by PCR and Sanger sequencing on a panel of 10 HH3 carriers for confirmation. PCR primers (Supplementary Table 7) were designed using the UMD3.1 bovine genome assembly with Primer3 software (Rozen and Skaletsky, 2000). PCR was performed using the Go-Taq Flexi system (Promega) according to the manufacturer's instructions on a Mastercycler pro thermocycler (Eppendorf). The resulting amplicons were purified and bidirectionally

sequenced by Eurofins MWG using conventional Sanger sequencing. Polymorphisms were detected with NovoSNP software (Weckx *et al.*, 2005).

As an additional control, 5,606 individuals (2,647 Holstein and 2,344 Montbeliarde and Normandy) were genotyped in duplicate for the g.95410507T>C mutation on BTA08 using the Illumina EuroG10k BeadChip, a custom SNP array developed for genomic selection purposes in the framework of the Eurogenomics Consortium.

2.4.5 Identification of a dominant mutation resulting in bulldog calf syndrome.

DNA from 12 bulldog calves and their 10 dams and Igale Masc was extracted from muscle and blood samples and from semen straws, respectively, using a standard phenol-chloroform method. A paired-end library with a 250-bp insert size was generated for two bulldog calves using the Illumina TruSeq DNA Sample Prep kit. Libraries were quantified using the QPCRLibrary Quantification kit (Agilent Technologies), quality controlled on a High-Sensitivity DNA Chip (Agilent Technologies) and sequenced on one HiSeq 2000 lane (Illumina), each with the Illumina TruSeq V3 kit (200 cycles). Mapping of reads and identification of polymorphisms was performed as previously reported. Polymorphisms found in regions with coverage above an arbitrary threshold of 50× were removed to avoid artifacts (as these regions are likely to be repeats). A 341-bp fragment spanning the g.32475732G>A candidate mutation on BTA05 was PCR amplified (primers listed in Supplementary Table 7) using Go-Taq Flexi DNA polymerase (Promega) according to the manufacturer's instructions. The resulting amplicon was bidirectionally sequenced by Qiagen using conventional Sanger sequencing, and polymorphism was detected with NovoSNP software (Weckx *et al.*, 2005). The candidate mutation was subsequently genotyped by digesting the same PCR product with BpmI endonuclease (New England BioLabs) according to the manufacturer's instructions. Whereas digestion of the wild-type allele is predicted to produce 166- and 175-bp fragments, the mutant allele, with disruption of a BmpI restriction site, is predicted to not be digested. After digestion, products were separated on a 2% agarose gel by electrophoresis and stained with ethidium bromide for visualization. To confirm mosaicism, a second PCR-RFLP analysis was performed. Products were separated by 2% NuSieve GTGTM agarose (FMC) gel electrophoresis, and the unrestricted fragment for Igale was recovered from the agarose plug using the Wizard SV Gel

and PCR Clean-up System protocol (Promega). Nested PCR (primers listed in Supplementary Table 7) was performed on the purified fragment using Go-Taq Flexi DNA polymerase (Promega) according to the manufacturer's instructions. Finally, the resulting amplicon was sequenced using Sanger sequencing and analyzed as previously described.

2.4.6 Imputation of sequence variants.

Imputation accuracy for a SNP on BTA29 was assessed using 5-fold cross-validation in 114 sequenced Holstein individuals. In the individuals to be imputed, all sequenced genotypes were set to missing, except for the markers that were on the Illumina BovineHD array. Imputation was performed using BEAGLE 3.3.2 (Browning and Browning, 2009) with default parameter settings. The genotypes for each individual were assumed to be unphased, and no relationships between individuals were used. Imputation accuracy (r) was calculated per SNP by the correlation between the observed and imputed genotypes. The correlation is calculated as:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

where x_i is the observed genotype for individual i and y_i is the estimated B allele dosage for individual i. In the case where no variation was found in the observed genotypes or in the estimated B allele dosage for one or more of the validation sets, the imputation accuracy was set to missing. The MAF for all SNPs was calculated using data for the 114 Holstein individuals.

The second study was performed using 118 Holstein (including 4 red Holstein), 15 Jersey and 43 Fleckvieh individuals, which were randomly divided into 5 cross-validation sets within each breed. The masking procedure and markers on BTA29 described above were used. Imputation using a single breed as the reference was performed using BEAGLE 3.3.2 with standard settings, and imputation using a combined-breed reference was performed using both BEAGLE and IMPUTE2 2.3.0 with N_e set to 100. For both methods, analysis was carried out using the unphased most likely genotypes from sequence data. For the combined-breed setups, the same cross-validation groups were used, but accuracy was calculated within breeds using the formula above.

2.4.7 Identification of candidate mutations for curly coat in the Fleckvieh and Montbeliarde breeds.

Fleckvieh cows with curly hair were identified during regular examination of first-crop daughters from artificial insemination. The total number of cows examined was 282,460, with an average of 87.7 daughters per bull. Because genotypes for females were not available, the proportion of curly daughters of a bull was used as a quantitative trait in a mixed model–based association study (as described for the mutation region associated with early-lactation milk fat content), with sequence variants from the 1000 bull genomes project imputed into a population of 3,222 bulls. Details on Fleckvieh sequence imputation are provided in the section on the mutation region associated with early-lactation milk fat content.

As a control, 5 curly-haired Montbeliarde bulls (1 grandsire, 2 sires and 2 sons), 19 straight-haired Montbeliarde founder bulls and 72 individuals from 9 different straight-haired breeds (8 Charolais, 8 Limousin, 8 Blond d'Aquitaine, 8 Vosgienne, 8 Abondance, 8 Rouge des Prés, 8 Normande and 8 Brown Swiss) were genotyped for the 5 markers with the lowest P value on BTA19 using PCR and Sanger sequencing. The 24 Montbeliarde individuals were also genotyped for 5 tagging SNPs associated with the curly allele at the BT05 locus. Details on the primers used are presented in Supplementary Table 7.

2.4.8 Identification of a candidate causal mutation for early-lactation milk fat content.

Sequence imputation in the Fleckvieh population. Of 28,272,001 filtered variants (SNPs and indels), 17,640,970 were polymorphic within the 43 sequenced Fleckvieh individuals. Data for these variants were extracted and subsequently phased with BEAGLE 3.2.1 (Browning and Browning, 2009). The resulting haplotypes were imputed for 4,563 Fleckvieh bulls (target population) using medium- and high-density genotype data. Within the target population, 3,292 bulls were genotyped with the Illumina BovineSNP50 BeadChip comprising 54,001 SNPs (medium density) and 1,271 bulls were genotyped with the Illumina BovineHD BeadChip comprising 777,962 SNPs (high density). After quality control (individual call rate > 95%, SNP call rate > 95%, MAF > 0.5%, no significant deviation from Hardy-Weinberg equilibrium), the medium- and high-density data sets included genotypes for 39,304 and 645,189 SNPs, respectively. In a first step, individuals with medium-density genotypes were imputed to high density using BEAGLE. Afterward,

the target population comprised 4,563 bulls with phased genotypes for 645,189 SNPs. In a second step, the phased sequence information was imputed in all target haplotypes with Minimac (Howie *et al.*, 2012). This prephasing-based approach yields good imputation accuracy in cattle, even if the number of reference haplotypes is small (Pausch *et al.*, 2013).

Genome-wide association study. We used EMMAX (Kang *et al.*, 2010) to fit the model $\boldsymbol{y} = \mu + Xb + Z\boldsymbol{u} + \boldsymbol{e}$, where $\boldsymbol{y}$ is a vector of phenotypes, μ is the mean, b is the additive allele substitution effect, X is a design matrix of allele dosages for the imputed variants, Z is a design matrix connecting phenotypes to animals, $\boldsymbol{u}$ is a vector of additive animal effects $\sim N(0, \sigma_a^2\, G)$, with σ_a^2 being the additive genetic variance and G being the genomic relationship matrix based on genotype information (VanRaden, 2008; Hayes *et al.*, 2009), and $\boldsymbol{e}$ is a vector of random residuals.

2.5 Appendix

Supplementary material can be found in the online version of the published paper or can directly be accessed via:

http://www.nature.com/ng/journal/v46/n8/extref/ng.3034-S1.pdf

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R. F. Brøndum *et al.* 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics 46: 858–865.

CHAPTER 3

ACCURACY OF IMPUTATION TO WHOLE-GENOME SEQUENCE DATA IN HOLSTEIN FRIESIAN CATTLE

RIANNE VAN BINSBERGEN[1, 2]
MARCO C.A.M. BINK[2]
MARIO P.L. CALUS[1]
FRED A. VAN EEUWIJK[2]
BEN J. HAYES[3]
INA HULSEGGE[1]
ROEL F. VEERKAMP[1]

[1] Animal Breeding and Genomics Centre, Wageningen Livestock Research, P.O. Box 338, 6700 AH Wageningen, the Netherlands
[2] Biometris, Wageningen University and Research Centre, P.O. Box 100, 6700 AC Wageningen, the Netherlands
[3] Biosciences Research Division, Department of Environment and Primary Industries, 1 Park Drive, Bundoora 3083, Australia

Genetics Selection Evolution (2014) 46:41

Abstract

Background: The use of whole-genome sequence data can lead to higher accuracy in genome-wide association studies and genomic predictions. However, to benefit from whole-genome sequence data, a large dataset of sequenced individuals is needed. Imputation from SNP panels, such as the Illumina BovineSNP50 BeadChip and Illumina BovineHD BeadChip, to whole-genome sequence data is an attractive and less expensive approach to obtain whole-genome sequence genotypes for a large number of individuals than sequencing all individuals. Our objective was to investigate accuracy of imputation from lower density SNP panels to whole-genome sequence data in a typical dataset for cattle.

Methods: Whole-genome sequence data of chromosome 1 (1737 471 SNPs) for 114 Holstein Friesian bulls were used. Beagle software was used for imputation from the BovineSNP50 (3132 SNPs) and BovineHD (40 492 SNPs) beadchips. Accuracy was calculated as the correlation between observed and imputed genotypes and assessed by five-fold cross-validation. Three scenarios S40, S60 and S80 with respectively 40%, 60%, and 80% of the individuals as reference individuals were investigated.

Results: Mean accuracies of imputation per SNP from the BovineHD panel to sequence data and from the BovineSNP50 panel to sequence data for scenarios S40 and S80 ranged from 0.77 to 0.83 and from 0.37 to 0.46, respectively. Stepwise imputation from the BovineSNP50 to BovineHD panel and then to sequence data for scenario S40 improved accuracy per SNP to 0.65 but it varied considerably between SNPs.

Conclusions: Accuracy of imputation to whole-genome sequence data was generally high for imputation from the BovineHD beadchip, but was low from the BovineSNP50 beadchip**.** Stepwise imputation from the BovineSNP50 to the BovineHD beadchip and then to sequence data substantially improved accuracy of imputation. SNPs with a low minor allele frequency were more difficult to impute correctly and the reliability of imputation varied more. Linkage disequilibrium between an imputed SNP and the SNP on the lower density panel, minor allele frequency of the imputed SNP and size of the reference group affected imputation reliability.

Keywords: genotype imputation, BovineSNP50, BovineHD, Beagle

3.1 Background

One advantage of using whole-genome sequence data over genotypes from SNP (single nucleotide polymorphisms) panels for genome-wide association studies (GWAS) and genomic prediction is that polymorphisms causing genetic differences can be included in whole-genome sequence data. Because the causative mutation is included, decay in linkage disequilibrium (LD) between a SNP and the causative mutation by recombination events is not an issue. Accordingly, testing variants directly associated with a given trait is possible and may lead to higher accuracy in GWAS and genomic predictions. Moreover, since there is no decay in LD when using sequence data compared to traditional smaller-sized marker panels, genomic selection across generations and across breeds may be improved (e.g. Meuwissen and Goddard, 2010; Li *et al.*, 2011b; Druet *et al.*, 2014).

Costs to generate whole-genome sequence data are decreasing rapidly. It is expected that, in the next few years, whole-genome sequence data will be widely available for crops and livestock, as is already the case for human studies (The 1000 Genomes Project Consortium, 2012). Despite the fact that costs of sequencing are decreasing, it is still expensive to sequence large numbers of individuals. A less expensive approach to produce sequence genotypes for a large number of individuals is to impute from lower density marker panels to whole-genome sequence data. In this case, a core set of individuals is fully sequenced, and the lower density genotypes of the remaining individuals will be imputed to whole-genome sequence genotypes using the sequenced individuals as reference (Marchini *et al.*, 2007; Browning and Browning, 2009; Howie *et al.*, 2009; Li *et al.*, 2010).

However, using sequence data may not lead to higher accuracy in genomic predictions and GWAS if the accuracy of imputation to sequence data is too low. Accuracy of imputation was studied in barley with 3200 SNPs (Iwata and Jannink, 2010), in maize with 35 000 SNPs (Hickey *et al.*, 2012), in sheep with 50 000 SNPs (Hayes *et al.*, 2012) and in cattle with 50 000 SNPs e.g. (Druet *et al.*, 2010) and 777 000 SNPs e.g. (VanRaden *et al.*, 2013), among others. The general tendency in those studies was that the accuracy of imputation increased with an increasing number of SNPs on the lower density marker panel, a decreasing distance between the imputed SNP and the nearest SNP on the lower density marker panel, an increasing

minor allele frequency (MAF) of imputed SNPs, an increasing level of LD (linkage disequilibrium), and an increasing number of close relatives between imputed and reference individuals. In all those studies, imputation was done from low-density panels to higher density panels but not to whole-genome sequence data.

In contrast to crops and livestock, human sequence data are available and accuracy of imputation to sequence data has been investigated (e.g. Fridley *et al.*, 2010; Li *et al.*, 2011a; Sung *et al.*, 2012), which showed that accuracy of imputation was influenced by reference group composition (e.g. size or populations included), number of markers on the lower density marker panel, and MAF of imputed SNPs. Moreover, according to Li *et al.* (2011a), these factors influenced accuracy of imputation especially in the case of SNPs with a MAF below 0.05. For imputation of SNPs with a MAF below 0.005, it was necessary that the reference group included at least 1200 individuals and for imputation of SNPs with a MAF between 0.005 and 0.05, only about 40% of the SNP genotypes were imputed with 1200 individuals in the reference group.

Crop and livestock populations differ from human populations, in extent of LD and population structure (Goddard and Hayes, 2009; The Bovine HapMap Consortium, 2009; Hamblin *et al.*, 2011). In cattle, effective population size of some individual breeds has decreased rapidly to about 100 due to intense selection (de Roos *et al.*, 2008; The Bovine HapMap Consortium, 2009; Qanbari *et al.*, 2010). Consequently, LD in cattle breeds extends on relatively long distances. This is also true for many other domestic animal and plant populations (e.g. dogs or barley), but not for human populations (Goddard and Hayes, 2009; Hamblin *et al.*, 2011). When using whole-genome sequence data, differences in extent of LD and population structure may affect imputation accuracies more in crop or livestock analyses than in human analyses.

The objective of this study was to investigate the accuracy of imputation of genotypes from SNP panels to whole-genome sequence data in a typical dataset of domestic animals and to gain insights on the factors that affect accuracy of imputation, such as number of sequenced individuals, number of SNPs on the lower density marker panel, location and MAF of the imputed SNPs. Because in practice true genotypes are unknown, it is important to understand the underlying factors that influence imputation accuracy. Holstein Friesian cattle data provided by the 1000 bull genomes project (Daetwyler *et al.*, 2014) was used in this study.

3.2 Methods

3.2.1 Genotypic data

Whole-genome sequence data of 114 Holstein Friesian bulls were provided by the 1000 bull genomes project (Run 2.0; Daetwyler *et al.*, 2014). Bulls that originated from Australia, Canada, Denmark, Finland, France, Germany, Sweden, The Netherlands, UK, and USA, were identified as key ancestors of the global Holstein Friesian population. Each bull was sequenced using Illumina HiSeq Systems (Illumina Inc., San Diego, CA). Alignment, variant calling, and quality controls were done in a multi-breed population with sequenced Holstein Friesian, Fleckvieh, Jersey, and Angus bulls as described by Daetwyler *et al.* (2014).Variants used in our study were SNPs and INDELs (both considered as SNPs here). Two alleles (A and B) per SNP were assumed with a value of 0, 1, or 2 for genotype AA, AB, or BB, respectively. To save computing time and space, only SNPs on *Bos taurus* autosome 1 (BTA1) were used. Similar results were expected for other chromosomes.

A set of sequence variants and genotypes that can be used to test imputation programs is available at request via www.1000bullgenomes.com .

3.2.2 Imputation

Beagle 3.3.2 software (Browning and Browning, 2009) with default parameter settings was used for imputation. No SNP edits were performed prior imputation. For each individual, the most likely genotypes were used and they were assumed to be unphased, for both the reference and validation sets. Moreover, it was assumed that all individuals were unrelated. Accuracy of imputation (*r*) was calculated as the correlation between observed and imputed genotypes. Imputed genotypes were assessed by estimated *B*-allele dosage, which had a value between 0 and 2 and was calculated using posterior genotype probabilities as estimated by Beagle: 0 * P(AA) + 1 * P(AB) + 2 * P(BB). SNPs with fixed observed genotypes or estimated *B*-allele dosages for one or more validation groups were removed. Accuracy of imputation ranged between -1 (opposite genotype imputed) and +1 (correct genotype imputed). An imputation accuracy with a value around 0 meant random imputation.

To assess imputation accuracy, five-fold cross validation was performed. Individuals were randomly divided in five groups, group 1 to 5, and each group was used as validation set once. For validation individuals, SNP genotypes for SNPs

corresponding to the Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, CA; 54 609 SNPs) or Illumina BovineHD BeadChip (Illumina Inc., San Diego, CA; 777 962 SNPs) were retained, while the remaining SNPs on the sequence panel were masked.

3.2.3 Scenarios

To study the effect of number of sequenced individuals on imputation accuracy, three scenarios were considered: S80, S60, and S40. Reference group in scenarios S80, S60 and S40 contained 80% (all, except validation individuals), 60% and 40% of the individuals, respectively. In scenarios S40 and S60, the two or three following groups were designated as reference group. For example for scenario S60, if individuals in group 1 were designated as validation individuals, then individuals in group 2, 3, and 4 were designated as reference individuals.

According to VanRaden *et al.* (2013), accuracy of imputation from 3K and 6K panels to the BovineHD beadchip was improved if the genotypes were imputed first to the BovineSNP50 and then to the BovineHD beadchip instead of directly to the BovineHD beadchip. To study if this stepwise imputation approach also improved accuracy of imputation from the BovineSNP50 beadchip to whole-genome sequence data, a stepwise imputation was studied in scenario S40. Individuals in the two following groups were reference individuals for imputation to the BovineHD beadchip (step 1) and individuals in the two previous groups were reference individuals for imputation to whole-genome sequence data (step 2). For example, if individuals in group 2 were designated as validation individuals, then individuals in group 3 and 4 were assigned to the reference group for step 1, and individuals in group 5 and 1 were assigned to the reference group for step 2.

3.2.4 Factors that affect imputation accuracy

Factors that can influence imputation accuracy per SNP are number of sequenced individuals, distance (in base pairs) and MAF difference between an imputed SNP and its nearest SNP on the lower density marker panel, and MAF of imputed SNPs. MAF was calculated for each SNP based on all 114 individuals. For graphical representation and to illustrate the average behavior of SNPs, SNPs were binned in groups of 1000 based on distance or MAF (difference), and these binned SNPs were used to study imputation reliability (r^2).

To investigate the relationship between imputation reliability for a SNP and the factors that may influence its value, a few simple functions were used. Although haplotypes (and not single SNPs) are used for imputation of missing SNPs, our first assumption was that imputation reliability is based on LD between known and unknown SNPs, and our second assumption was that MAF together with number of sequenced individuals will affect imputation reliability.

Two functions were used to model LD between two SNPs: one was based on distance (Sved, 1971) and one was based on difference in MAF (Miller, 2013). The first function describes LD decay ($r_{dist}{}^2$) based on effective population size (*Ne*) and distance of an imputed SNP to its nearest SNP on the lower density marker panel (*c*; in Morgan):

$$r_{dist}{}^2 = \frac{1}{4 * Ne * c + 1}$$

Ne was assumed to be equal to 100 or 1000 and for distances, it was assumed that 10^6 base-pairs (1 Mb) are equal to 1 centiMorgan (cM) (Gautier *et al.*, 2007; Kim and Kirkpatrick, 2009). The second function describes the general upper limit for LD ($r_{dMAF}{}^2$) based on difference in MAF between an imputed SNP and its nearest SNP on the lower density marker panel (dMAF; Miller, 2013):

$$r_{dMAF}{}^2 = \frac{1 - 4dMAF}{2dMAF + 1}$$

If two SNPs differ in MAF, LD between those SNPs is expected to be low (Lewontin, 1995; Mueller, 2004).

These two functions do not account for the MAF of imputed SNPs or number of reference individuals. With a low number of reference individuals, the probability that individuals carry the rare allele of a SNP with a low MAF is lower, thus increasing the number of reference individuals may increase imputation reliability of this SNP. To our knowledge, there is no theoretical function that describes the relationship between imputation reliability or LD and MAF of imputed SNPs or number of reference individuals. Therefore, an empirical function was derived by fitting a Michaelis-Menten function (Johnson and Goody, 2011) on the data:

$$r_{MAF}{}^2 = \frac{V_{max} * MAF}{K_m + MAF}$$

where $r_{MAF}{}^2$ is the imputation reliability, V_{max} is the estimate of the upper limit of $r_{MAF}{}^2$ and K_m is the deflection point, i.e. the estimated MAF when $r_{MAF}{}^2$ = $1/2V_{max}$. The Michaelis-Menten function is often used in studies on enzyme kinetics

that describe the rate of enzymatic reactions based on substrate concentration (Johnson and Goody, 2011). This function was chosen because of its simplicity (two meaningful parameters) and its agreement with the observed data (starting from 0, it increases rapidly at the beginning and asymptotically approaches its maximum).

The three functions mentioned each explain a part of the imputation reliability. For overall imputation reliability($r_{total}{}^2$) the functions were multiplied:

$r_{total}{}^2 = r_{dist}{}^2 * r_{dMAF}{}^2 * r_{MAF}{}^2$.

In the functions for $r_{dist}{}^2$ and $r_{dMAF}{}^2$, the nearest SNP on the lower density marker panel was used although it may not be the SNP that has the highest LD with the imputed SNP. To take this into account, for each SNP $r_{dist}{}^2 * r_{dMAF}{}^2$ was estimated for the five nearest SNPs on the lower density marker panel and, for each imputed SNP, SNPs on the lower density marker panel that had the highest value for $r_{dist}{}^2 * r_{dMAF}{}^2$ were selected. Next, the parameters V_{max} and K_m were estimated by fitting $r_{MAF}{}^2$. Finally, $r_{total}{}^2$ was calculated and imputed SNPs were grouped with 1000 SNPs into bins with similar values of $r_{total}{}^2$ and plotted against the observed r^2 from the sequence data.

3.3 Results

3.3.1 Whole-genome sequence data

BTA1 is the largest bovine chromosome and contains approximately 160.10^6 bp. In the current 1000 bull genomes dataset, 1737 471 SNPs (of which 5.5% were INDELs) were called on BTA1 based on a multi-breed population. Of these SNPs, 76.8% showed variation within the 114 Holstein Friesians. The BovineSNP50 and BovineHD panels contained respectively 3514 and46 499 SNPs on BTA1, however, not all these SNPs were found in the sequence data. For the BovineSNP50 panel, 3132 SNPs (0.18% of the SNPs in the sequence data) and the BovineHD panel, 40 492 SNPs (2.33% of the SNPs in the sequence data) were found in the sequence data. Figure 3.1 presents a Venn diagram of the numbers of SNPs on BTA1 in the two lower density marker panels and in the whole-genome sequence data and numbers of overlapping SNPs.

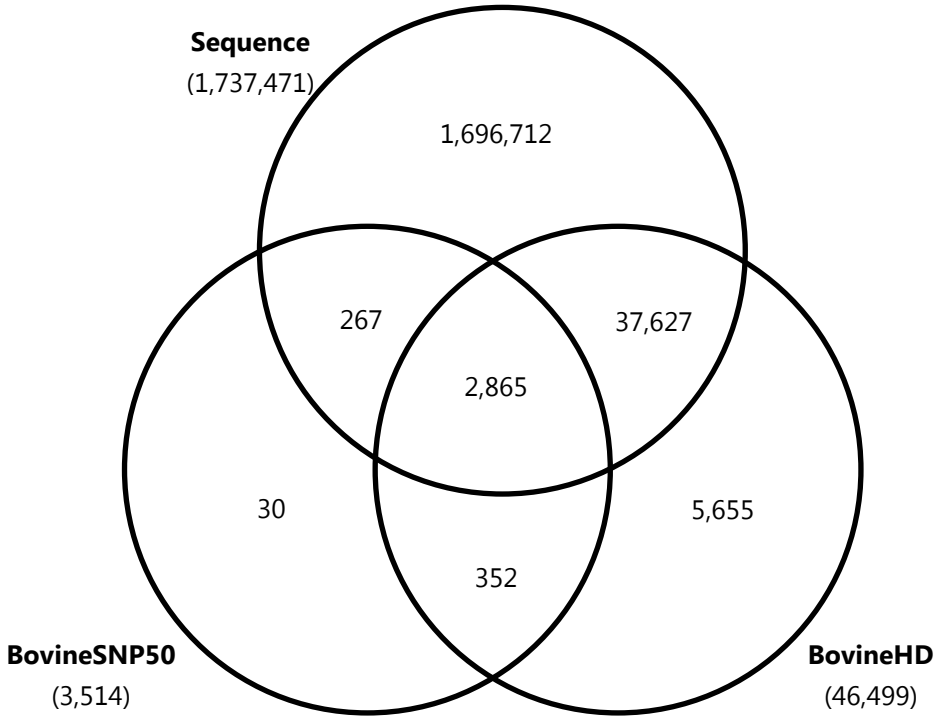

Figure 3.1 Number of SNPs on BTA 1. Venn diagram showing number of SNPs on BTA1 in the two lower density marker panels (BovineSNP50 and BovineHD) and in sequence data and overlapping numbers.

3.3.2 Accuracy of imputation

Mean accuracy of imputation per SNP was assessed by cross-validation. For imputation from the BovineSNP50 beadchip to sequence data, it ranged between 0.37 for scenario S40 and 0.46 for S80, and for imputation from the BovineHD beadchip to sequence data, it ranged between 0.77 for scenario S40 to 0.83 for S80 (Table 3.1). Standard deviations ranged from 0.36 to 0.37 for imputation from the BovineSNP50 beadchip, and from 0.27 to 0.29 for imputation from the BovineHD beadchip. In comparison to direct imputation from the BovineSNP50 beadchip to sequence data, stepwise imputation from the BovineSNP50 to the BovineHD beadchip and then to sequence data improved accuracy per SNP from 0.28 to 0.65 for scenario S40. However, it was still lower than the accuracy of imputation from

the BovineHD panel to sequence data (0.77). Accuracy per SNP for stepwise imputation was found to be similar to the product of imputation accuracies for the two steps.

Mean accuracy of imputation per individual was higher than mean accuracy per SNP. For imputation from the BovineSNP50 panel and from the BovineHD panel to sequence data, mean accuracies ranged from 0.78 for scenario S40 to 0.95 for S80, and from 0.93 for scenario S40 to 0.95 for S80, respectively (Table 3.2). Reasons for this difference are discussed below. For imputation from either of the lower density marker panels, standard deviation was 0.04 for all scenarios. As for accuracy per SNP, imputation accuracy per individual was improved with stepwise imputation from the BovineSNP50 beadchip to sequence data for scenario S40 and reached a value similar to the product of imputation accuracies of each step.

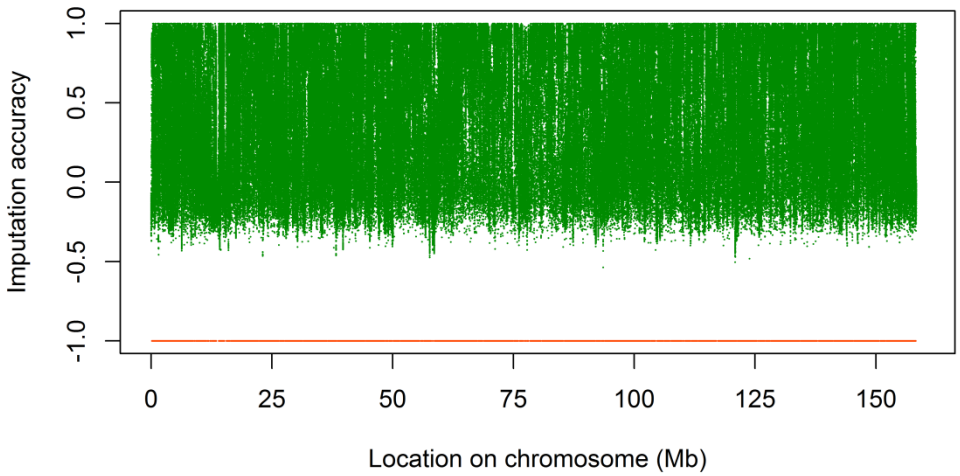

Figure 3.2 Accuracy of imputation from the BovineSNP50 beadchip on BTA1. Location on BTA1 versus accuracy of imputation from the BovineSNP50 beadchip to whole-genome sequence data for scenario S80; each green dot represents a SNP; orange dots at -1 are locations of SNPs of the BovineSNP50 beadchip

3.3.3 Factors that influence imputation accuracy

The range of variation for imputation accuracies per SNP was large (Table 3.1). In Figure 3.2 and Figure 3.3, this variation is illustrated for all SNPs on BTA1 for scenario S80. More SNPs had an accuracy above 0.5 for imputation from the BovineHD than from the BovineSNP50 beadchip. However, even with imputation from the BovineHD panel, SNPs from some regions of the genome were still

imputed with low accuracy. For example, around the position 75.10^3 Mb there is a region in which the distance between imputed SNPs and SNPs on the BovineHD panel is large and for which imputation was difficult (Figure 3.3B). This region contained SNPs that are on the BovineHD panel, but since they did not segregate in the sequence data, no genotypes were available.

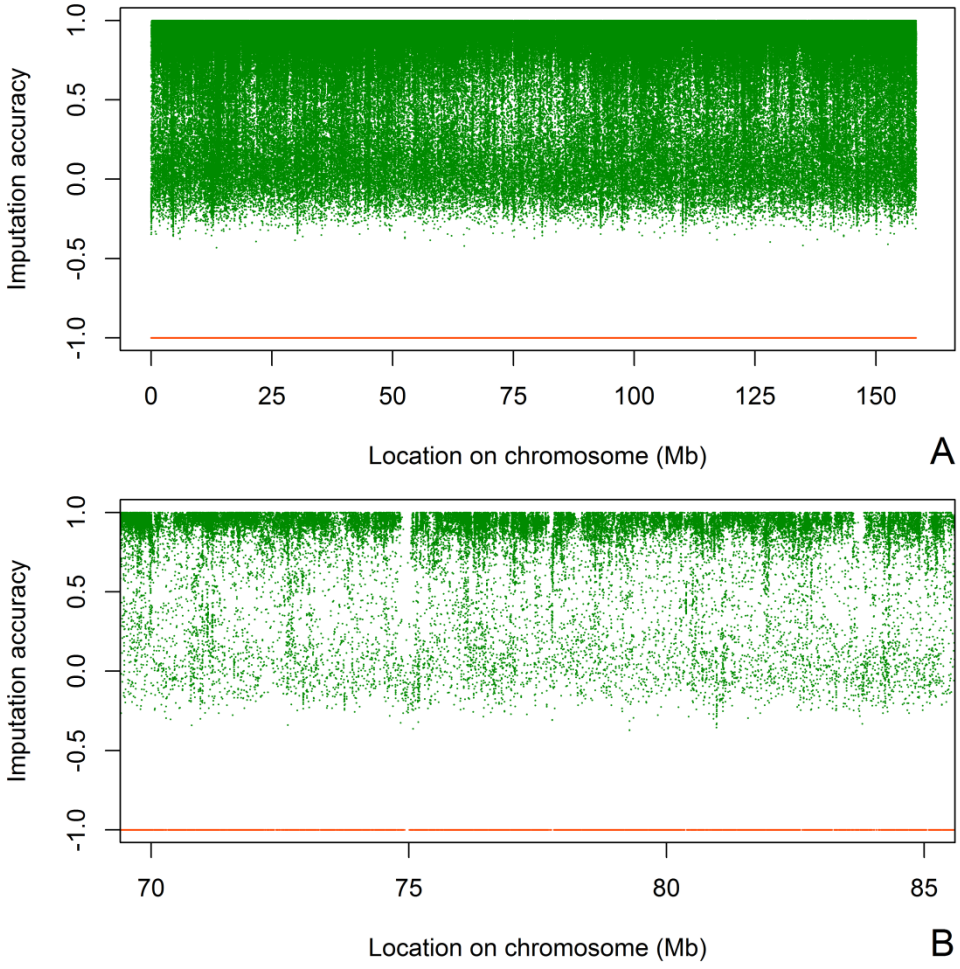

Figure 3.3 Accuracy of imputation from the BovineHD beadchip on BTA1. (A) for the complete BTA1. **(B)** for the region between70 and 85 Mb on BTA1. Location on BTA1 versus accuracy of imputation from the BovineHD beadchip to whole-genome sequence data for scenario S80; each green dot represents a SNP; orange dots at -1 are locations of SNPs of the BovineHD beadchip

Table 3.1 Mean accuracy of imputation per SNP. Mean, standard deviation (SD), minimum and maximum accuracy of imputation per SNP on BTA1 for different combinations of scenarios and lower density marker panels; for scenario S40, accuracy of stepwise imputation is also shown for step 1 (BovineSNP50 to BovineHD), step 2 (BovineHD to sequence), and overall; number of SNPs used for analyses are presented in the last column

		Mean	SD	Minimum	Maximum	Nb SNPs
S80	**BovineHD**	0.83	0.27	-0.43	1.00	744 896
	BovineSNP50	0.46	0.37	-0.54	1.00	768 907
S60	**BovineHD**	0.81	0.27	-0.37	1.00	736 216
	BovineSNP50	0.43	0.36	-0.58	1.00	780 388
S40	**BovineHD**	0.77	0.29	-0.33	1.00	739 859
	BovineSNP50	0.37	0.36	-0.40	1.00	764 439
2-step	**Step 1**	0.83	0.15	-0.17	1.00	32 880
	Step 2	0.77	0.29	-0.33	1.00	739 859
	Overall	0.65	0.30	-0.41	1.00	764 912

Table 3.2 Mean accuracy of imputation per individual. Mean, standard deviation (SD), minimum and maximum accuracy of imputation per individual on BTA1 for different combinations of scenarios and lower density marker panels; for scenario S40, accuracy of stepwise imputation is also shown for step 1 (BovineSNP50 to BovineHD), step 2 (BovineHD to sequence), and overall; number of SNPs used for analyses are presented in the last column

		Mean	SD	Min	Max	Nb SNPs
S80	**BovineHD**	0.95	0.04	0.70	0.97	744 896
	BovineSNP50	0.80	0.04	0.61	0.85	768 907
S60	**BovineHD**	0.94	0.04	0.70	0.97	736 216
	BovineSNP50	0.79	0.04	0.61	0.85	780 388
S40	**BovineHD**	0.93	0.04	0.69	0.96	739 859
	BovineSNP50	0.78	0.04	0.60	0.85	764 439
2-step	**Step 1**	0.92	0.07	0.53	0.99	32 880
	Step 2	0.93	0.04	0.69	0.96	739 859
	Overall	0.86	0.07	0.53	0.95	764 912

Figure 3.4 shows the mean imputation reliability versus distance to the nearest SNP on the BovineHD beadchip for the three scenarios. Imputation reliability (imputation accuracy squared) decreased with increasing distance between imputed SNP and nearest SNP on the BovineHD panel. This decrease in imputation reliability follows the decay in LD, described as $r_{dist}{}^2$, for *Ne* = 1000. Even at very small distances, the observed imputation reliability is lower than $r_{dist}{}^2$. In addition to this distance effect, reference group size has an effect. Since imputations from the BovineHD and BovineSNP50 panels showed similar patterns for distance and all other factors, only the results for the imputation from the BovineHD panel are shown.

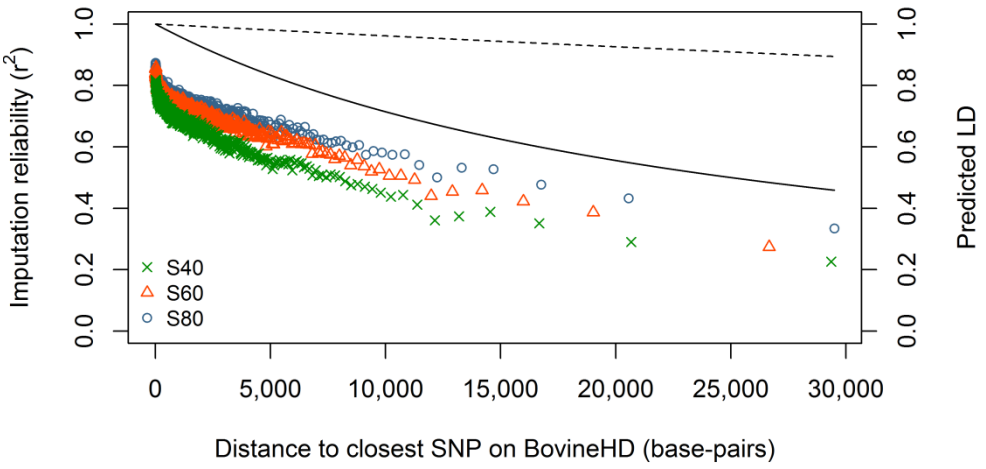

Figure 3.4 Distance to the nearest SNP on the BovineHD beadchip versus mean imputation reliability for imputation from the BovineHD panel to whole-genome sequence data on BTA1 for the three scenarios (S40, S60, and S80); SNPs were grouped in bins of 1000 SNPs with similar distance; the predicted LD ($\boldsymbol{r_{dist}{}^2}$) was calculated with assumed effective population sizes (*Ne*) of 100 (dashed line) and 1000 (solid line).

The difference in MAF between imputed SNPs and their nearest SNPs on the BovineHD beadchip determines the maximum LD between two SNPs. Figure 3.5 shows this MAF difference versus $r_{dMAF}{}^2$ and versus mean imputation reliability for imputation from the BovineHD beadchip for all three scenarios. For differences in MAF below 0.05, imputation reliability was below $r_{dMAF}{}^2$, which was in agreement with expectation based on maximum LD. For larger differences in MAF, observed imputation reliabilities were above estimations from $r_{dMAF}{}^2$. This pattern implies that other SNPs than only the nearest SNP on the BovineHD panel influenced imputation reliability.

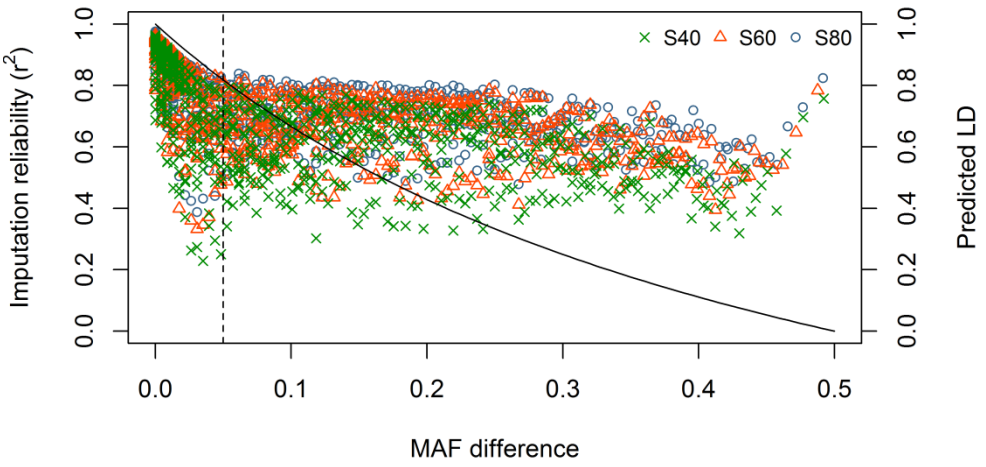

Figure 3.5 Differences in MAF with the nearest SNP on the BovineHD beadchip versus mean imputation reliability. Differences in MAF between imputed SNP and the nearest SNP on the BovineHD beadchip versus predicted LD $(\boldsymbol{r_{dMAF}{}^2})$ and versus mean imputation reliability for imputation from the BovineHD panel to whole-genome sequence data on BTA1 for the three scenarios (S40, S60, and S80); SNPs were grouped in bins of 1000 SNPs with similar MAF differences

The effect of MAF of imputed SNPs on imputation reliability is shown in Figure 3.6, with a Michaelis-Menten curve fitted for each scenario separately. Imputation reliability increased with increasing MAF. This increase in imputation reliability was more pronounced at a MAF below 0.2. The estimated value for the upper limit of $r_{MAF}{}^2$ (V_{max}) was 1.01 (SE = 0.007) for scenario S40, 0.98 for S60 (SE = 0.005), and 0.95 (SE = 0.004) for S80. The maximum value of $r_{MAF}{}^2$ at the maximum MAF value (MAF = 0.5) was 0.881 for scenario S40, 0.893 for S60, and 0.886 for S80. The estimated MAF when $r_{MAF}{}^2$ = 1/2V_{max}, or at the deflection point K_m was equal to 0.073 (SE = 0.002) for scenario S40, 0.049 (SE = 0.001) for S60 and 0.036 (SE = 0.001) for S80.

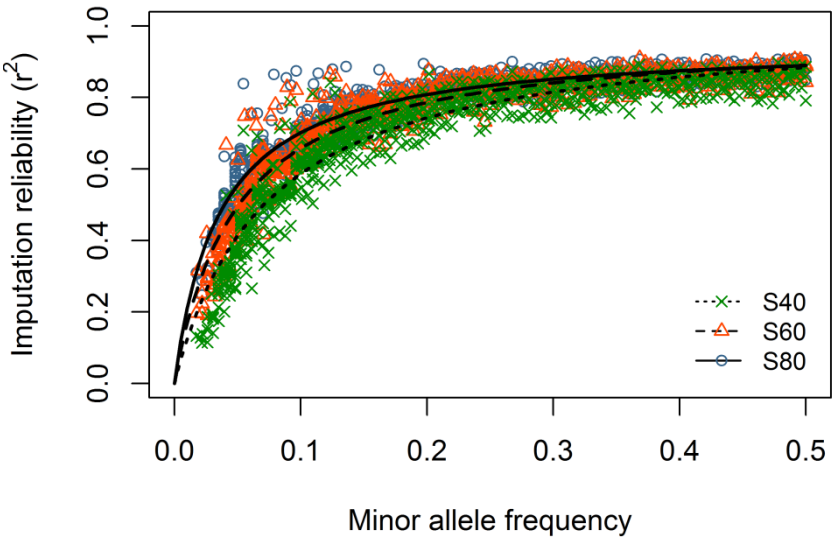

Figure 3.6 Effect of MAF of imputed SNP and number of reference individuals on reliability of imputation. Combined effect of MAF of imputed SNPs and scenario (S40, S60, and S80) on reliability of imputation from the BovineHD beadchip to whole-genome sequence data on BTA1; SNPs per scenario were grouped in bins of 1000 SNPs with similar MAF; for each scenario a Michaelis-Menten function was fitted.

Figure 3.7 shows the overall estimation of imputation reliability ($r_{total}{}^2$, *Ne* = 1000) against observed imputation reliability for the three scenarios (S40, S60, S80). The estimated $r_{total}{}^2$ followed the observed reliabilities closely, although the estimated $r_{total}{}^2$ were higher than the observed reliabilities. At low $r_{total}{}^2$, the observed imputation reliability deviated more from estimated $r_{total}{}^2$. In particular, scenarios with a higher number of individuals showed larger observed imputation reliabilities compared to the estimated $r_{total}{}^2$.

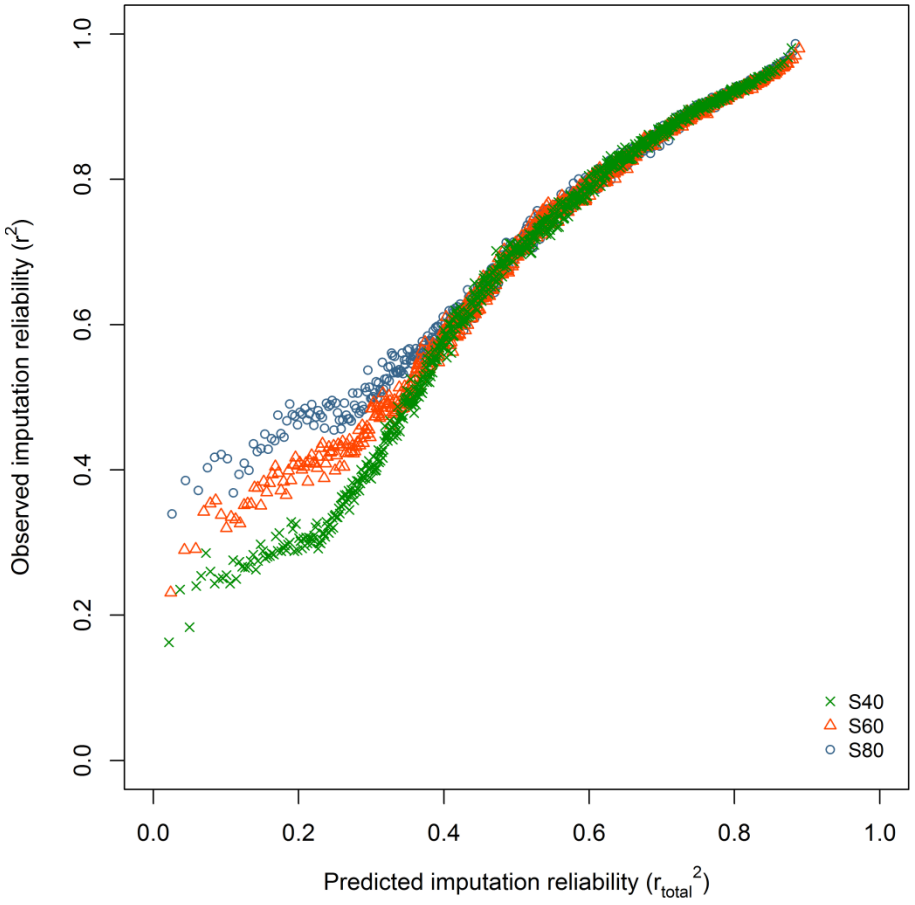

Figure 3.7 Overall prediction of imputation reliability versus observed imputation reliability. Overall prediction of imputation reliability ($\boldsymbol{r_{total}{}^2}$, Ne = 1000) plotted against observed imputation reliability for imputation from the BovineHD panel to whole-genome sequence data on BTA1 for three scenarios (S40, S60, and S80); SNPs were grouped in bins of 1000 SNPs with similar $\boldsymbol{r_{total}{}^2}$.

3.4 Discussion

3.4.1 Imputation from the lower density panel

Our objective was to investigate accuracy of imputation from the lower density SNP panels to whole-genome sequence data in Holstein Friesian cattle. Accuracy of imputation was defined as the correlation between observed genotypes and the imputed *B*-allele dosages. Mean accuracy of imputation per SNP to whole-genome sequence data was equal to 0.46 with 0.18% of SNPs known (BovineSNP50), and 0.83 with 2.33% of SNPs known (BovineHD). We chose to use the correlation between observed and imputed genotypes to measure accuracy of imputation, whereas most studies used percentage of correctly imputed SNPs. Compared to correlation between observed and imputed genotypes, percentage of correctly imputed SNPs does not account for the (low) MAF of imputed SNPs. A necessary condition for correlation between two random variables is that both variables show variation. Therefore, SNPs with fixed observed genotypes or estimated *B*-allele dosages for one or more validation groups were removed. This might have caused a positive bias in the results, because of removal of monomorphic loci with poor imputation. In other studies (e.g. Hao *et al.*, 2009; Hayes *et al.*, 2012; VanRaden *et al.*, 2013), criteria such as MAF greater than 0.01 were used in data editing procedures. If this type of criteria had been applied to the sequence data in our study, a large number of SNPs (987 514) would have been removed, which is similar to what occurred with the criterion chosen here.

Previous studies showed that increasing the number of close relatives between imputed and reference individuals increased imputation accuracy (Iwata and Jannink, 2010; Zhang and Druet, 2010; Hayes *et al.*, 2012; Hickey *et al.*, 2012). The sequenced bulls in this study were key ancestors of the global Holstein Friesian population and in general, were not very closely related. In fact, in some cases, they were chosen to be as little related as possible, in order to maximize sequencing effort of unique chromosome segments. A genomic relationship matrix (Yang *et al.*, 2010) was constructed based on SNPs found on BTA1. About 90% of the off-diagonals were below 0.125 and 0.5% were above 0.5 (results not shown). In practice, these sequenced bulls will be used as reference individuals to impute genotypes of other individuals in the current population, which might be their

progeny or otherwise closely related individuals. Therefore, it is expected that, in practice, imputation accuracies will be higher than those estimated in this study.

SNPs used in this study were called in a larger multi-breed population than the 114 Holstein individuals included here. Ideally, to better mimic the reality and answer the question on how many individuals need to be sequenced, the number of reference individuals used in the three scenarios should also be used for variant calling. This is important since the set of individuals used for variant calling influences the called genotypes and therefore a bias might be introduced in this study. However, we expect that the effect on the results is small, because we disregarded SNPs that did not show variation in either the reference or validation set. These are also SNPs that will not be called if only the Holstein individuals are used for variant calling. Another deviation from a real situation is that, for imputation, we assumed that the called genotypes from the sequence data were true genotypes, while it would have been more correct to use the probabilities of inferred genotypes from the sequence data as starting point for imputation. Therefore, imputation accuracies estimated in this study may differ slightly from accuracies obtained from "true genotypes".

Mean imputation accuracy per SNP from the BovineSNP50 panel to whole-genome sequence data was below 0.46. Our results showed that an alternative approach, i.e. using stepwise imputation from the BovineSNP50 to the BovineHD panel and then to sequence data, also yielded high accuracies of imputation. For example, in scenario S40, accuracy of the stepwise imputation was higher (0.65) than that of direct imputation from the BovineSNP50 beadchip to sequence data (0.37) or even than that of direct imputation from the BovineSNP50 beadchip in scenario S80 (0.46). Such a high accuracy with the stepwise approach was unexpected, because less information was available in the reference set. In the two-step approach, 20% of the individuals had genotypes similar to those of the BovineSNP50 panel (validation individuals), 40% had genotypes similar to those of the BovineHD panel (reference individuals step 1), and 40% had sequenced genotypes (reference individuals step 2). Whereas, in scenario S80, with direct imputation from the BovineSNP50 panel to sequence data, all reference individuals (80% of all individuals) had sequenced genotypes. VanRaden *et al.* (2013) found an increase in imputation accuracy of about 2% when imputation was done from 3000 SNPs to 50 000 SNPs and then to 777 000 SNPs compared to direct imputation

from 3000 SNPs to 777 000 SNPs. Although less information is used, the reason why there is this increase in imputation accuracy is not clear. However, one reason could be that the imputation algorithm has problems with selecting the correct haplotypes since there are multiple possible matches between sequence haplotypes and a BovineSNP50 haplotype, whereas there are less possible matches when BovineHD genotypes are added in between. In this case, there is a higher probability of selecting the long range haplotypes in the first step, and the short range haplotypes in the second step, which increases accuracy of imputation.

In cattle, many individuals with BovineHD genotypes are available. Using those individuals to impute BovineSNP50 genotypes to BovineHD genotypes may increase the accuracy gained in the first step, which would result in even higher accuracies when using the two-step approach than those obtained here. In some species, this is not a realistic scenario because no high-density marker panel is available yet i.e. for pig. Developing these high-density panels and re-genotyping individuals can be expensive, especially if the end goal is to impute to sequence genotypes. In a scenario in which no high-density panel is available, it might be more cost effective to sequence additional animals and use the two-step approach by masking part of the SNPs of the individuals used for the first imputation step. This will mimic a high-density marker panel, and according to the results reported here, the overall imputation accuracy would be higher than that obtained by direct imputation from the lower density SNP chip. An improvement of this step-wise approach could be to use information of all individuals in the reference population in both steps instead of using disjoint reference sets as was done in this study, to mimic dairy cattle breeding practice. In the former case, the expected advantage is that all the genotype information will be available in the last step, while with disjoint datasets, the masked genotype information of individuals in the first step is not used in the second step. Moreover, it would be interesting to investigate the use of more than two steps because there may be an optimum number of steps to reach the highest accuracy.

In genomic selection, it is important to know the imputation accuracy per individual, because there is a direct relation with the accuracy of genomic prediction (Mulder *et al.*, 2012) and therefore the response of selection. In the present study, mean imputation accuracy per individual was higher compared to mean imputation accuracy per SNP, which was also reported by Mulder *et al.*

(2012). They argued that allele frequencies bias imputation accuracy per individual and suggested to subtract mean genotype per SNP from observed and imputed genotypes. We tested this hypothesis and showed it had a small effect i.e. the mean accuracy of imputation from the BovineHD panel per individual in scenario S80 decreased only by 0.04 to reach 0.90. After standardization for the genotype variance per SNP, mean accuracy of imputation per individual in scenario S80 decreased furthermore to 0.87. This standardized mean accuracy per individual is still higher compared to the mean accuracy per SNP, however, the remaining bias is small and might be explained by a correlation between imputations of markers within a haplotype within an individual (Mulder *et al.*, 2012).

3.4.2 Imputing SNPs with a low MAF

Using whole-genome sequence data for genomic prediction and GWAS is interesting because the actual polymorphisms that cause genetic differences are potentially included in the data (e.g. Meuwissen and Goddard, 2010; Li *et al.*, 2011b; Druet *et al.*, 2014). The distribution of allele frequencies of causal mutations is not known, but it is hypothesized that those mutations may have a low MAF (Druet *et al.*, 2014). To calculate imputation accuracy, all SNPs with fixed observed genotypes or estimated *B*-allele dosages for one or more validation groups were removed. The remaining numbers of SNPs per scenario and per SNP chip are in Table 3.1. In the case of imputation from the BovineHD panel in scenario S80, 744 896 SNPs remained and 992 575 SNPs were removed from the dataset. It is possible that removing these SNPs without changing the allele dosage affected the results. Of the removed SNPs, 40.6% had a MAF of 0, which could have been easily imputed with a 100% accuracy, 56.1% had a MAF between 0 and 0.1and their imputation accuracy could have been affected by their low MAF only, and the remaining 3.3% had a MAF above 0.1, which could have been difficult to impute for other reasons than their low MAF. However, it is unlikely that these 3.3% SNPs could affect the average imputation accuracy of common markers because of their small number. Although many loci with a low MAF in the observed genotypes were removed, among the remaining SNPs those with a lower MAF were more difficult to impute correctly and the reliability of imputation varied more than for the SNPs with a higher MAF. These findings may potentially limit the benefit of using imputed sequence data for genomic prediction and GWAS. However, decay in imputation

reliability for SNPs with a lower MAF was smaller in the scenarios with more reference individuals than those with less reference individuals, which confirms results with human data (Browning and Browning, 2009). In large-sized reference populations, there is more chance to have multiple allele copies to construct the haplotypes (Li *et al.*, 2011a). Moreover, Howie *et al.* (2011) showed that a multi-population reference panel can improve imputation accuracy for SNPs with a low MAF, because a low-frequency allele in one population can be more frequent in another population. Since it is expected that, in the near future, more individuals from more different breeds will be sequenced in cattle, it is assumed that imputation accuracy of SNPs with a low MAF will improve.

Still, in species with a small number of sequenced individuals, imputation of SNPs with a low MAF may remain an issue. In such a situation, it might be beneficial to use another algorithm for imputation, such as IMPUTE (Marchini *et al.*, 2007) or MaCH (Li *et al.*, 2010). It is claimed that these methods perform better compared to Beagle when the number of reference individuals is low (Browning, 2008; Nothnagel *et al.*, 2009) and for SNPs with a low MAF (Pei *et al.*, 2008). All three methods use Hidden Markov models, but IMPUTE and MaCH model genotypes on a set of haplotypes without clustering, whereas Beagle uses haplotype clustering strategies and therefore may miss SNPs with a low MAF (Browning, 2008; Pei *et al.*, 2008). Clustering strategies as in Beagle reduce computer time and memory use compared to IMPUTE and MaCH, which is an advantage when handling large datasets (Nothnagel *et al.*, 2009).

3.4.3 Imputation reliability per SNP

Although the assumption that the polymorphisms responsible for genetic differences are included in the dataset may be true for sequence data, for imputed sequence data it is important to know if polymorphisms are imputed correctly. Beagle calculates an allelic R^2 measure, which predicts accuracy of imputation per SNP. Allelic R^2 is the squared correlation between allele dosage of the most likely imputed genotype and allele dosage of the true imputed genotype (Browning and Browning, 2009) and the closer these are, the more accurate the imputation is for the SNP. The correlation between the allelic R^2 measure from Beagle and true imputation reliability that we calculated was equal to 0.79 for imputation from the BovineHD beadchip to sequence data in scenario S80 (results not shown). Of the

622 862 SNPs with estimates for both measures, 67,2% showed a difference between the allelic R^2 measure from Beagle and true imputation reliability of less than 0.1, although the maximum difference between both measures was 0.78. This indicates that the allelic R^2 measure provided by Beagle gives a good indication of imputation reliability in general, although in specific cases it may severely underestimate imputation reliability.

In human studies, imputed genotypes did not result in a high increase in power in GWAS compared to lower density marker panels (Hao *et al.*, 2009; Huang *et al.*, 2009; Pasaniuc *et al.*, 2012). Therefore, it is important to understand the underlying factors that affect imputation reliability and to take those factors into account when imputing genotypes. An important factor that influences imputation reliability is the LD between the imputed SNP and the SNP on the lower density marker panel. This may reduce the advantage of using imputed sequence data for genomic predictions or GWAS, compared to true sequence data. The advantage with true sequence data is the lack of dependency on LD between an SNP and the causal mutation in the sequence data, assuming that the true causal variant was accurately identified in the data. Our results showed that successful imputation of the causal mutation depended on the LD between the SNP on the lower density marker panel and the causal mutation. Hence, causal mutations that are poorly tagged by the low-density SNP panel will also be difficult to detect for reliable imputation.

In the current Holstein Friesian population, the effective population size is estimated to be around 100 (de Roos *et al.*, 2008; Qanbari *et al.*, 2010). However, Figure 3.4 shows that the decay in imputation accuracy based on a *Ne* of 1000 seemed more appropriate for our data than a *Ne* of 100. Hayes *et al.* (2003) reported that LD at very short distances is related to effective population sizes in the past, while LD at longer distances is related to current effective population sizes. In our study, LD was calculated on very short distances, which suggests that a historical value should be used for *Ne*, rather than the current value of 100. Another reason for imputation reliability to decay more quickly than that expected from the decay in LD based on a *Ne* of 100 is that other factors also affected imputation reliability, or that the factors interacted with respect to their effect on accuracy. For example, when the SNP selected on the high-density panel and the SNP in the sequence are close, their MAF may be comparable, while as the distance between

them increases the difference in MAF may also increase. Since these factors, distance and MAF, have a multiplicative effect, the decay in imputation reliability is larger than that expected from the decay in LD based on a *Ne* of 100. This expectation is confirmed by the resemblance between the combined functions for *Ne* of 100 (results not shown) and the combined functions for *Ne* of 1000 (Figure 3.7).

Another factor that affected LD was the difference in MAF, which at first sight may be an unexpected indicator for imputation accuracy, especially since haplotypes are used for imputation. However, as shown in other studies (Lewontin, 1995; Mueller, 2004; Miller, 2013) the difference in MAF determines the mathematical upper limit of the LD between two SNPs. At extreme differences in MAF, alleles at the different SNPs cannot match, even if the distance between SNPs is small. For example, the maximum possible correlation obtained for two random binary variables with a MAF of 0.45 and 0.05, respectively, is 0.06. Thus, for two SNPs at the same distance, LD may differ and they may be in different haplotypes used for imputation. This could be particularly important since the SNPs included in the SNP panels are not randomly selected and generally have a high MAF.

Imputation reliability was also affected by the MAF of the imputed SNPs and by the number of sequenced individuals. Our results indicate that, if causal mutations have a low MAF, a large-sized reference group is required to impute those mutations correctly and to benefit from using sequence data, which confirms previous reports (Clark *et al.*, 2011; Druet *et al.*, 2014). Extrapolation of K_m using a power function (R^2 = 0.999) showed that, with more than 500 reference individuals, the increase in imputation reliability was expected to be small (results not shown). This agrees with other cattle studies that used lower density marker data and showed that, with more than 1000 reference individuals, the increase in imputation accuracy is expected to be small (Druet *et al.*, 2010; Zhang and Druet, 2010).

The goal of imputation is to assemble a large group of individuals with phenotypic information and sequence genotypes for genomic prediction or GWAS. For power calculations in GWAS, imputation reliability (not only overall imputation reliability but also imputation reliability per SNP because of the variation between SNPs) should be taken into account when imputed genotypes are used (Marchini *et al.*, 2007). Our results show that functions that estimate LD based on distance only or on the difference in MAF between the imputed SNP and the closest SNP on the

lower density marker panel did not provide a good indication of imputation reliability. When these functions were combined with an empirical derived function that corrects for MAF of the imputed SNPs and size of the reference group, a much better indication of imputation reliability was obtained but it was still not perfect (Figure 3.7). The same functions also held for BTA29, even when using estimates for V_{max} and K_m based on BTA1 (results not shown). Hence within this population and dataset, the predictions hold across chromosomes, at least on average since bins of 1000 SNPs were used. However, these functions could be further improved. For example, currently the functions are based on the use of an individual SNP (the closest SNP or the SNP in highest LD of the five closest SNPs) to estimate imputation reliability, whereas a program like Beagle uses haplotypes for imputation. Moreover, instead of choosing the closest SNP, a more distant SNP might be in higher LD with the imputed SNP. Therefore, using all SNPs or haplotypes is likely to estimate imputation reliability better than the functions used here. However, taking all SNPs into account or using haplotypes will make estimation more time-consuming and less generic applicable. Further research using simulation is necessary to investigate the generality of the estimations and the obtained imputation reliability. However, our study shows that the functions described above provide a good indication of the factors that affect imputation reliability per SNP.

Obviously, imputation reliability does not rely only on LD, MAF, and reference group size. Other factors, such as genotyping errors (Browning, 2008), or degree of relationship between validation and reference groups (Iwata and Jannink, 2010; Zhang and Druet, 2010; Hickey *et al.*, 2012), are also important. It has been reported that increasing the number of close relatives in the reference group increased accuracy of imputation and that this increase was more pronounced when the differences between number of SNPs genotyped in the validation and reference populations were large (such as the differences between BovineSNP50 or BovineHD and sequence data) (Hickey *et al.*, 2012).

3.5 Conclusions

Accuracy of imputation to whole-genome sequence data was generally high for imputation from the BovineHD beadchip, but was low for imputation from the BovineSNP50 beadchip. Stepwise imputation from the BovineSNP50 to the BovineHD beadchip and to sequence data substantially improved accuracy of imputation. SNPs with a lower MAF were more difficult to impute correctly and led to more variation in reliability of imputation. Functions that estimate LD based on distance only or on the difference in MAF between the imputed SNP and the closest SNP on the lower density marker panel did not provide a good indication of imputation reliability. However, when these functions were combined with an empirical derived function that corrects for MAF of the imputed SNPs and size of the reference group, estimation of imputation reliability was greatly improved.

3.6 Acknowledgements

The authors want to acknowledge the 1000 bull genomes consortium for providing the data, John Hickey for his useful comments, and the Breed4Food project (program "Kennisbasis Dier", code: KB-12-006.03-004-ASG-LR) for financial support.

CHAPTER 4

GENOMIC PREDICTION USING IMPUTED WHOLE-GENOME SEQUENCE DATA IN HOLSTEIN FRIESIAN CATTLE

RIANNE VAN BINSBERGEN[1, 2]
MARIO P.L. CALUS[1]
MARCO C.A.M. BINK[2]
FRED A. VAN EEUWIJK[2]
CHRIS SCHROOTEN[3]
ROEL F. VEERKAMP[1]

[1] Animal Breeding and Genomics Centre, Wageningen Livestock Research, P.O. Box 338, 6700 AH Wageningen, the Netherlands
[2] Biometris, Wageningen University and Research Centre, P.O. Box 100, 6700 AC Wageningen, the Netherlands
[3] CRV, Arnhem, The Netherlands

Genetics Selection Evolution (2015) 47:71

Abstract

Background: In contrast to currently used single nucleotide polymorphism (SNP) panels, the use of whole-genome sequence data is expected to enable the direct estimation of the effects of causal mutations on a given trait. This could lead to higher reliabilities of genomic predictions compared to those based on SNP genotypes. Also, at each generation of selection, recombination events between a SNP and a mutation can cause decay in reliability of genomic predictions based on markers rather than on the causal variants. Our objective was to investigate the use of imputed whole-genome sequence genotypes versus high-density SNP genotypes on (the persistency of) reliability of genomic predictions using real data.

Methods: Highly accurate phenotypes based on daughter performance and Illumina BovineHD Beadchip genotypes were available for 5503 Holstein Friesian bulls. The BovineHD genotypes (631 428 SNPs) of each bull were used to impute whole-genome sequence genotypes (12 590 056 SNPs) using the Beagle software. Imputation was done using a multi-breed reference panel of 429 sequenced individuals. Genomic estimated breeding values for three traits were predicted using a Bayesian stochastic search variable selection (BSSVS) model and a genome-enabled best linear unbiased prediction model (GBLUP). Reliabilities of predictions were based on 2087 validation bulls, while the other 3416 bulls were used for training.

Results: Prediction reliabilities ranged from 0.37 to 0.52. BSSVS performed better than GBLUP in all cases. Reliabilities of genomic predictions were slightly lower with imputed sequence data than with BovineHD chip data. Also, the reliabilities tended to be lower for both sequence data and BovineHD chip data when relationships between training animals were low. No increase in persistency of prediction reliability using imputed sequence data was observed.

Conclusions: Compared to BovineHD genotype data, using imputed sequence data for genomic prediction produced no advantage. To investigate the putative advantage of genomic prediction using (imputed) sequence data, a training set with a larger number of individuals that are distantly related to each other and genomic prediction models that incorporate biological information on the SNPs or that apply stricter SNP pre-selection should be considered.

Keywords: genotype imputation, BovineHD, GBLUP, Bayesian

4.1 Introduction

Genomic selection is increasingly applied in breeding programs for livestock and plant species (e.g. Hayes *et al.*, 2009; Heffner *et al.*, 2009; Goddard *et al.*, 2010; Jannink *et al.*, 2010). Genomic selection relies on the prediction of genomic estimated breeding values (GEBV) of individuals or lines using marker genotype information only, by applying genomic prediction models that are based on training individuals that have both phenotypic and genotypic data. In most breeding programs, single nucleotide polymorphism (SNP) marker panels are used. With SNP panels, the level of linkage disequilibrium (LD) between SNPs and the actual causal variant (e.g. SNP, insertion, deletion, etc.) influences the reliability of genomic prediction. In this paper, these causal variants will be considered as quantitative trait loci (QTL). At each generation of selection, recombination events between a SNP and the QTL can cause a decay in the reliability of genomic predictions (Muir, 2007). Typically, a decrease in reliability of GEBV prediction in cattle with 50k SNP genotypes has been observed when the additive-genetic relationships between training animals and validation animals decrease (Habier *et al.*, 2010; Pszczola *et al.*, 2012). Moreover, this decay in reliability was greater when the size of the training set was smaller (Habier *et al.*, 2010). This decay could become a problem for dairy cattle since sons from young bulls that are selected on their GEBV without daughter information are now entering breeding programs. These sons' GEBV are estimated based on a training set of progeny-tested bulls that are two generations older (i.e. their grand sires) and therefore the reliability of their genomic breeding values will be lower compared with those of the previous generation.

On average, increasing the number of SNPs in a panel increases the level of LD between a SNP and a QTL and this should be beneficial for genomic prediction. Studies using real data, have shown that genomic prediction using an array with approximately 777 000 (imputed) SNPs resulted in a small gain in genomic prediction reliability compared to an array with approximately 50 000 SNPs (Erbe *et al.*, 2012; Su *et al.*, 2012; Ertl *et al.*, 2014). However, even with 777 000 SNPs, predictions still depend on LD between SNPs and QTL. In contrast to the SNP panels currently used, whole-genome sequence data are expected to include the causal mutations that underlie QTL (Meuwissen and Goddard, 2010), which means

that predictions should no longer depend on LD between SNPs and QTL. Inclusion of the causal mutations allows the effect of the QTL on a given trait to be estimated directly, which should increase the reliability of genomic predictions compared to using SNP genotypes, as well as the persistency of the reliability of predictions across generations and even across breeds (Meuwissen and Goddard, 2010; Clark *et al.*, 2011; Macleod *et al.*, 2013).

However, identifying the QTL and obtaining a higher persistency of reliabilities of genomic predictions over generations probably requires a large training set of thousands of sequenced individuals. Without a large number of training individuals, QTL effects might be estimated with too much error and thus, there will be little advantage of using sequence data (Druet *et al.*, 2014). Sequencing many individuals is still too expensive but instead imputed sequence data can be used, especially since many animals that are genotyped using SNP panels are available in livestock populations.

The 1000 bull genome project (Daetwyler *et al.*, 2014) aims at sequencing a number of key ancestor bulls in the beef and dairy cattle population at medium coverage. These sequenced animals can be used as reference animals to impute other animals that are genotyped with 50k or 777k SNP panels to the whole-genome sequence level. A reliability of 0.83 was obtained for imputation from 777k SNP panels to sequence data with a reference set of 91 Holstein Friesian animals with whole-genome sequence data (van Binsbergen *et al.*, 2014a). Moreover, adding individuals of other breeds in a relatively large reference set will further increase imputation accuracy. In particular, it was reported that low MAF (minor allele frequency) variants that segregate in other breeds can benefit from combining different breeds together (Bouwman and Veerkamp, 2014; Brøndum *et al.*, 2014). Therefore, imputation to sequence data using SNP genotypes is an attractive and cost-effective approach to obtain a large training set of sequenced individuals, and to investigate the benefit of using sequence data for relevant populations.

Many methods are available for genomic prediction, most of which are based on linear regression (see de los Campos *et al.* (2013) for review). These methods can differ in the underlying assumptions about the distribution of SNP effects. With a genome-enabled best linear unbiased prediction (GBLUP) model it is assumed that the a priori variance of SNP effects is equal, so a large number of

SNPs, each with a small effect, are fitted in the model (infinitesimal model). Consequently, it is expected that GBLUP does not take full advantage of sequence data, since it will allocate the same variance to SNPs without effect and to those that are causal, although only a very small proportion of the SNPs is expected to be causal. Alternatively, methods such as BayesB (Meuwissen *et al.*, 2001), BayesC (Habier *et al.*, 2011) and Bayesian stochastic search variable selection (BSSVS) (Verbyla *et al.*, 2009; Calus, 2014) assume that the a priori variance of the effects of many SNPs is very small or zero, while it is large for only a few SNPs. Because of this mixture of the prior distributions of SNP effects, these methods could benefit from sequence data. Simulation studies using bovine sequence data confirmed this expectation (e.g. Meuwissen and Goddard, 2010; Clark *et al.*, 2011; Macleod *et al.*, 2013). However, Ober et al. (Ober *et al.*, 2012) concluded that predictions from BayesB were not better than predictions from a method equivalent to GBLUP when using real sequence genotypes of *Drosophila melanogaster*, although the size of the training set size (~120 observations) was relatively small. Moreover, the advantage of Bayesian methods over GBLUP was shown to be greatly influenced by the size and distribution of the simulated QTL effects (Meuwissen and Goddard, 2010; Clark *et al.*, 2011; Macleod *et al.*, 2013).

Since the use of whole-genome sequence data for genomic prediction in livestock populations, and its impact on the reliability of genomic prediction and persistency across generations have been mainly studied with simulated data, the objective of this study was to investigate (the persistency of) the reliability of genomic predictions based on imputed whole-genome sequence genotypes versus 777k SNP genotypes for real dairy cattle data.

4.2 Methods

4.2.1 Phenotypes

De-regressed proofs and associated weights (effective daughter contributions, EDC) were available for somatic cell score (SCS), interval between first and last insemination (IFL), and protein yield (PY) for 5503 Holstein Friesian bulls provided by CRV (Arnhem, the Netherlands). De-regressed proofs (DRP) were calculated according to VanRaden *et al.* (2009):

$$\mathrm{DRP} = \mathrm{PA} + (\mathrm{EBV} - \mathrm{PA}) * \left(\frac{\mathrm{EDC_{EBV}}}{\mathrm{EDC_{prog}}}\right)$$

where EBV is the estimated breeding value of a bull for a trait available from the national evaluations, and PA is the parent average of the bull for that trait. Effective daughter contribution, EDC_{EBV}, represents the effective number of daughters with phenotypes that contributed to the estimated breeding value of a bull (Fikse and Banos, 2001) and was calculated according to VanRaden and Wiggans (1991) as $\alpha * \mathrm{REL}_{\mathrm{EBV}}/(1-\mathrm{REL}_{\mathrm{EBV}})$, where REL_{EBV} is the published reliability for EBV and $\alpha = (4-h^2)/h^2$, where h^2 is the heritability of the trait. $\mathrm{EDC}_{\mathrm{prog}} = \mathrm{EDC}_{\mathrm{EBV}} - \mathrm{EDC}_{\mathrm{PA}}$, where $\mathrm{EDC}_{\mathrm{PA}} = \alpha\mathrm{REL}_{\mathrm{PA}}/(1-\mathrm{REL}_{\mathrm{PA}})$ and $\mathrm{REL}_{\mathrm{PA}} = (\mathrm{REL}_{\mathrm{sire}} + \mathrm{REL}_{\mathrm{dam}})/4$ (VanRaden and Wiggans, 1991). As the number of daughters with phenotypic information for a trait increases, the reliability of the EBV of a bull and EDC_{EBV} increase. The average EDC_{EBV} (and its range) for animals in the training set was equal to 266 (24 – 971) for SCS, 643 (47 – 4851) for IFL, and 245 (24 – 693) for PY. The pedigree information for the 5503 bulls in this study included 39 917 animals.

4.2.2 Genotypes

In total, 551 bulls were genotyped with the Illumina BovineHD BeadChip (Illumina Inc., San Diego, CA) and the other 4952 bulls were genotyped with a 50k SNP panel and imputed to BovineHD (734 403 SNPs). Imputation from the 50k to the BovineHD SNP panel was performed with Beagle 3.3.0 (Browning and Browning, 2007; Browning and Browning, 2009), using a reference set of 1333 animals genotyped with the BovineHD SNP panel. For this first step, the error rate of imputation was low (Schrooten *et al.*, 2014). For each bull, BovineHD genotypes were subsequently imputed to whole-genome sequence genotypes with Beagle version 4 (Browning and Browning, 2013). The following default parameter settings in Beagle were used: five iterations for initial burn-in, five iterations for phasing, and five iterations for imputation. Imputation was performed for sliding windows of 24 000 SNPs in the sequence data, with an overlap of 3000 SNPs between sliding windows. No pedigree information was used in the imputation procedure. The sex chromosomes were excluded.

Whole-genome sequence data (28 336 153 SNPs) of 429 animals that were provided by the 1000 bull genomes project (Run 3.0) were used as reference data for imputation. All these animals, except two, were males and originated from 15 dairy and beef breeds (1 to 121 animals per breed), among which there were four major breeds, with 121 Holstein, 87 Simmental, 54 Angus, and 43 Brown Swiss

animals. Each animal was sequenced with the Illumina HiSeq System (Illumina Inc., San Diego, CA). Alignment, variant calling, and quality controls were as described by Daetwyler *et al.* (2014). The average number of sequence genotypes was equal to 9.6 per animal and ranged from 3.0 to 44.5. To assess the accuracy of genotype calling, concordance with BovineHD genotypes was calculated as the proportion of identical genotypes between the BovineHD and sequence data and ranged from 67.5 to 99.9% (on average 94.8%) for the 303 animals with BovineHD genotypes. After correcting sequence genotypes with Beagle, average concordance increased to 98.3% (range: 74.1 to 99.9%). Note that most animals in this whole-genome sequence dataset were only used as reference animals for imputation and not for genomic prediction, except for 57 bulls that had genotypes in both datasets.

After imputation, non-informative SNPs were removed from the dataset, i.e. SNPs with less than two alleles, SNPs with a minor allele frequency lower than 0.005 and SNPs with an estimated imputation reliability lower than 0.05 (only for the imputed sequence data). Imputation reliability was predicted by Beagle software as the estimated squared correlation between the estimated allele dosage (0*P(AA) + 1*P(AB) + 2*P(BB)) and the true allele dosage (estimated from posterior genotype probabilities) (Li *et al.*, 2010). In general, the imputation reliability predicted by Beagle gives a good indication of the true reliability for imputation from BovineHD to sequence data (van Binsbergen *et al.*, 2014a). The thresholds for these selection criteria were chosen so that monomorphic SNPs and SNPs that are likely to be imputed incorrectly were removed.

To evaluate the effect of imputation on genomic prediction, a third genotype panel (ImputedHD) was generated by randomly selecting SNPs from the imputed sequence data. The number of selected SNPs per chromosome was the same as for the BovineHD genotype dataset, and did not include SNPs that were in the BovineHD genotype dataset.

4.2.3 Genomic prediction

GEBV for the three traits were predicted based on two sets of genotypes: the original BovineHD genotypes and imputed whole-genome sequence genotypes. In both cases, the most likely genotypes were used for prediction. Genomic prediction was performed using two types of linear regression models: GBLUP and BSSVS.

GBLUP

The GBLUP model was:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of de-regressed proofs of all individuals, μ is the overall mean, $\mathbf{1}$ is a vector of ones, $\mathbf{Z}$ is an incidence matrix that links records to bulls, $\mathbf{g}$ is a matrix of the genomic breeding values of all individuals, and $\mathbf{e}$ contains the random residuals. Genomic breeding values were assumed to be distributed as $\mathbf{g}|\mathbf{GRM}, \sigma_g^2 \sim N(\mathbf{0}, \mathbf{GRM}\sigma_g^2)$, where **GRM** is the genomic relationship matrix, and σ_g^2 is the additive genetic variance picked up by the markers. Diagonal and off-diagonal values of the **GRM** were calculated following Yang *et al.* (2010) as:

$$G_{jk} = \frac{1}{N}\sum_i G_{ijk} = \frac{1}{N}\sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1-p_i)}$$

where G_{ijk} is the estimated relationship between individuals *j* and *k* at locus *i*, and *N* is the number of SNPs. The SNP genotypes (x_i) were coded as 0, 1 or 2, and p_i is the allele frequency of the allele for which the homozygote genotype was coded as 2. Residual effects were assumed to be distributed as$\mathbf{e}|\mathbf{D}, \sigma_e^2 \sim N(\mathbf{0}, \mathbf{D}\sigma_e^2)$, where $\mathbf{D}$ is a diagonal matrix containing $1/\mathrm{EDC}_{\mathrm{EBV}}$ on the diagonals, and σ_e^2 is the residual variance.

After calculation of the genomic relationship matrix, the GBLUP model was fitted using the ASReml 4 software (Gilmour *et al.*, 2014). ASReml software was used to estimate variance components (restricted maximum likelihood estimation, REML), with BLUP of the random effects as 'byproducts'. Therefore, it might be more appropriate to call this method GREML. However, since our main objective was to predict genomic values; we used the terminology GBLUP.

BSSVS

The BSSVS model (Calus, 2014) was:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \mathbf{X}\boldsymbol{\alpha} + \mathbf{e}$$

where $\mathbf{u}$ is a vector that contains the polygenic effects of all bulls ($\mathbf{u}|\mathbf{A}, \sigma_u^2 \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$, where $\mathbf{A}$ is the numerator relationship matrix derived from the pedigree), $\mathbf{X}$ is a matrix that contains the allele dosage (0, 1, or 2) for all SNPs (rows) for all bulls (columns), $\boldsymbol{\alpha}$ is a vector that contains the (random) allele substitution effects for all SNPs. The prior for μ was a constant and both σ_u^2 and σ_e^2 had a flat, uninformative prior distribution.

An important aspect of the BSSVS is that the prior distribution for each allele substitution effects for each locus j (α_j) depends on the variance for the allele substitution effects (σ_α^2) and the QTL indicator I_j, which is sampled for each locus j and takes the value 0 (1) if the SNP was included in the model with a small (large) effect:

$$\alpha_j|I_j,\sigma_\alpha^2 = \begin{cases} \sim N\left(0,\frac{\sigma_\alpha^2}{100}\right) \; when \; I_j = 0 \\ \sim N(0,\sigma_\alpha^2) \; when \; I_j = 1 \end{cases}$$

The prior distribution for I_j was: $p(I_j) = \text{Bernoulli}(1-\pi)$. For both the BovineHD and the imputed sequence datasets, the same number of SNPs (885) was assumed to have a large effect, therefore π was assigned a value equal to $(n_{total} - 885)/n_{total}$, where n_{total} is the total number of SNP effects. The prior distribution for σ_α^2 was: $p(\sigma_\alpha^2) = \chi^{-2}(v_a, S_a^2)$, with $v_a = 4.2$ degrees of freedom (Meuwissen *et al.*, 2001; Habier *et al.*, 2011), and scale parameter $S_a^2 = \frac{\tilde{\sigma}_\alpha^2(v_a-2)}{v_a}$, where $\tilde{\sigma}_\alpha^2 = \left(\frac{100}{100+\pi(1-100)}\right)\frac{\sigma_g^2}{n_{total}}$ (de los Campos *et al.*, 2013).

The conditional posterior densities of the BSSVS model are described in the Appendix (4.7.1). The additive genetic variance (σ_g^2) was estimated as the sum of the posterior mean variances explained by the SNPs (σ_{SNP}^2) and estimated variance of the polygenic effect included in the BSSVS model (σ_u^2), where $\sigma_{\text{SNP}}^2 = \sum_{j=1}^{n_{total}} \alpha_j^2$. The BSSVS model was implemented using Gibbs sampling, using right-hand-side updating as described in Calus (2014), and was run in three chains per trait of 80 000 cycles, with the first 10 000 cycles disregarded for burn-in. Burn-in length was chosen based on a preliminary study using a similar dataset (Van Binsbergen *et al.*, 2014b). Convergence of the BSSVS model was monitored by plotting the total SNP variance for each cycle of the Gibbs sampler (See Appendix: Figure S4.1). For each trait, the results (variances and BLUPs) of three chains were combined.

Pedigree BLUP

For comparison, BLUP based on pedigree information only was also performed. Following the notation above, the model was:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \mathbf{e}$$

Similar to GBLUP, the BLUP model was applied using ASReml 4 software (Gilmour *et al.*, 2014).

4.2.4 Prediction reliability

The reliability of genomic prediction was evaluated by assigning all 5503 bulls to either the training or validation set based on year of birth, according to the protocol used to validate genomic prediction in practice. Bulls born before 2001 (3416 bulls) were assigned to the training set and bulls born between 2001 and 2008 (2087 bulls) to the validation set. The validation animals were split into smaller subgroups (see below) to ensure that the number of animals in these subgroups was sufficient, and a relatively large number of validation animals were chosen. Reliability of genomic prediction was calculated for the validation animals as the squared correlation between de-regressed proof and the EBV for the different traits. Furthermore, the regression coefficient of the DRP on the EBV was calculated to evaluate the bias of predictions. A regression coefficient of 1 indicates no bias.

Persistency of the reliability of genomic prediction across generations was evaluated by splitting the validation bulls into three non-overlapping groups based on the presence of close relatives in the training set. The first group consisted of 1643 bulls with their sire and maternal grandsire in the training set (SMGS); the second group consisted of 113 bulls with their sire in the training set, but no maternal grandsire (SIRE); and the third group consisted of 329 bulls with no sire in the training set, but had one or both grandsires in the training set (GS). Two animals had no sire and no grandsires in the training set, and therefore were excluded from these analyses.

4.3 Results

4.3.1 Descriptive results

After editing SNPs for MAF and imputation reliability, the final BovineHD and ImputedHD genotype dataset consisted of 631 428 SNPs and the imputed sequence genotype dataset of 12 590 056 SNPs. In the final datasets, the average minor allele frequency (MAF) was equal to 0.27 with a median of 0.28 for the BovineHD dataset, 0.17 with a median of 0.13 for the ImputedHD dataset and 0.19 with a median of 0.16 for the imputed sequence dataset. The distribution of SNPs across the different classes of MAF is in Figure 4.1. Imputation reliability estimated by Beagle was on average 0.77 and ranged from 0.05 to 1.00, with a median of 0.89. Across prediction methods, the additive genetic variance (sum of polygenic and total SNP variance for BSSVS; SNP variance for GBLUP; polygenic variance for BLUP) ranged from 17.0 to 20.2 for SCS, from 15.9 and 19.6 for IFL and from 285.4 to 341.1 for PY (Table 4.1). As expected for de-regressed proofs, estimates of residual variance were very small, and therefore heritability estimates for all traits were close to 1 (Table 4.1).

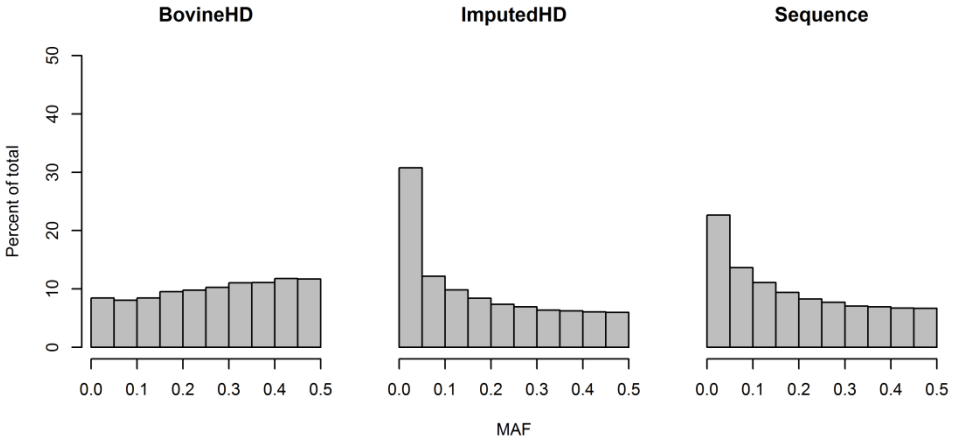

Figure 4.1 Distribution of minor allele frequencies (MAF) among 5503 individuals for different genotype panels

Table 4.1 Estimates of genetic parameters. Estimates of additive genetic variance (σ_g^2), heritability (h^2), regression coefficient (b), and prediction reliability (r^2) for somatic cell score (SCS), interval between first and last insemination (IFL), and protein yield (PY) using four types of data and two prediction methods.

Trait	Genotype data	Method	σ_g^2	h^2	b [1]	r^2 [2]
SCS	Pedigree	BLUP	20.22	0.97	1.00	0.33
	BovineHD	GBLUP	16.97	0.90	0.96	0.52
	BovineHD	BSSVS	18.55	0.95	0.99	0.52
	ImputedHD	GBLUP	17.41	0.93	1.00	0.50
	ImputedHD	BSSVS	18.37	0.98	1.05	0.51
	Sequence	GBLUP	17.09	0.93	1.03	0.49
	Sequence	BSSVS	18.82	0.98	1.04	0.50
IFL	Pedigree	BLUP	19.60	1.00	0.92	0.27
	BovineHD	GBLUP	15.90	0.94	0.83	0.39
	BovineHD	BSSVS	18.01	0.99	0.92	0.40
	ImputedHD	GBLUP	16.29	0.95	0.86	0.37
	ImputedHD	BSSVS	17.20	1.00	0.97	0.39
	Sequence	GBLUP	16.13	0.96	0.88	0.37
	Sequence	BSSVS	17.71	1.00	0.95	0.39
PY	Pedigree	BLUP	341.05	1.00	0.82	0.26
	BovineHD	GBLUP	295.05	0.94	0.86	0.47
	BovineHD	BSSVS	306.53	0.99	0.89	0.48
	ImputedHD	GBLUP	307.33	0.97	0.89	0.44
	ImputedHD	BSSVS	285.36	1.00	0.95	0.45
	Sequence	GBLUP	300.68	0.98	0.92	0.44
	Sequence	BSSVS	293.73	1.00	0.95	0.45

[1] Standard error of the regression coefficient ranged from 0.02 to 0.03

[2] Standard error of the prediction reliability was 0.02.

4.3.2 Prediction reliabilities

Prediction reliabilities ranged from 0.26 to 0.52 on average (Table 4.1). Overall, reliabilities were highest for SCS and lowest for IFL, except for pedigree-based BLUP, for which PY had the lowest reliability. For all traits, pedigree-based BLUP gave the lowest reliabilities and GBLUP performed less well than BSSVS. For both genomic prediction methods, reliabilities were highest when the BovineHD genotype data was used. Correlations between predicted breeding values using the different datasets and different genomic prediction methods were high and ranged from 0.95 to 1.00 (see Appendix: Table S4.1). For SCS, the coefficients of regression of the original phenotypes on the predicted breeding values were close to 1.00 (ranged from 0.96 to 1.05; Table 4.1). For IFL and PY, a slight overestimation of the breeding values was observed, since the regression coefficients ranged from 0.82 to 0.97 (Table 4.1). Using imputed sequence data, the overestimation for IFL and PY was less than when using BovineHD data, i.e. the regression coefficients were closer to 1.00. Plots of the de-regressed proofs versus the GEBV (for the two methods using the three types of data) for the 2087 validation animals and three traits are in the Appendix (Figure S4.2, Figure S4.3, and Figure S4.4).

To evaluate the reliability of genomic predictions across generations, the validation bulls were divided into groups based on the presence of (grand)parents in the training set: sire and maternal grandsire (SMGS); only sire (SIRE); no sire, but one or two grandsires (GS). As expected, in most cases, the SMGS group had the highest prediction reliability and the GS group the lowest (Table 4.2). Overall, across those groups, the largest decay in prediction reliability was found for IFL. Moreover, for IFL, the decay in prediction reliability was larger with both ImputedHD data and imputed sequence data (in both cases, the decay was equal to -35% for GS compared to SMGS) than with BovineHD data (-25% for GS compared to SMGS). For SCS and PY, this difference was much smaller (Table 4.2). Overall, there was no clear benefit of using sequence data on the persistency of reliability across generations, even when BSSVS was used.

Table 4.2 Estimated prediction reliability per pedigree group. Estimates of prediction reliability for somatic cell score (SCS), interval between first and last insemination (IFL), and protein yield (PY). Validation animals were divided based on the presence of relatives in the training set: sire and maternal grandsire (SMGS); only sire (SIRE); no sire, but one or two grandsires (GS).

Trait	Genotype data	Method	SMGS[1]	SIRE[2] (% of SMGS)	GS[3] (% of SMGS)
SCS	Pedigree	BLUP	0.35	0.33 (94%)	0.23 (67%)
	BovineHD	GBLUP	0.53	0.50 (95%)	0.45 (85%)
	BovineHD	BSSVS	0.53	0.51 (95%)	0.46 (86%)
	ImputedHD	GBLUP	0.51	0.52 (101%)	0.42 (83%)
	ImputedHD	BSSVS	0.52	0.52 (102%)	0.44 (85%)
	Sequence	GBLUP	0.50	0.53 (104%)	0.43 (85%)
	Sequence	BSSVS	0.51	0.53 (103%)	0.44 (87%)
IFL	Pedigree	BLUP	0.29	0.16 (55%)	0.15 (51%)
	BovineHD	GBLUP	0.40	0.34 (85%)	0.30 (75%)
	BovineHD	BSSVS	0.42	0.34 (80%)	0.31 (74%)
	ImputedHD	GBLUP	0.39	0.32 (81%)	0.25 (65%)
	ImputedHD	BSSVS	0.41	0.31 (75%)	0.27 (65%)
	Sequence	GBLUP	0.39	0.32 (83%)	0.25 (65%)
	Sequence	BSSVS	0.41	0.32 (78%)	0.27 (65%)
PY	Pedigree	BLUP	0.30	0.30 (101%)	0.24 (81%)
	BovineHD	GBLUP	0.48	0.48 (100%)	0.45 (95%)
	BovineHD	BSSVS	0.49	0.49 (101%)	0.45 (91%)
	ImputedHD	GBLUP	0.45	0.43 (96%)	0.41 (91%)
	ImputedHD	BSSVS	0.47	0.47 (100%)	0.41 (88%)
	Sequence	GBLUP	0.45	0.45 (99%)	0.42 (93%)
	Sequence	BSSVS	0.46	0.45 (98%)	0.42 (90%)

[1] Standard error of prediction reliability for the SMGS set was 0.02

[2] Standard error of prediction reliability for the SIRE set ranged from 0.06 to 0.08

[3] Standard error of prediction reliability for the SMGS set ranged from 0.03 to 0.05.

4.3.3 Individual SNP effects

For both genomic prediction methods, the (persistency in) reliabilities were highest when BovineHD genotype data were used compared to imputed sequence data. However, the additive genetic variances explained when imputed sequence data or BovineHD data was used were similar (Table 4.1). In Figure 4.2, Figure 4.3, and Figure 4.4, the individual SNP effects are plotted (as % of σ_g^2) for BSSVS using BovineHD data, ImputedHD data, and imputed sequence data. These Manhattan-plots do not show similar genome-wide association results as typically seen from single-SNP analyses. Instead, the Manhattan-plots represent the variances explained by a single SNP, conditional on fitting all other SNPs simultaneously. Therefore, SNP effects are much smaller than those obtained when only one SNP is fitted. Still, the Figures show that when BovineHD data and ImputedHD data are used for SCS and PY, it is possible to detect some regions on the genome that explain greater levels of variance, e.g. on chromosomes 15 and 22 (SCS) and chromosome 14 (PY). For BovineHD data, 26 SNPs had a SNP variance greater than 0.003%, with a maximum of 0.05%, most of these SNPs were located in a 1.8 Mb region at the beginning of chromosome 14. With imputed sequence data, no clear region could be detected with large SNP effects on the traits, but it should be noted that with imputed sequence data, there are 20 times more SNPs. For a fair comparison with BovineHD data, SNPs in the imputed sequence data were grouped in windows of 20 neighboring SNPs and the sum of the variances of the neighboring SNPs per window was plotted. However, still we did not detect any clear regions with an increased level of explained variance (results not shown).

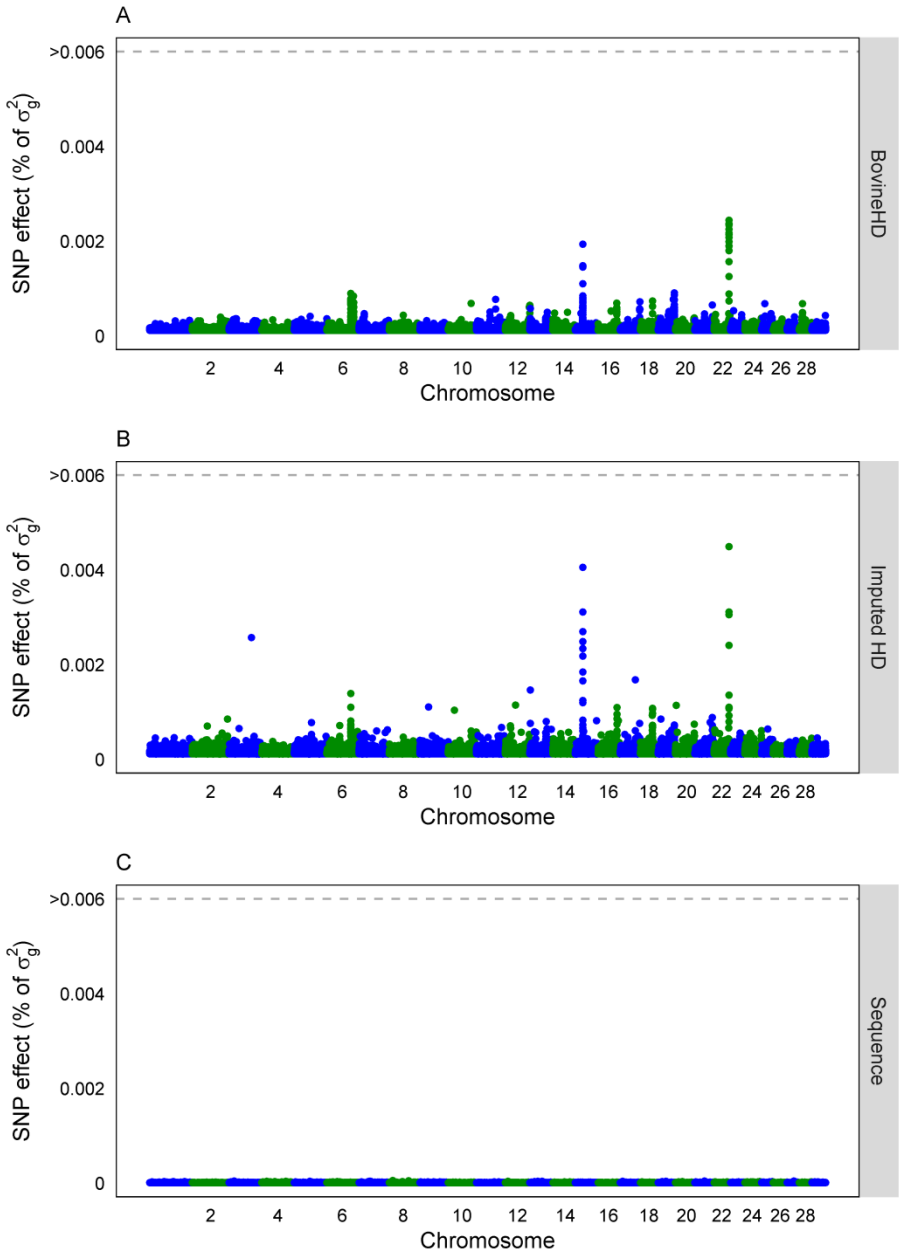

Figure 4.2 Manhattan plot with estimated SNP effects (% of σ_g^2) for somatic cell score (SCS) using the BSSVS model. Estimated SNP effects (% of $\boldsymbol{\sigma_g^2}$) based on the BSSVS model for somatic cell score using BovineHD data **(A)**, ImputedHD data **(B)**, and imputed sequence data **(C)**.

4.4 Discussion

Our objective was to investigate the reliability of genomic prediction based on imputed whole-genome sequence genotypes versus high-density SNP genotypes using real cattle data. Our hypothesis was that the use of sequence data in genomic prediction would result in higher reliability and higher persistency of reliability across generations. The rationale was that sequence data include the causal mutations that underlie QTL and that the effects of these mutations are estimated directly and not via effects of associated SNPs. This has been shown using simulated data (Meuwissen and Goddard, 2010; Clark *et al.*, 2011; Macleod *et al.*, 2013) but not yet with real data. Contrary to our expectation, the results did not show a higher (persistency of) reliability of genomic prediction using imputed sequence data compared to BovineHD SNP genotypes, although a relatively large training dataset with highly accurate phenotypes based on many daughters was used. While we did not expect a large gain in prediction reliability, we did expect to see a small gain as has been reported in studies comparing genomic prediction using 50 000 and 777 000 SNPs (Erbe *et al.*, 2012; Su *et al.*, 2012; Ertl *et al.*, 2014). Moreover, studies that used simulated whole-genome sequence data have claimed an increase in reliability using sequence data (e.g. Meuwissen and Goddard, 2010). The main improvement we expected was for persistency of reliability, when comparing the reliability observed in the lower related validation subset (GS) compared to more closely related validation sets. However, no increase in persistency of reliability was observed with imputed sequence data compared to the BovineHD data. In this study, our approaches were close to those used for genomic prediction in dairy cattle, including a training set of closely related animals, a pre-imputation step, and standard genomic prediction methods. Apparently, these approaches are not optimal to capitalize on the potential provided by sequence data. Below, we will discuss several factors that can explain this result.

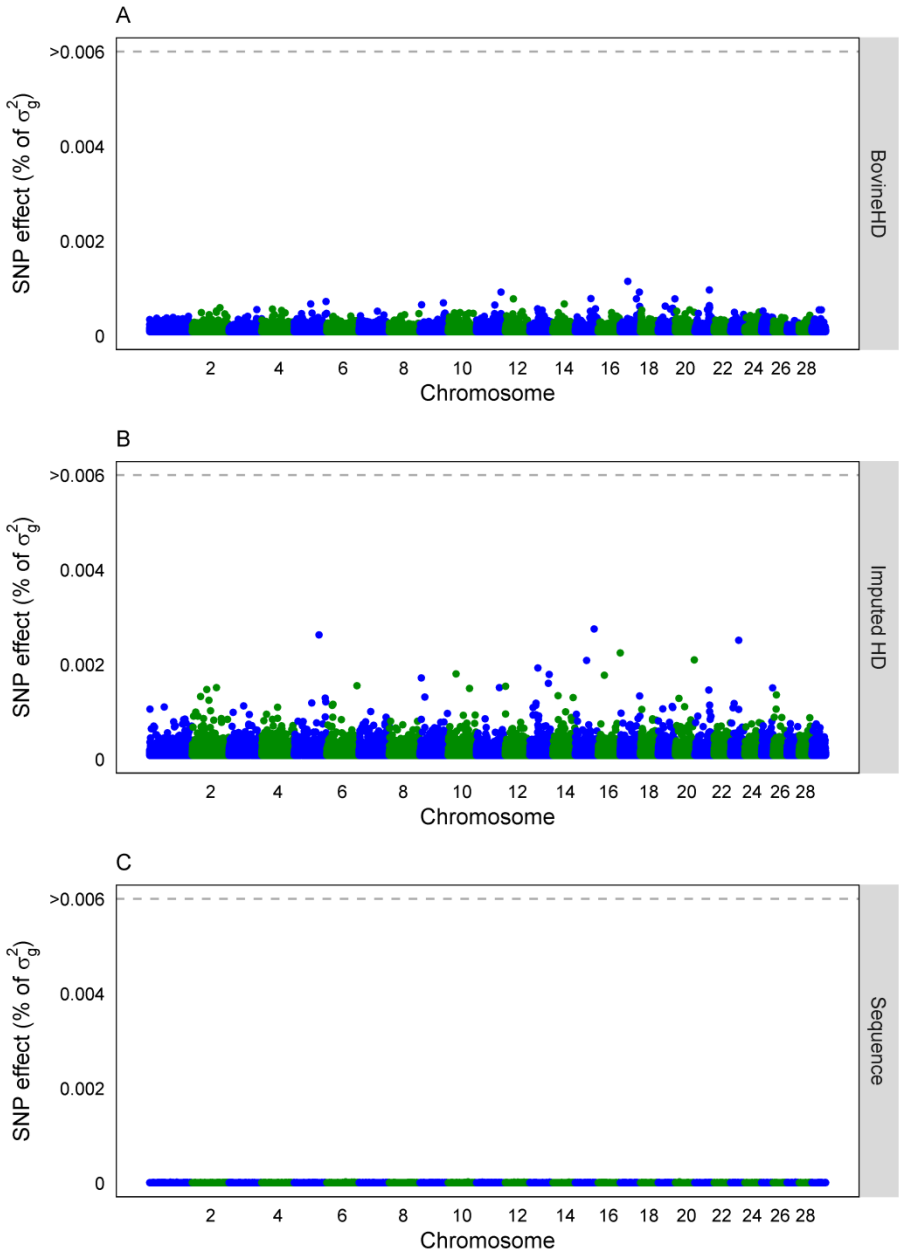

Figure 4.3 Manhattan plot with estimated SNP effects (% of σ_g^2) for interval between first and last insemination (IFL) using the BSSVS model. Estimated SNP effects (% of σ_g^2) based on the BSSVS model for interval between first and last insemination using BovineHD data **(A)**, ImputedHD data **(B)**, and imputed sequence data **(C)**.

4.4.1 Dataset

Our results did not show an advantage of using imputed sequence data compared to BovineHD genotype data for genomic prediction. Using imputed sequence data, both genomic prediction methods missed some QTL or QTL regions e.g. Figure 4.2, Figure 4.3, and Figure 4.4. A reason for this could be the structure of the dataset. Animals in the training set used in this study were closely related to each other. For example, the training set included 2878 father-son relationships. Close relationships between animals cause long range LD between a SNP and the QTL. Long range LD is useful for genomic prediction of animals that are closely related to those of the training population. However, when the aim is to find the precise location of QTL based on sequence data, long-range LD between the training animals is unfavorable, for instance to increase accuracy of genomic prediction across generations or populations. In a simulation study using dairy cattle data, it was concluded that using a training set with animals that have a low average relationship is beneficial for genomic prediction (Pszczola *et al.*, 2012). Altogether, a training set with less related individuals (e.g. multiple breeds) might be required to increase the advantage of using sequence data for genomic prediction. However, because of the way breeding programs operate currently and because relationships contribute significantly to prediction accuracy, in practice, it may not be possible to avoid this problem, other than by using training populations that include multiple breeds or lines.

In this study, 3416 individuals were used to estimate the effects of over 12 million SNPs. Thus, the number of SNPs (*p)* was much larger than the number of observations (*n*), which might be a second limitation of the current training set. With a dataset that is too small, the QTL effects might be estimated with too much error, which reduces the advantage of using sequence data compared to SNP genotypes for genomic prediction (Druet *et al.*, 2014). The Manhattan plots in Figure 4.2, Figure 4.3, and Figure 4.4 suggest that the effect of the potential QTL was spread across multiple SNPs. Increasing the number of individuals in the training dataset or pre-selecting SNPs based on other sources of information (Wimmer *et al.*, 2013) might be necessary to increase prediction reliability based on sequence data, as reported by Hayes *et al.* (2014). These authors obtained a very small increase of 2% in prediction reliability using imputed sequence data compared to BovineHD.

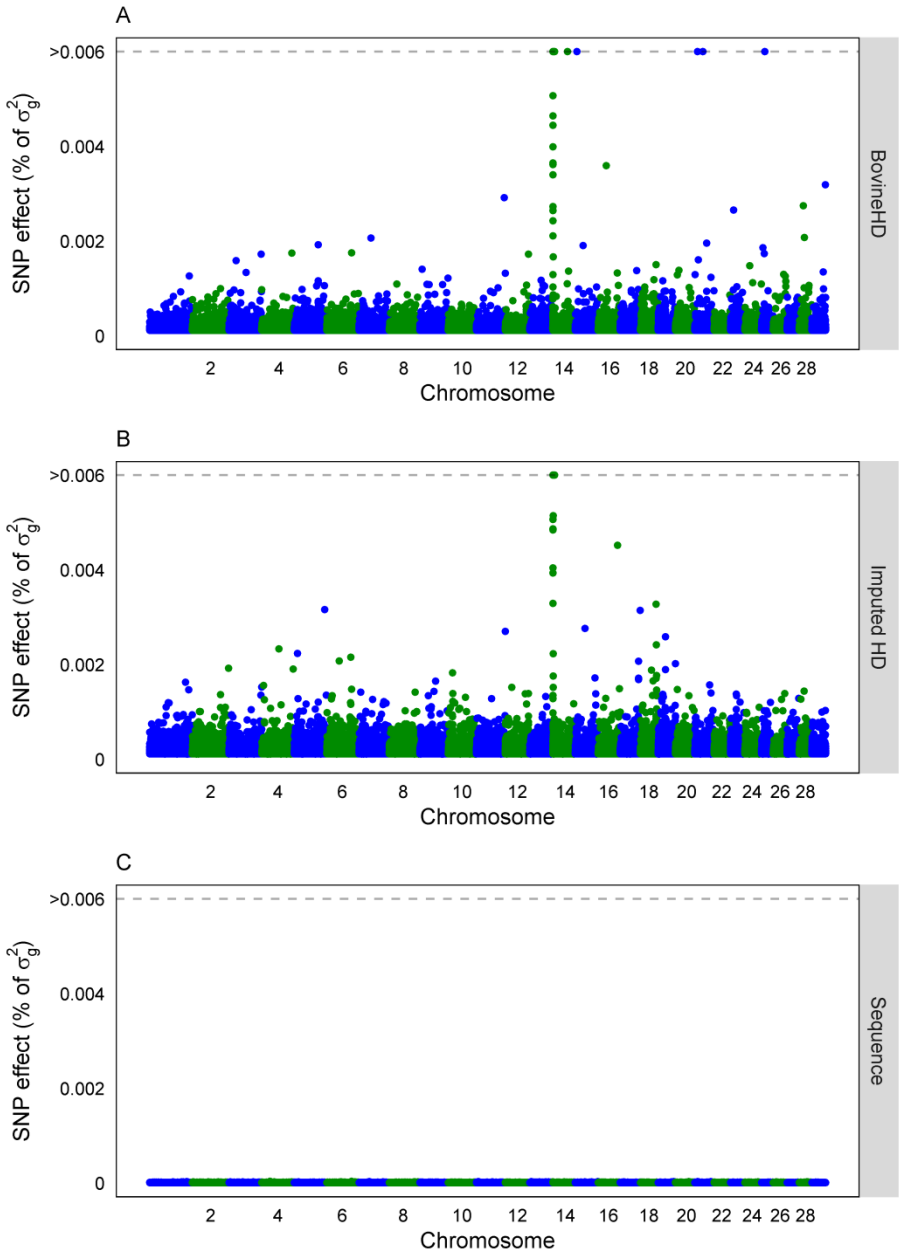

Figure 4.4 Manhattan plot with estimated SNP effects (% of σ_g^2) for protein yield (PY) using the BSSVS model. Estimated SNP effects (% of $\boldsymbol{\sigma_g^2}$) based on the BSSVS model for protein yield using BovineHD data **(A)**, ImputedHD data **(B)**, and imputed sequence data **(C)**.

However, they applied strict *a-priori* filtering steps for the SNPs and ended up with around 1.7 million variants, which is a factor 7 less than in our study. Also, their training set consisted of 16 214 bulls and cows, compared to the 3416 bulls used here. Thus, to benefit from the advantage of using sequence data compared to BovineHD genotype data for genomic prediction, it is necessary to aim for a large training set with a small average relationship between the animals, and possibly to pre-select SNPs based on functional information.

4.4.2 Pre-imputation step

Apart from the size and structure of the training dataset, the quality of the pre-imputation step could also impact the advantage of using sequence data for genomic prediction. To really benefit from imputed whole-genome sequence data compared to BovineHD data, imputation accuracy should be greater than the LD (measured as r^2) between a BovineHD SNP and the QTL. To test the possible effect of imputation, genomic prediction using a dataset of randomly selected SNPs from the imputed sequence data (ImputedHD) was compared with genomic prediction using the BovineHD dataset. Depending on the trait and method, a reduction of 0.01 to 0.03 in prediction reliability was found. A reduction in the reliability of GEBV with imputed genotypes has also been reported for studies on dairy cattle that used imputation from a few hundred SNPs to 50k SNPs, (e.g. Weigel *et al.*, 2010; Khatkar *et al.*, 2012; Mulder *et al.*, 2012; Segelke *et al.*, 2012; Chen *et al.*, 2014), which showed that the magnitude of the imputation errors was larger and the reliability of genomic prediction was lower compared to imputation from a 3k or 6k panel to a 50k panel. It has also been shown that the influence of imputation errors depends on the trait studied, e.g. traits that are influenced by a few large QTL were more affected than traits that are influenced by many QTL (Chen *et al.*, 2014). Moreover, van Binsbergen et al. [16] reported that the accuracy of imputation from BovineHD to sequence data ranged from 0.77 to 0.83 when the number of animals per breed ranged from 45 to 91. In this study, since 429 individuals from multiple breeds were used as reference animals, the accuracy of imputation was expected to be higher (Bouwman and Veerkamp, 2014; Brøndum *et al.*, 2014; van Binsbergen *et al.*, 2014a). Although the accuracy of imputation was relatively high, imputation errors will have some effect. However, based on the results with the ImputedHD data, we believe that the errors in the pre-imputation step were probably a small

factor in the reduction of the advantage of using sequence data compared to BovineHD data for genomic prediction.

The reason why imputation can reduce the accuracy of prediction is that imputed genotypes are called with increased uncertainty. In this study, SNPs that were likely to be imputed incorrectly were removed from the genotype dataset, using a low threshold of 0.05 for estimated imputation reliability to minimize the risk of removing potential causal mutations. With such a low threshold, there is still uncertainty about the genotype calling of imputed SNPs and potential causal mutations, although the mean imputation reliability was equal to 0.77. To take the effect of uncertainty in genotype calling on imputation accuracy into account, we considered the possibility of using the genotype probability instead of the most likely genotype for genomic prediction, which is expected to increase the reliability of genomic prediction (Mulder *et al.*, 2012). However, using genotype probabilities, saved as real or double precision values, would increase computation requirements by a factor 4 or even 8 compared to using the integer values (0, 1, and 2) used in our study. With the currently available resources, using genotype probabilities was not feasible.

4.4.3 Genomic prediction methods

A third reason why imputed sequence information did not improve prediction reliability could be parameterization of the BSSVS model. In the BSSVS model used here, we assumed that the prior distribution for α_j depended on the variance σ_α^2 and the QTL indicator I_j, which was sampled for each SNP taking a value of 0 if the SNP was included in the model with a small $\left(\frac{\sigma_\alpha^2}{100}\right)$ effect or 1 if the SNP was included with a large effect (σ_α^2). With imputed sequence data, each cycle included about 12 million SNPs with a small effect. Combined together, these small SNP effects might explain a very large part of the variance and, thus, the larger QTL remained undetected by the model. A way to decrease the variance explained by the SNPs with a small effect could be to include only SNPs with large effects and set all other SNP effects to zero as:

$$\alpha_j|\pi,\sigma_\alpha^2 = \begin{cases} 0 & when\ I_j = 0 \\ \sim N(0,\sigma_\alpha^2) & when\ I_j = 1 \end{cases}.$$

This model is also known as BayesC (Habier *et al.*, 2011). Compared to BSSVS, BayesC will save computing time, since, in each cycle, for a large proportion

of the SNPs, part of the calculations can be skipped as soon as I_j is sampled to be 0. Also, instead of two distributions, with large and (close to) zero effects, it might be useful to derive SNP effects from more distributions, which is done in methods such as BayesR (Erbe *et al.*, 2012).

It was assumed that both genotype panels had the same number of underlying QTL, i.e. the chosen π was larger for the imputed sequence dataset compared to the BovineHD dataset. However, due to LD between closely linked SNPs, the number of SNPs with a large effect might be larger for imputed sequence data than for the BovineHD data. Therefore, it might be better to use the same π for analyses using imputed sequence data as that for BovineHD analyses. Ultimately, the combination of the chosen π value and the parameterization of the model defines *a priori* the distribution of the effects (Daetwyler *et al.*, 2010), and thereby controls the posterior distribution of the effects. For instance, a study based on a 50k genotype dataset showed that the maximum SNP variances achieved with BSSVS with a π value of 0.999 were up to ten times as large as those achieved with BayesC with a π value of 0.9 (Calus *et al.*, 2014). To overcome this, π could be treated as unknown (Habier *et al.*, 2011).

Due to the computation requirements of genomic prediction applied to imputed sequence data, it was unrealistic to test many different settings and models. For example, with the BSSVS model, one chain of 80 000 cycles took approximately 85 days on a High Performance Linux cluster containing Intel(R) Xeon(R) CPU E5-2660 with a clock speed of 2.20 GHz. GBLUP was less time demanding (~6 hours), but required ~600 GB of RAM to store the genotypes. Due to efficient storing of the genotypes in the right-hand-side algorithm (Calus, 2014), the BSSVS model required less memory (~32 GB of RAM). These large computer requirements prevent fine tuning of the models used, but, at the same time, empirical studies have shown only small differences in prediction accuracy between available linear models (de los Campos *et al.*, 2013). The size of the training set used and the relationships between the individuals are probably more important factors than the choice of the model (de los Campos *et al.*, 2013). Therefore, it might be more beneficial to focus on the properties of the training set, than to test many different settings and models.

With 12 million SNPs, convergence of the Gibbs sampler can be rather low. Convergence of the BSSVS model was visually inspected by plotting the total SNP

variance for each cycle of the Gibbs sampler (see Appendix: Figure S4.1). The pattern of the estimated SNP variance components across the cycles appeared to be quite stable. For a simple check, EBV were also calculated after 40 000 cycles and 60 000 cycles. For the three traits analyzed here, the correlation between these EBV and the final EBV after 80 000 cycles was higher than 0.999 (results not shown). Based on these assessments, we believe that the model did converge and that the potential impact of Monte Carlo errors was probably small.

It should be noted that in contrast to the GBLUP model, the BSSVS model includes pedigree data and uses a spike-slab prior for the SNP-effects, i.e. priors are mixtures of two densities: one with small variance (the spike) and one with large variance (the slab). The GBLUP model was based on equally weighted markers and did not include the pedigree separately. Therefore, the comparison between BSSVS and GBLUP involves not only two different models but also two different input sets and this could make interpretation of the results difficult. However, we tested the GBLUP model by including a polygenic component for SCS using the three types of genotype data (see Appendix: Table S4.2). Due to the high correlation between the pedigree-based relationship matrix and genomic relationship matrix, the model had difficulties to converge. Including a polygenic component gave less residual error variance and therefore a slightly higher heritability. In addition to a higher heritability, the model also introduced more bias in predictions. However, prediction reliabilities were similar to those obtained with the GBLUP model without a polygenic component. Due to the convergence issues and similar prediction reliabilities, the GBLUP model without a polygenic component was used in this study.

4.4.4 SNP pre-selection

As shown in Table 4.1, predictions using imputed sequence data had similar additive genetic variance as predictions using the BovineHD data but, at the same time, the Manhattan plots using the sequence data in Figure 4.2, Figure 4.3, and Figure 4.4 did not reveal any regions with large effects. This suggests that the effect of the potential QTL was spread across multiple SNPs that were in high LD with the QTL. A way to overcome this problem is to pre-select SNPs based on annotation information or their putative regulatory role (Hayes *et al.*, 2014; Macleod *et al.*, 2014). Incorporation of this biological information has shown potential for the

detection of QTL (Macleod *et al.*, 2014) but did not result in higher reliability of genomic prediction (Hayes *et al.*, 2014). Improving the accuracy of this biological information might improve detection of QTL and also increase the prediction reliability (Hayes *et al.*, 2014).

To test if reliability of genomic prediction increased by giving certain SNPs a higher prior, we included some SNPs as fixed effects in the GBLUP model. For SCS, the three SNPs (on chromosomes 6, 15, and 22) that explained the most variance in the BovineHD analysis (Figure 4.2) were selected. For PY, a SNP in *DGAT1* (*diacylglycerol O-acyltransferase 1*) (Chr14:1802266) was selected, since *DGAT1* is known to have a major effect on milk production traits in Holstein Friesian cattle (Grisart *et al.*, 2002; Jansen *et al.*, 2013). For SCS, the prediction reliability did not change. However, for PY the prediction reliability increased from 0.47 to 0.51 for the BovineHD data and from 0.44 to 0.49 for the imputed sequence data. This suggests that pre-selecting SNPs and treating them as fixed effects or giving them a high prior might improve prediction reliability. However, this will be true only for SNPs that have a substantially large effect on the trait, such as *DGAT1*.

4.5 Conclusions

Our results did not show an advantage of using imputed sequence data compared to BovineHD genotype data for genomic prediction. To investigate whether using (imputed) sequence data compared to BovineHD genotype data can be an advantage for genomic prediction, the use of a large set of animals with small average relationships, along with other properties of the training set used, should be considered. Genomic prediction models that incorporate biological information of the SNPs, or use a stricter SNP pre-selection procedure, might also increase the advantage of using (imputed) sequence data for genomic prediction.

4.6 Acknowledgements

The authors want to acknowledge CRV and the 1000 bull genomes consortium for providing the data, and the Breed4Food project (program "Kennisbasis Dier", code: KB-12-006.03-004-ASG-LR) for financial support. MCAMB also acknowledges the financial support from Kennisbasis project KB-17-003.01-002 "Genomic breeding tools and databases".

4.7 Appendix

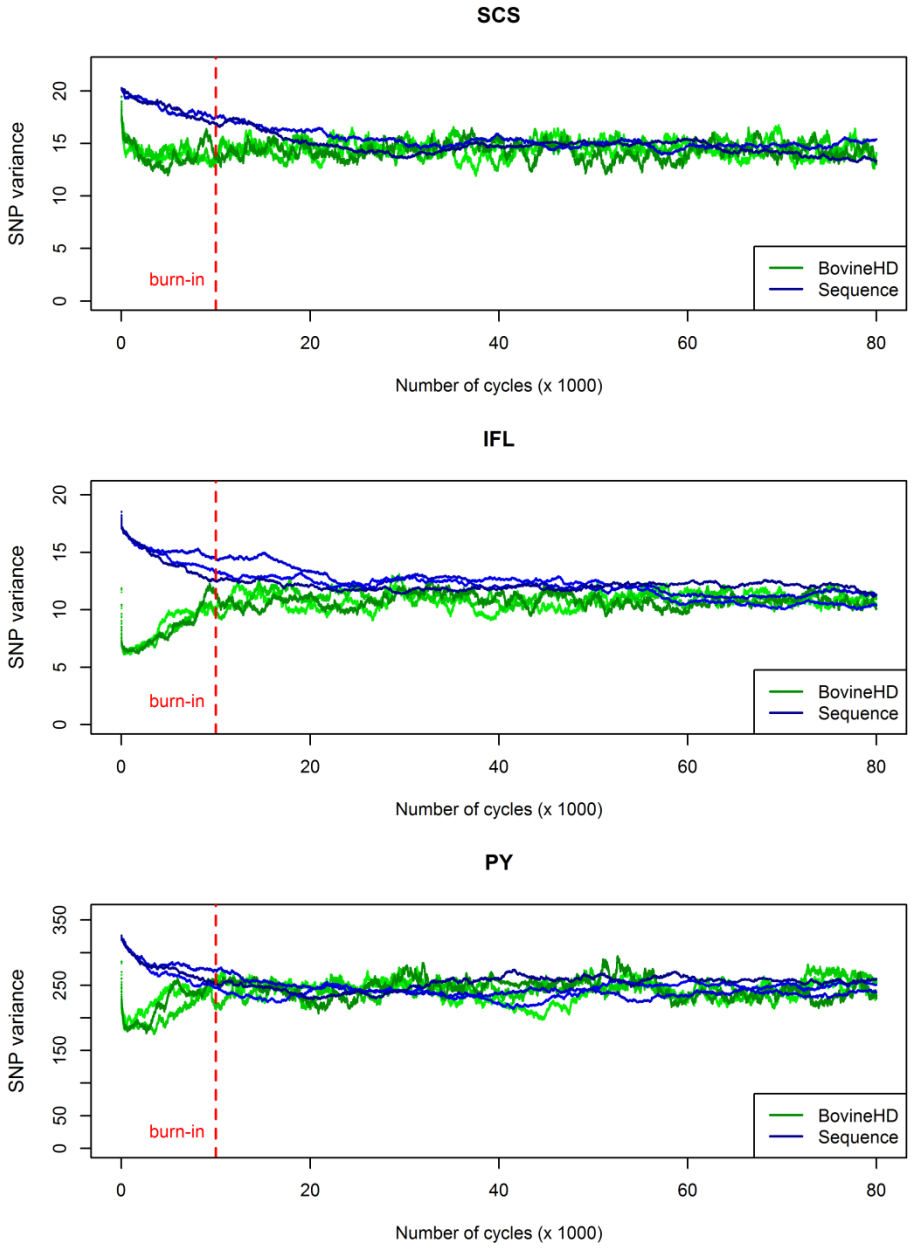

Figure S4.1 SNP variance components across cycles for the BSSVS model. Values are shown for somatic cell score (SCS), interval between first and last insemination (IFL), and protein yield (PY) for the three replicates using BovineHD data and imputed sequence data

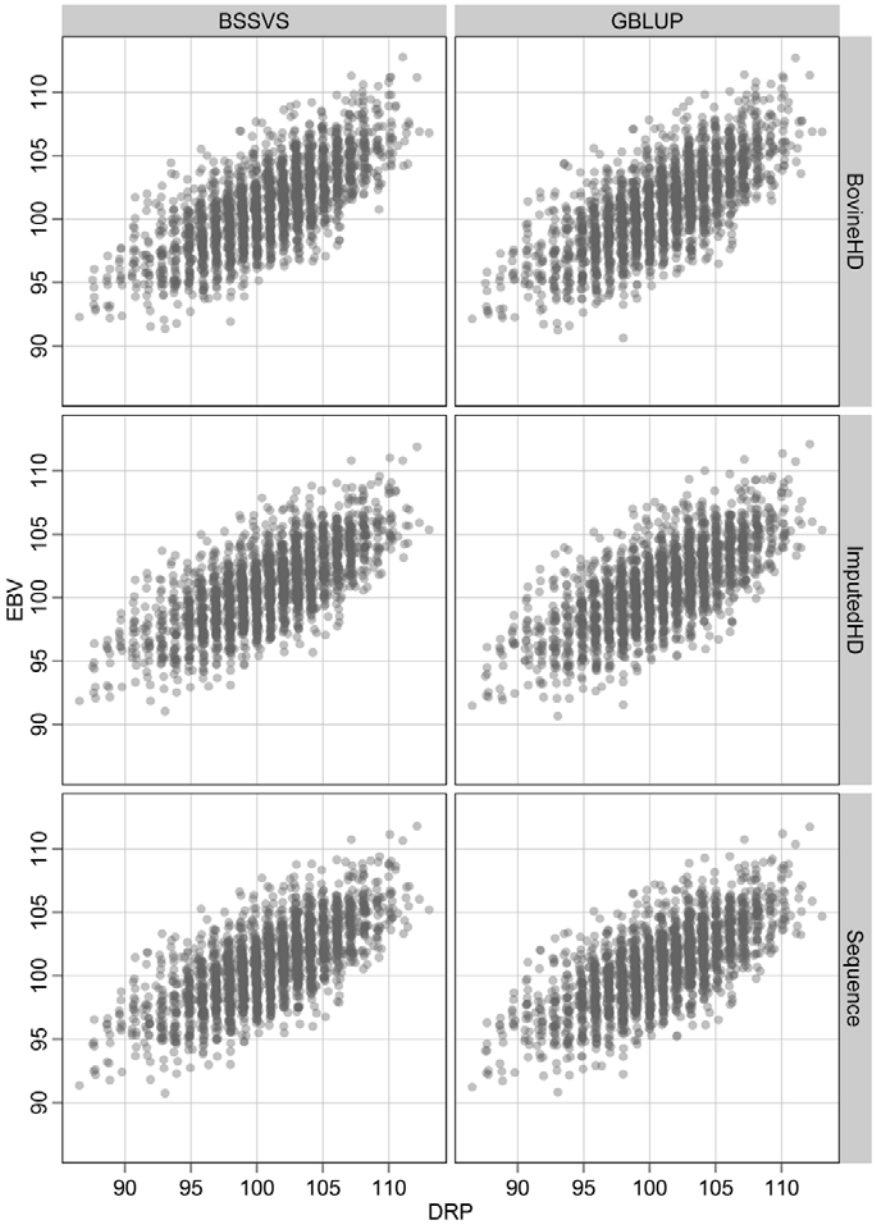

Figure S4.2 Original versus predicted breeding values for somatic cell score. Original de-regressed proofs (DRP) versus the estimated genomic breeding values (EBV) for the two methods (GBLUP and BSSVS) using the three types of data (BovineHD, ImputedHD, and imputed sequence) for the 2087 validation animals

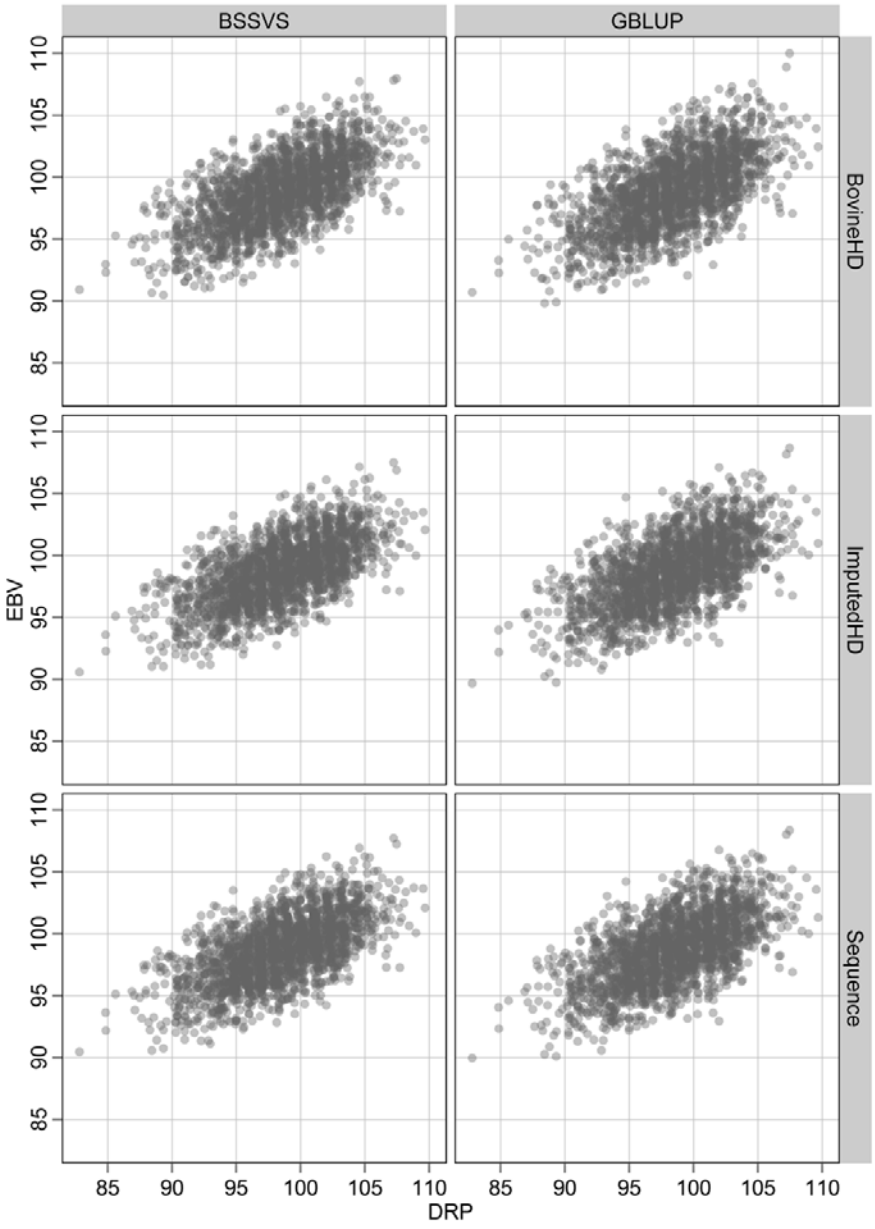

Figure S4.3 Original versus predicted breeding values for interval between first and last insemination. Original de-regressed proofs (DRP) versus the estimated genomic breeding values (EBV) for the two methods (GBLUP and BSSVS) using the three types of data (BovineHD, ImputedHD, and imputed sequence) for the 2087 validation animals.

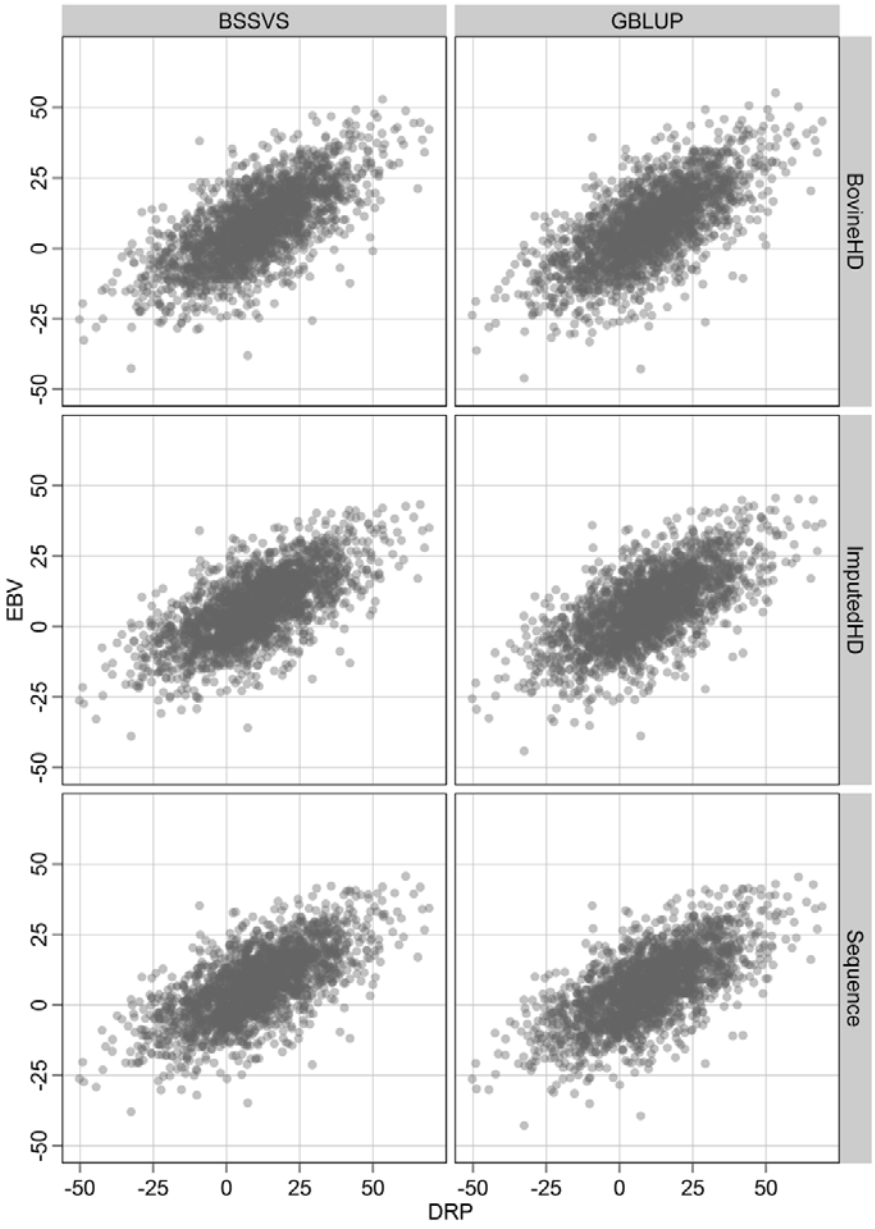

Figure S4.4 Original versus predicted breeding values for protein yield. Original de-regressed proofs (DRP) versus the estimated genomic breeding values (EBV) for the two methods (GBLUP and BSSVS) using the three types of data (BovineHD, ImputedHD, and imputed sequence) for the 2087 validation animals.

Table S4.1 Correlations between predictions of the different models

		Pedigree - BLUP	BovineHD - GBLUP	ImputedHD - GBLUP	Sequence - GBLUP	BovineHD - BSSVS	ImputedHD - BSSVS
SCS	BovineHD - GBLUP	0.751					
	ImputedHD - GBLUP	0.721	0.975				
	Sequence - GBLUP	0.720	0.973	0.996			
	BovineHD - BSSVS	0.792	0.995	0.969	0.967		
	ImputedHD - BSSVS	0.765	0.973	0.993	0.993	0.976	
	Sequence - BSSVS	0.765	0.974	0.992	0.995	0.977	0.998
IFL	BovineHD - GBLUP	0.700					
	ImputedHD - GBLUP	0.678	0.967				
	Sequence - GBLUP	0.680	0.967	0.996			
	BovineHD - BSSVS	0.795	0.978	0.948	0.949		
	ImputedHD - BSSVS	0.796	0.954	0.972	0.973	0.977	
	Sequence - BSSVS	0.771	0.961	0.979	0.983	0.977	0.997
PY	BovineHD - GBLUP	0.729					
	ImputedHD - GBLUP	0.717	0.965				
	Sequence - GBLUP	0.715	0.964	0.996			
	BovineHD - BSSVS	0.766	0.992	0.959	0.959		
	ImputedHD - BSSVS	0.758	0.959	0.988	0.989	0.968	
	Sequence - BSSVS	0.762	0.964	0.991	0.995	0.967	0.994

Table S4.2 Estimates of genetic parameters for GBLUP model including polygenic component. Estimates of additive genetic variance ($\boldsymbol{\sigma_g^2}$), heritability ($\boldsymbol{h^2}$), regression coefficient ($\boldsymbol{b}$), and prediction reliability ($\boldsymbol{r^2}$) for somatic cell score using three types of genomic data.

Data	$\boldsymbol{\sigma_g^2}$	$\boldsymbol{h^2}$	$\boldsymbol{b}$ [(1)]	$\boldsymbol{r^2}$ [(2)]
BovineHD	18.44	0.95	1.12	0.51
ImputedHD	19.00	0.99	1.14	0.49
Sequence	18.82	0.98	1.16	0.49

[1] Standard error of the regression coefficient ranged between 0.02 and 0.03

[2] Standard error of the prediction reliability was 0.02.

4.7.1 Conditional posterior densities BSSVS model

The conditional posterior density of α_j is:

$$N\left(\widehat{\alpha_j};\ \frac{\omega_j \widehat{\sigma}_e^2}{\mathbf{x}_j' \boldsymbol{D^{-1}} \mathbf{x}_j + \lambda_j}\right)$$

where $\widehat{\alpha_j}$ is the conditional mean of the allele substitution effect at locus j, $\lambda_j = \frac{\omega_j \widehat{\sigma}_e^2}{\widehat{\sigma}_\alpha^2}$, where $\omega_j = 1$ (if $I_j = 1$) or $\omega_j = 100$ (if $I_j = 0$). The conditional posterior density of σ_α^2 was: $\sigma_\alpha^2|\alpha \sim \chi^{-2}(\nu_\alpha + n, S_\alpha^2 + \boldsymbol{\omega}'\widehat{\boldsymbol{\alpha}}^2)$, where $\widehat{\boldsymbol{\alpha}}^2$ is a vector of squares of the current estimates of the allele substitution effects of all loci, that is weighted by vector $\boldsymbol{\omega}$. The conditional posterior distribution of I_j was:

$$\Pr(I_j = 1) = \frac{\mathrm{f}(r_j|I_j = 1)(1-\pi)}{\mathrm{f}(r_j|I_j = 0)\pi + \mathrm{f}(r_j|I_j = 1)(1-\pi)}$$

where $r_j = \mathbf{x}_j'\mathbf{D^{-1}}\mathbf{y}^* + \mathbf{x}_j'\mathbf{D^{-1}}\mathbf{x}_j\widehat{\alpha}_j$ where $\mathbf{y}^*$ are the conditional DRPs (i.e. for each SNP $\mathbf{y}$ minus the sum of the estimated SNP effects of the other SNPs), and $f(r_j|I_j = \delta)$ is the probability density function, giving the probability that $I_j = 0(\delta = 0)$ or $I_j = 1(\delta = 1)$, and is proportional to $\frac{1}{\sqrt{v}}\mathrm{e}^{-\frac{r_j^2}{2v}}$, where $v = \left(\mathbf{x}_j'\mathbf{D^{-1}}\mathbf{x}_j\right)^2 \frac{\sigma_{\alpha_j}^2}{\omega_j} + \mathbf{x}_j'\mathbf{D^{-1}}\mathbf{x}_j\sigma_e^2$. The conditional posterior density of σ_u^2 and σ_e^2 were inverse-χ^2 distributions, respectively $\sigma_u^2|u \sim \chi^{-2}(m-2, \mathbf{u}'\mathbf{A^{-1}}\mathbf{u})$ and $\sigma_e^2|e \sim \chi^{-2}(m-2, \mathbf{e}'\mathbf{D^{-1}}\mathbf{e})$, where m is the number of animals in the pedigree.

CHAPTER 5

UTILIZING WHOLE-GENOME SEQUENCE DATA TO INCREASE POWER OF QTL DETECTION IN TOMATO RECOMBINANT INBRED LINES (*S. LYCOPERSICUM* X *S. PIMPINELLIFOLIUM*)

RIANNE VAN BINSBERGEN[1, 2]
RICHARD FINKERS[3]
MARCO C.A.M. BINK[1,4]
MARIO P.L. CALUS[2]
RICHARD G.F. VISSER[3]
ROEL F. VEERKAMP[2]
FRED A. VAN EEUWIJK[1]

[1] Wageningen University and Research, Biometris, P.O. Box 16, 6700 AA Wageningen, the Netherlands
[2] Wageningen University and Research, Animal Breeding and Genomics, P.O. Box 338, 6700 AH Wageningen, the Netherlands
[3] Wageningen Plant Research, Plant Breeding, P.O. Box 386, 6700 AJ Wageningen, the Netherlands
[4] Hendrix Genetics Research, Technology & Services B.V., P.O. Box 114, 5830 AC Boxmeer, the Netherlands

Abstract

Statistical properties of procedures for mapping quantitative trait loci (QTL) in bi-parental populations are well known and it is straightforward to derive the power for detecting QTLs of a given magnitude for a particular marker density. Based on standard theory for linkage disequilibrium decay in bi-parental mapping populations, it is not expected that increasing marker density of marker panels with few thousand single nucleotide polymorphisms (SNPs) to whole-genome sequence data will provide additional power to detect QTLs. We challenged this expectation by investigating whether the use of whole-genome sequence data would reveal additional QTL, relative to a SNP array panel, when performing QTL analyses in a tomato population of recombinant inbred lines. We analysed a subset of a population of 100 recombinant inbred lines (RILs) that had phenotypic data for fruit weight and soluble solid content were available. For 49 individuals SNP array data (1,053 SNPs), and sequence data (430,431 SNPs) were also available. Despite the low number of RILs, more QTLs were found when using the sequence data, indicating that the linkage disequilibrium properties in the RIL population deviated from those expected. In the absence of sequence data for offspring RILs, a recommended strategy imputes sequence data of parents into offspring and then perform QTL mapping based on the imputed data.

Keywords: Imputation, SNP array panel; QTL mapping; Fruit weight; Soluble solid content; Recombination

5.1 Introduction

Tomato (*Solanum lycopersicum*) is an economically important crop across the world that in addition serves as an important model species for plant development, fruit ripening, and disease resistance (Menda *et al.*, 2013). For tomato breeding it is advantageous that crossing of cultivated tomato genotypes and wild relatives is relatively easy. Using a sufficient number of markers, QTLs can be detected by linkage between markers and QTLs (Lynch and Walsh, 1998; Hu and Xu, 2008; Takuno *et al.*, 2012). The correlation between a marker and QTL is related to the probability that an odd number of crossovers occurs between a marker and QTL. This probability is called recombination frequency (Lynch and Walsh, 1998). When it is assumed that crossovers occur randomly and are independent, the number of crossovers can be predicted from the observed recombination frequency by applying Haldane's mapping function (Haldane, 1919). However, this mapping function does not take into account interference between loci, i.e. presence of a crossover in one region affects frequency of crossovers in adjacent regions. For example, the Kosambi mapping function (Kosambi, 1944) allows for modest interference. These mapping functions have been applied in interval mapping approach to estimate QTL positions and effects.

Nowadays, SNP marker panels are available with a few thousand markers spread over the genome, for example for tomato (Sim *et al.*, 2012; Víquez-Zamora *et al.*, 2013). As the length of the genetic map of tomato is approximately 1,400 centiMorgan (Young *et al.*, 1988), those panels have on average more than one SNP per centiMorgan. This should be sufficient to locate recombination events and find QTL, especially in a bi-parental RIL population (Hu and Xu, 2008; Takuno *et al.*, 2012). Therefore, increasing SNP density to whole-genome sequence data is expected to give no additional power with respect to QTL mapping.

In contrast, Spindel *et al.* (2013) found additional QTL in a rice RIL population when using sequence data containing 30,984 SNPs compared to QTL mapping using a subset of 1,464 SNPs (including a previously unreported QTL for aluminium tolerance). This outcome triggers the question if the typical recombination patterns assumed by Haldane and Kosambi mapping functions are different from the true recombination patterns. With whole-genome sequence data all genetic variation in a population is captured, and therefore all recombination events are captured.

While with SNP array data information about these events is obtained from a limited number of markers and, for example, double crossovers in-between markers cannot be observed (Leu and Sen, 2014).

The benefit of QTL mapping for whole genome sequence data appears to be promising for a wide range of species. Like in the study of Spindel *et al.* (2013), also in outbred populations sequence data showed added value for fine mapping QTL compared to a SNP array panel, e.g. Arabidopsis (Alonso-Blanco *et al.*, 2016), human (The 1000 Genomes Project Consortium, 2010) and cattle (Daetwyler *et al.*, 2014).

Unfortunately, sequencing many individuals is still expensive. A cost effective approach is to use previous obtained SNP array genotype data, together with sequence data from a relatively small group of individuals and to use genotype imputation to obtain sequence data for large segregating populations. This is a frequently used approach in human genetics (Marchini and Howie, 2010) and livestock breeding (Cleveland and Hickey, 2013; Druet *et al.*, 2014; van Binsbergen *et al.*, 2014), but yet underutilized in plants. Several imputation algorithms that can handle large amounts of sequence data were designed for outbred populations (e.g. Browning and Browning, 2007; Li *et al.*, 2010; Howie *et al.*, 2011; Sargolzaei *et al.*, 2014; VanRaden *et al.*, 2015). However, these algorithms were not designed to deal with population structures and high inbreeding levels that are often present in plant populations. Only recently, software (PlantImpute) has been released that can perform genotype imputation from low- to high density marker panels in bi-parental populations (Hickey *et al.*, 2015).

With respect to RIL populations, an interesting approach would be to only sequence the parents of the population and use SNP array genotype data, together with genotype imputation to obtain sequence data information for the population of RILs. This imputed sequence data could then be used for QTL mapping. This two-step approach might be a good alternative for sequencing large number of individuals, but is only useful if sequence data have additional power in contrast to SNP array genotype data. Therefore, our objective was to investigate whether the use of sequence data would reveal additional QTL, relative to a SNP array panel, when performing QTL analyses in a tomato RIL population. We also investigated whether the use of imputed sequence data has additional power in comparison to a SNP array panel.

5.2 Materials and methods

5.2.1 Genotypic data

A population of 100 RILs was available from a cross between *Solanum lycopersicum* (cv. Moneymaker) and *Solanum pimpinellifolium* CGN15528 (Voorrips *et al.*, 2000). For 58 out of the 100 lines low-coverage sequence (Illumina HiSeq 2000) data for the F8 plants (Víquez-Zamora *et al.*, 2014) was available under study PRJEB6659 in the European Nucleotide archive. From the 7,786,934 variants only bi-allelic SNPs mapped on chromosome 1-12 with minor allele frequency ≥ 0.20, and sequencing coverage between 3 and 20 were kept. This upper limit for sequencing coverage was applied to remove outliers (more than 3 standard deviations from median), which could be due to sequencing or alignment errors. In a RIL population the number of heterozygotes is expected to be very low, i.e. close to zero. Therefore, 7 individuals with more than 50% heterozygous genotypes on one of the chromosomes were removed. For the other individuals heterozygous genotype calls were recoded to missing. Finally, SNPs with more than 20% of the genotype calls missing were removed from the dataset. After these edits 2,787,027 SNPs remained for 51 RIL individuals.

Parents. Together with the RIL population, also the *S. pimpinellifolium* (PIMP) parent was sequenced with a mean sequencing coverage of 7.1. The same edits as described before were done for this parent. Sequence data for the other parent (Moneymaker) came from a different dataset (PRJEB5235 in the European Nucleotide archive; The 100 Tomato Genome Sequencing Consortium *et al.* (2014)), and contained fewer variants (315,351) at higher sequencing coverage (mean depth was 33.7). Mapping was done against the *S. Lycopersicum* cv. Heinz SL2.40 reference genome sequence. Due to the close relationship between Moneymaker and Heinz, relatively few variants were found compared to the PIMP parent and the RIL population. Again only bi-allelic SNPs mapped on chromosome 1-12 were kept with mean sequencing coverage between 3 and 60 to remove outliers. After these edits 192,620 SNPs of the Moneymaker sequence remained.

In an ideal case both parents and the RIL population should have the same SNPs genotyped. Unfortunately, in the present study sequence data for the Moneymaker parent contained 192,620 SNPs, while sequence data of the PIMP parent and the RIL population contained 2,787,027 SNPs. Therefore, the missing

SNPs of the Moneymaker parent were imputed based on genotypes in the RIL population and the PIMP parent by following a few simple steps. First, per SNP segregating alleles in the RIL population and genotypes of the PIMP parent were defined. Then this information was used to construct the genotype of Moneymaker as the opposite genotype compared to PIMP parent. In case the genotype of the PIMP parent was missing, then also the genotype of Moneymaker was set to missing. The SNPs that were already present in the original Moneymaker sequence data were also imputed as described above to check the accuracy of the procedure.

Genetic map. To compare QTL mapping for SNP and sequence data we needed a genetic map for the sequence variants. No genetic map was initially available for the low-coverage sequence data. However, the F6 plants of the RIL population were also genotyped using a custom made SNP array as described by Víquez-Zamora *et al.* (2013). This SNP array panel contained 1,969 informative SNPs across the 12 chromosomes, and for 712 out of these 1,969 SNPs the genetic position (in cM) was available next to the physical position (in Mb). By using interpolation the genetic positions of all SNPs could be predicted

5.2.2 Genotype imputation

In practice, it could be possible that no sequence data are available for the whole RIL population, but only for the parents. In that case, it might be beneficial to impute the sequence data of parents into the RIL population (assuming the RIL population is genotyped using a SNP array panel). To test this approach, all sequenced genotypes of the RIL population were masked, except for the positions that were present on the SNP array panel. Then genotype imputation of the sequence data into the RIL population was performed using sequence data of the parents as reference.

Genotype imputation was performed using PlantImpute (Hickey, et al. 2015). The algorithm was used with default settings, except for the number of iterations per run that was set equal to 5 (default was 10), to decrease computation time. Furthermore, also because of computational limitations, only the chromosome arms were imputed and chromosome arms were split in multiple runs if the number of sequenced SNPs was greater than 30,000. These runs included 2,500 overlapping SNPs on sequence data and at least one overlapping SNP on the SNP array panel.

PlantImpute was run on a High Performance Linux cluster containing 48 slim nodes (64 GB RAM) containing 16 cores (Intel(R) Xeon(R) CPU E5-2660) with clock speed of 2.20 GHz, and two fat nodes (1 TB RAM) containing 64 cores (AMD Opteron(tm) Processor 6376) with clock speed of 2.30 GHz. For each run, 16 cores were allocated of 4 GB RAM each (64 GB in total). For a few more memory demanding jobs, 16 cores of a fat node were allocated of 10 GB RAM each (160 GB in total).

The software outputs genotype probabilities. A threshold of 0.9 was imposed to assign the most likely genotype for an individual at a specific position. If no genotype with a probability above 0.9 was present, then that position was coded as missing. Imputation quality was assessed for SNPs with non-missing genotypes, and expressed as the proportion of genotypes imputed correctly

5.2.3 QTL analyses

Our objective was to investigate the added power of sequence data relative to a SNP array panel when performing QTL analyses for fruit weight and soluble solid content. Phenotypic data was available for 98 out of 100 RIL individuals. The phenotypes were obtained during an experiment (Voorrips *et al.*, 2000) on F7 lines. The lines were grown in two plots with each four plants in a greenhouse trial (PPO, Naaldwijk, The Netherlands). The traits studied were average fruit weight (in grams) and soluble solid content (° brix), which were measured on the average number of five ripe fruits per plant.

The QTL analyses included all individuals with phenotypic data, SNP array panel data, and sequence data available. Only SNP genotypes on the short- and long arms of the chromosomes were used, as the excluded centromere region was expected to contain no recombination events and therefore was of less interest. The code used for single marker QTL analysis doesn't tolerate missing marker genotypes, however all datasets included some missing values. Therefore, a simple imputation script was used to fill in the missing genotypes by applying the following rules. Per individual a missing genotype for SNP *i* was imputed using the closest SNP (based on physical distance) without a missing genotype. The genotype of SNP *i* was assumed to be the same as this closest SNP genotype. After this step, SNPs with a minor allele frequency below 0.20 were removed from the data.

Linear regression. Single marker QTL analysis was performed with R (version 3.3.2), using an adapted code from Sikorska *et al.* (2013), for the three available genotype datasets: SNP array panel; original true sequence data; and imputed sequence data. The QTL analyses were performed by fitting the following linear regression model for each SNP:

$$\mathbf{y} = \mathbf{s}\boldsymbol{\beta} + \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

where $\mathbf{s}$ and $\mathbf{X}$ are the incidence vector and matrix relating the phenotypes to SNP (β) and cofactor (γ) effects, and $\boldsymbol{\epsilon}$ is a vector containing the residuals. The SNP genotypes were coded as the number of Moneymaker alleles (0 or 2). In our model, γ includes only the intercept (mean) and $\mathbf{X}$ includes only a column of ones. The initial plan was to perform multiple iterations where γ included also the effects of the QTL found in the previous genome scan. However, due to the low number of individuals no additional QTL were found in subsequent analyses.

To speed up the calculations, the variables $\mathbf{s}$ and $\mathbf{y}$ are transformed as given by the equations:

$$\mathbf{s}^* = \mathbf{s} - \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\mathbf{T}\mathbf{s}$$

$$\mathbf{y}^* = \mathbf{y} - \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\mathbf{T}\mathbf{y}$$

These transformations simplify the model to:

$$\mathbf{y}^* = \boldsymbol{\beta}\mathbf{s}^* + \boldsymbol{\epsilon}^*$$

With this model for each SNP the regression coefficient $\hat{\beta}$ can be computed as follows:

$$\widehat{\boldsymbol{\beta}} = \frac{\sum_{\mathbf{i}}^{\mathbf{n}} \tilde{\mathbf{y}}_{\mathbf{i}}^* \mathbf{s}_{\mathbf{i}}^*}{\sum_{\mathbf{i}}^{\mathbf{n}} \mathbf{s}_{\mathbf{i}}^{*\mathbf{2}} - \mathbf{n}(\bar{\mathbf{s}}^*)^{\mathbf{2}}}$$

where $\tilde{\mathbf{y}}_{\mathbf{i}}^{*}$ is the centered $\mathbf{y}^{*}$ for individual i, and $\bar{\mathbf{s}}^{*}$ is the mean SNP genotype after transformation. For the calculation of the p-values, it was assumed that the test statistic had a normal distribution.

Interval mapping. In addition to linear regression for single marker QTL analysis, exclusively for the SNP array panel, interval mapping was applied (Lander and Botstein 1989). Interval mapping can be interpreted as a simple approach to imputation using a regular mapping function like Haldane's or Kosambi's. The EM algorithm (Lander and Botstein, 1989) was used by the "scanone" function in R/qtl (Broman *et al.*, 2003). Genotype probabilities were calculated using a Haldane mapping function (Haldane, 1919).

5.3 Results

5.3.1 Genotype data

Sequence data were available for 51 RIL individuals and consisted of 2,787,027 SNPs after the edits on minor allele frequency, sequencing coverage, proportion heterozygotes, and proportion missing. There was variation shown in sequencing depth across individuals and across SNPs. On average the mean sequencing depth per individual (over all SNPs) was 6.8, with a range between 4.5 and 12.8. Mean sequencing depth per SNP (over all individuals) was on average 7.3, and ranged between 4.4 and 14.3. Across the genome there were no large differences shown in sequencing depth between the regions on the chromosomes, in other words, there were no regions with a much higher or lower coverage compared to the rest of the genome.

No genetic map was available for the low-coverage sequence data. Therefore, interpolation based on the SNP array panel was used to predict the genetic positions of the sequenced SNPs. In Figure 5.1 the interpolated genetic positions were plotted against the physical positions for 712 SNPs for which this information was available. Based on the regression pattern each chromosome was divided into three regions: short- and long arm (being recombination hotspots), and the centromere (recombination cold-spot). For the recombination hotspots an increase in genetic distance was shown with increasing physical distance. While for the centromere region no increase (or very small) in genetic distance was shown with increasing physical distance (flat line).

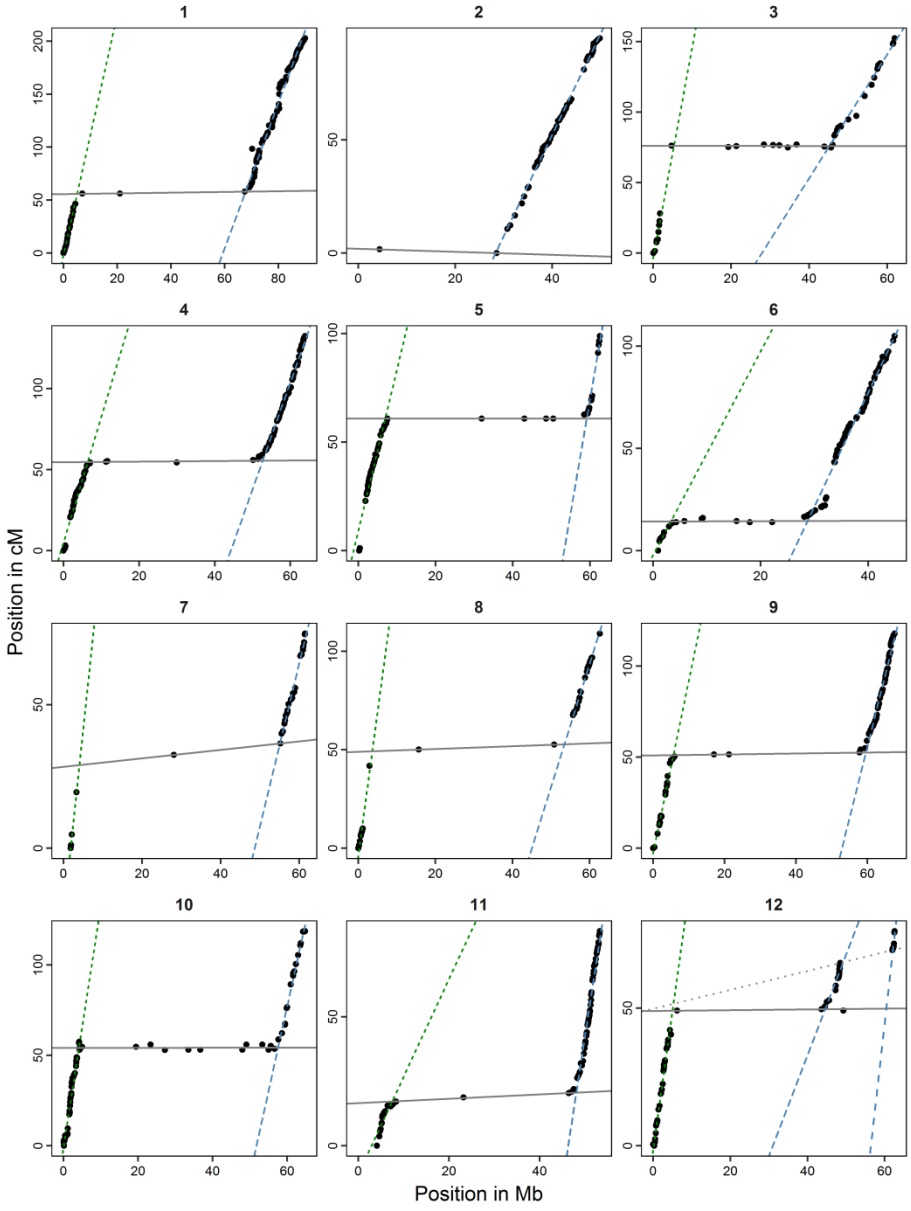

Figure 5.1 Genetic positions (in cM) versus the physical positions (in Mb) for 712 SNPs (black dots) on the SNP array. Based on the pattern each chromosome was divided into three regions: short arm, centromere, long arm. Per region linear regression was performed (lines), and the obtained regression coefficients were used to predict genetic positions for all SNPs

Chromosome 2 and 12 showed some deviation from this pattern: chromosome 2 did not contain a short arm and the long arm of chromosome 12 showed a pattern containing less recombination. This pattern was likely an assembly error in version 2.4 of the tomato genome as shown before (e.g. Víquez-Zamora *et al.*, 2013). Therefore, in the present study, this part of chromosome 12 was also regarded as centromere region. For each region on every chromosome a linear regression of the genetic positions on the physical positions of 712 SNPs from the array panel was performed (Figure 5.1). Based on the region-specific regression coefficients, the genetic positions of the sequenced SNPs were predicted.

After assigning every SNP to a specific region, it turned out that 83.4% of the sequenced SNPs were located in the centromere regions (Table 5.1 and Figure 5.2). This high percentage is in contrast to the SNP array panel, which consisted of 1,969 SNPs with 36.5% in the centromere regions Table 5.1 and Figure 5.3). When disregarding the centromere regions, the distribution of the SNPs was similar for SNP array panel and sequence data. The short arms contained 28.3% of the sequenced SNPs (excluding centromeres) and 29.0% of the SNPs on the SNP array panel (excluding centromeres), for the long arms these percentages were 71.7% and 71.0%, respectively.

The sequence data might contain sequencing errors, especially at low depth resulting in erroneous SNP calls, whereas it is expected that the calls from the SNP array contain less errors. To check the quality of the sequence data, the genotype calls of all 1,663 SNPs included in the sequence data and SNP array panel were compared for the 51 individuals that were genotyped for both. Across these individuals, 81.9% of the called genotypes were the same between the SNP array panel and sequence data; for 3.2% the opposite homozygote was called; and the other 14.9% were missing in one or both of the genotype datasets. These results show a good consistency in genotype calls between the SNP array panel and sequence data.

Table 5.1 Number of SNPs on sequence data, SNP array panel and number of SNPs in common between SNP array panel and sequence data per region on the chromosome (short arm (S), centromere (C), long arm (L)).

Chr.	Sequence				SNP array panel				Sequence and SNP array panel			
	S	**C**	**L**	**Total**	**S**	**C**	**L**	**Total**	**S**	**C**	**L**	**Total**
1	12,482	261,437	46,630	**320,549**	33	8	94	**135**	30	5	87	**122**
2	-	109,644	39,127	**148,771**	-	1	159	**160**	-	-	140	**140**
3	23,517	165,746	43,139	**232,402**	16	20	12	**48**	11	17	12	**40**
4	12,231	198,776	19,947	**230,954**	74	174	119	**367**	64	133	93	**290**
5	14,357	325,700	12,586	**352,643**	72	284	21	**377**	64	232	17	**313**
6	7,002	129,935	32,555	**169,492**	9	41	166	**216**	9	36	136	**181**
7	8,988	306,230	21,323	**336,541**	5	1	49	**55**	4	1	45	**50**
8	7,805	316,216	19,482	**343,503**	16	2	45	**63**	13	2	42	**57**
9	11,388	136,557	16,170	**164,115**	42	13	97	**152**	38	11	87	**136**
10	8,885	174,248	14,209	**197,342**	6	90	12	**108**	6	71	8	**85**
11	16,274	110,397	11,483	**138,154**	29	16	55	**100**	25	16	47	**88**
12	11,517	89,155	51,889	**152,561**	52	69	67	**188**	44	57	60	**161**
Total	**134,446**	**2,324,041**	**328,540**	**2,787,027**	**354**	**719**	**896**	**1,969**	**308**	**581**	**774**	**1,663**

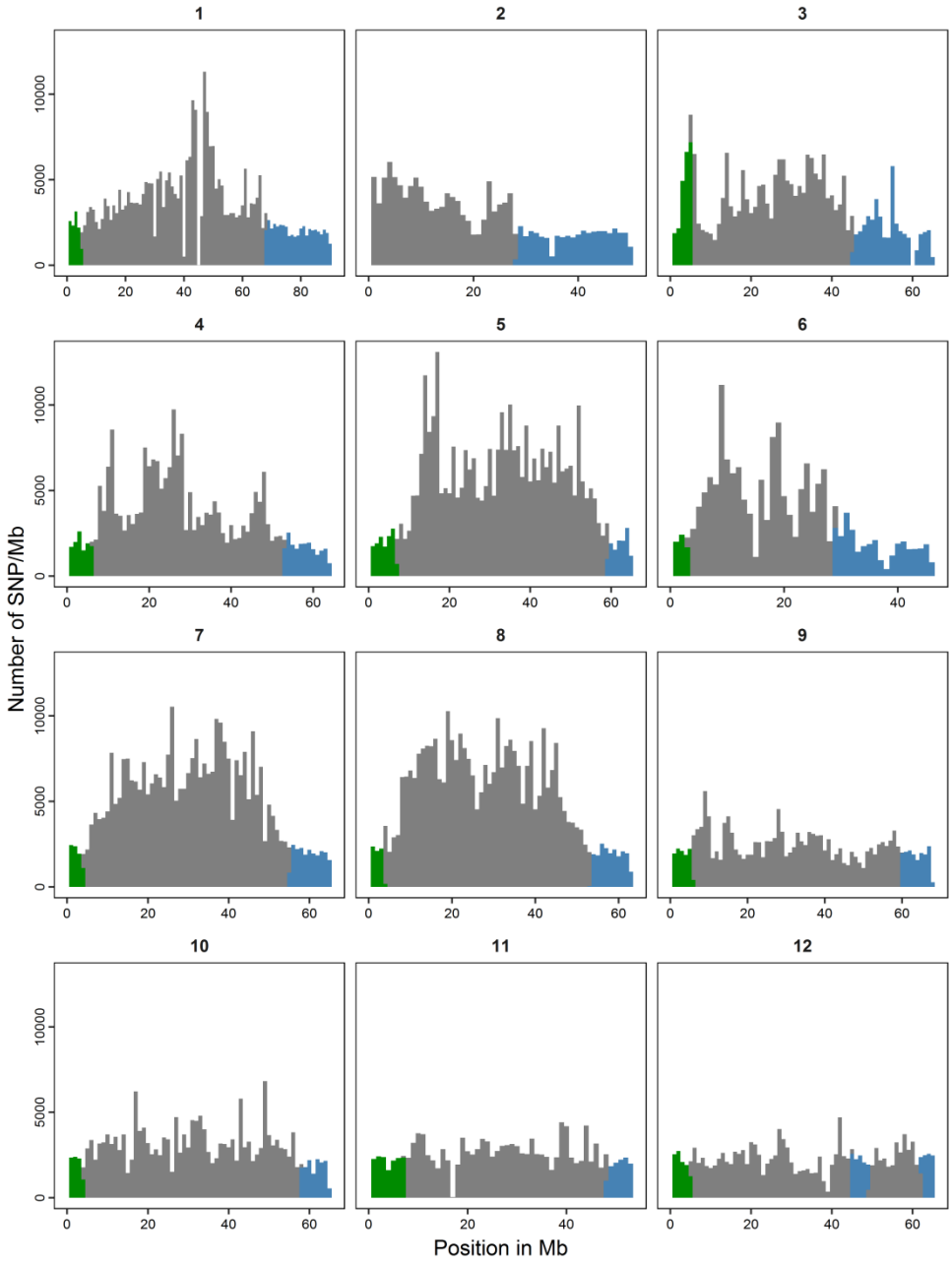

Figure 5.2 Frequency distribution of SNPs per 1 Mb region (physical position) across the chromosomes for sequence data. The colours represent the region on the chromosome: short arm (green), centromere (grey), long arm (blue)

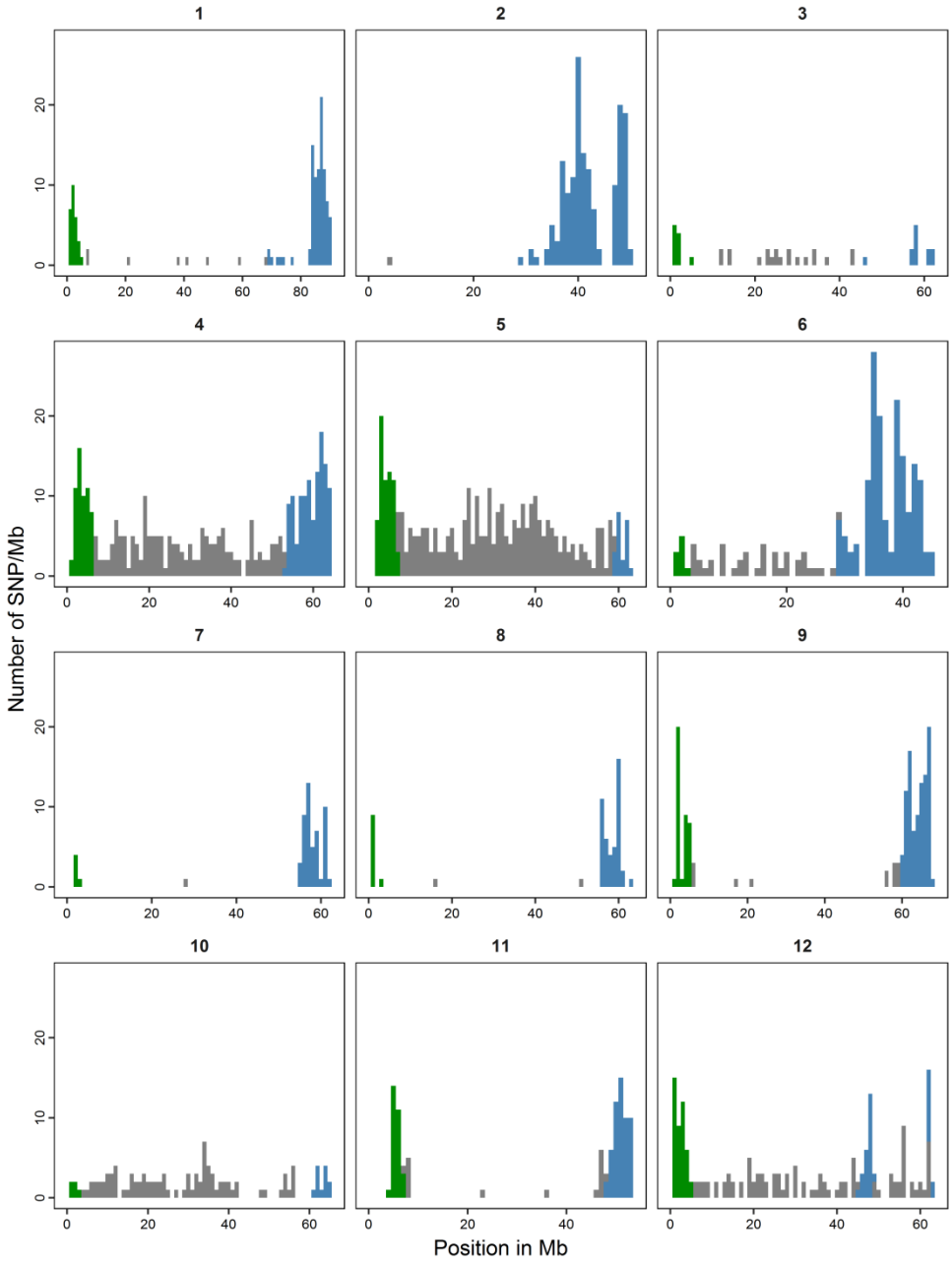

Figure 5.3 Frequency distribution of SNPs per 1 Mb region (physical position) across the chromosomes for SNP array panel. The colours represent the region on the chromosome: short arm (green), centromere (grey), long arm (blue)

5.3.2 Recombination events

To have insight in the recombination pattern of the sequence data, first the genotypes of both parents should be known. Sequence data of both parents were available, however as described before the Moneymaker sequence data contained less SNPs compared to Heinz than the PIMP parent and the RIL population compared to Heinz (due to close relationship of Moneymaker and the Heinz 1605 accession used to construct the reference genome). The Moneymaker sequence data contained 192,620 SNPs, whereas in the PIMP parent and the RIL population 2,787,027 SNPs were called. The 'missing' SNPs of the Moneymaker parent were constructed based on genotypes in the RIL population and the PIMP parent. Between the original Moneymaker sequence and constructed sequence dataset 37,348 SNPs were in common. Of these overlapping SNPs most (35,621 SNPs) were constructed correct; 1 SNP was constructed wrongly (opposite homozygote); 125 SNPs were constructed as homozygote but were heterozygote in the original sequence data; and 1601 SNPs could not be constructed. That implies the constructed SNPs were consistent for 99.6%, and therefore this constructed Moneymaker sequence data were used in subsequent analyses.

With the sequence data of both parents, the parental allele frequencies for the non-missing genotypes in the RIL population could be calculated. On average 50% of the non-missing alleles came from the PIMP parent and 50% from the Moneymaker parent. However, as shown in Figure 5.4 there is some variation across the genome. This variation was shown in both the sequence data and the SNP array panel, and the pattern for both genotype datasets was similar. Except for chromosome 7, where the SNP array panel did not contain any SNPs at the start and end of the chromosome. These regions did contain SNPs from the sequence data, and as shown in Figure 5.4 their allele frequencies deviate from the expected allele frequency based on the closest SNP on the SNP array panel.

Graphical genotypes including the SNPs on the chromosomal arms were shown in Figure 5.5. In the sequence data more recombination events were observed compared to the SNP array panel. For example, chromosome 3 contained a large region with no recombination on the SNP array panel, but with some recombination events in the sequence data.

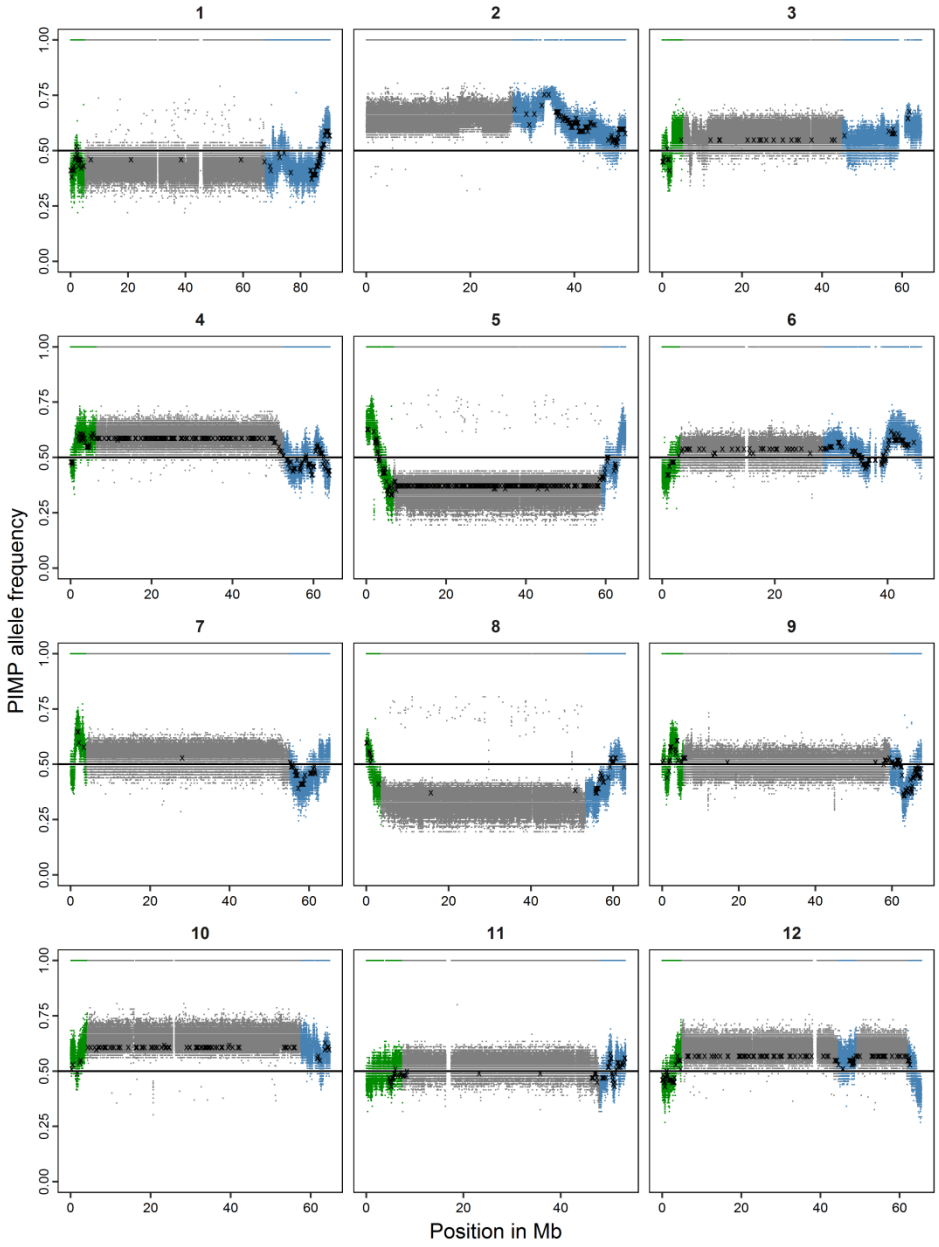

Figure 5.4 Allele frequency of the S. pimpinellifolium (PIMP) parent per SNP in the RIL population for the sequence data (coloured dots) and SNP array panel (black crosses) versus the physical position on the chromosomes. The colours represent the region on the chromosome: short arm (green), centromere (grey), long arm (blue)

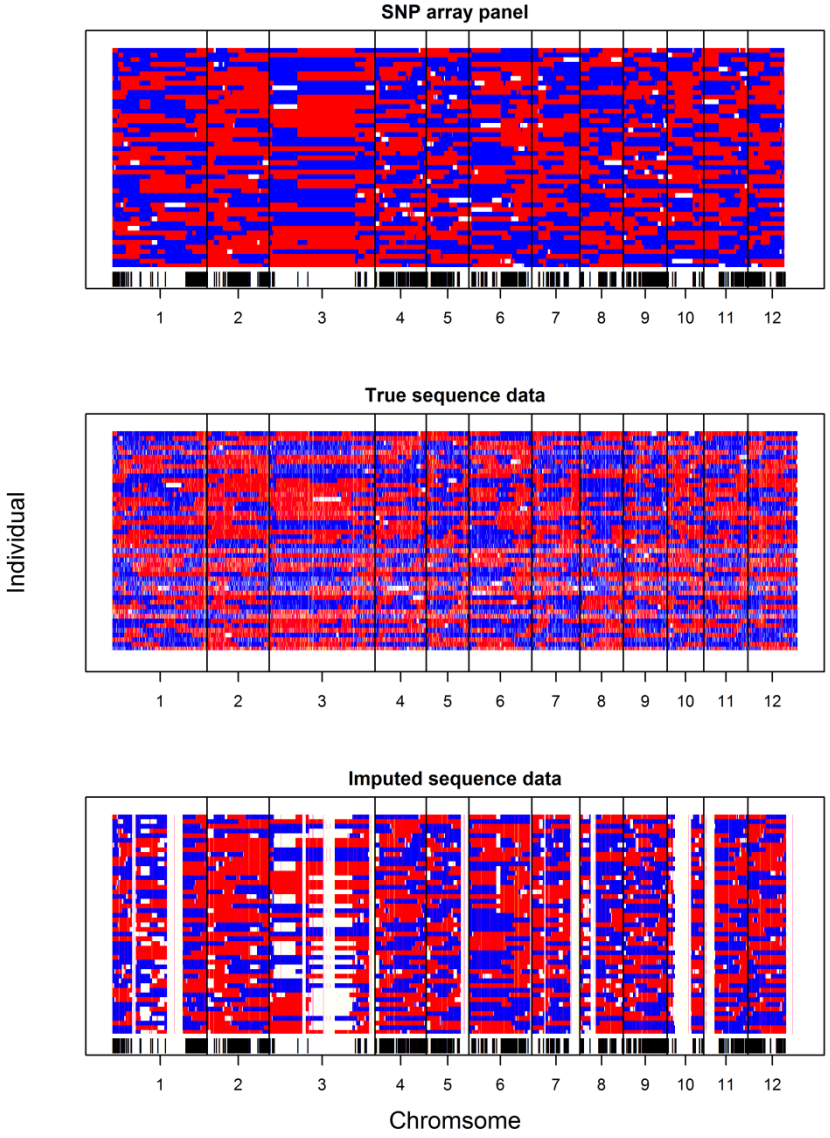

Figure 5.5 Graphical genotypes of the three datatypes: SNP array panel (1,053 SNPs), true- and imputed sequence data (430,431 SNPs). Each row presents an individual. Each position on the x-axis is one SNP in the sequence data. Only the short and long arms of the chromosomes are shown. Red indicates that the SNP genotype call was the same as S. pimpinellifolium, blue indicates that the SNP genotype call was the same as Moneymaker, and white indicates a missing genotype. The black lines below the plots represent the positions of the SNPs on the SNP array panel

5.3.3 Genotype imputation

In practice, it could be possible that no sequence information is available for the whole RIL population, but only for the parents. To simulate this situation, all sequenced genotypes of the RIL population were masked, except for the positions that were present on the SNP array panel. Then the masked positions were imputed with PlantImpute using the sequence data of the two parents as reference. In Table 5.2 the mean accuracy of imputation using PlantImpute is shown. Due to computational issues only the chromosomal arms were imputed and the largest arms (containing over 30,000 SNPs) were split in two separate runs. Similarly, the long arm of chromosome 12 was split in two runs; one run contained SNPs before the recombination cold spot and one run with SNPs after this recombination cold spot (Figure 5.1). In total 28 runs were performed. The number of SNPs per run ranged between 7,002 and 31,604 SNPs on sequence data, and between 2 and 109 on SNP array panel (Table 5.2).

Imputation quality was assessed by calculation of the proportion of genotypes imputed correctly of the SNPs with an assigned genotype (i.e. the most likely genotype with a probability above 0.9). Per region between 2.6% and 75.5% of the SNPs were not imputed, as there was no genotype with a probability above 0.9. In Figure 5.5 these regions with missing genotypes are clearly shown as white blocks. Most of these regions without imputed genotypes were on locations with (almost) no variation in the SNP array panel and with variation (recombination) in the sequence data, for example chromosome 3.

For the imputed SNPs, the proportion of genotypes imputed correctly was on average 0.989. Per run this proportion ranged between 0.879 (short arm chromosome 7) and 0.999 (short arm chromosome 9). In Figure 5.6 the proportion genotype imputed correctly per SNP is shown. For most SNPs all genotypes were imputed correctly, but there were some regions with poor accuracy of imputation. Most of these latter regions contained relatively few SNPs on the SNP panel or were at the beginning or end of a chromosomal region. For example, on chromosome 1 around 109 and 160 cM a clear decrease in proportion of genotypes imputed correctly was shown. Those regions contained only a few SNPs on the SNP array panel: at 103 cM, 119 cM and 165 cM. Likewise, the starts of chromosome 7 and 11 showed a decrease in proportion of genotypes imputed correctly. In both cases, the sequence data contained many SNPs (4521 SNPs and

Table 5.2 Per region (short arm (S) and long arm (L)) on each chromosome the number of SNPs on sequence data, the number of SNP in common between sequence data and SNP array panel, proportion SNPs imputed correctly (mean and standard deviation (SD)), the percentage of SNPs not imputed, and the number of SNPs with proportion imputed correctly below 0.75 and below 1.00. Due to computational limitations, for chromosome 1, 2, 3, and 6 the long arm was split in two parts (L1 and L2) including 2500 overlapping SNPs.

Chr.	Region	# SNPs sequence	# SNPs array	Mean	SD	% SNPs not imputed	# SNPs < 0.75	# SNPs < 1.00
1	S	12,482	30	0.997	0.018	4.5%	3	982
1	L1	25,000	6	0.990	0.025	14.4%	10	3,510
1	L2	24,130	81	0.993	0.026	30.0%	1	1,237
2	L1	25,000	88	0.997	0.013	3.1%	13	1,953
2	L2	16,627	67	0.992	0.015	2.6%	1	3,935
3	S	23,517	11	0.991	0.034	13.1%	2	1,686
3	L1	31,604	2	0.976	0.031	25.7%	5	10,548
3	L2	14,035	11	0.978	0.040	25.5%	0	2,332
4	S	12,231	64	0.997	0.011	3.5%	1	1,200
4	L1	19,947	93	0.998	0.009	3.2%	3	1,435
5	S	14,357	64	0.998	0.021	4.4%	21	606
5	L1	12,586	17	0.991	0.023	39.5%	2	1,559
6	S	7,002	9	0.994	0.022	4.0%	5	1,482
6	L1	25,000	109	0.997	0.024	3.2%	24	1,638
6	L2	10,055	36	0.986	0.029	8.1%	0	1,537
7	S	8,988	4	0.879	0.123	11.2%	2,065	5,760
7	L1	21,323	45	0.992	0.028	29.2%	4	1,630
8	S	7,805	13	0.988	0.024	12.4%	2	2,068
8	L1	19,482	42	0.995	0.015	16.2%	1	2,652
9	S	11,388	38	0.999	0.009	3.9%	2	349
9	L1	16,170	87	0.997	0.019	3.5%	6	1,564
10	S	8,885	6	0.956	0.074	41.3%	42	2,080
10	L1	14,209	8	0.987	0.029	48.5%	0	1,906
11	S	16,274	25	0.976	0.055	46.1%	1	2,282
11	L1	11,483	47	0.995	0.015	4.6%	2	1,821
12	S	11,517	44	0.995	0.021	3.9%	5	710
12	L1	10,848	20	0.994	0.016	4.0%	1	1,996
12	L2	8,486	11	0.992	0.017	75.5%	1	635

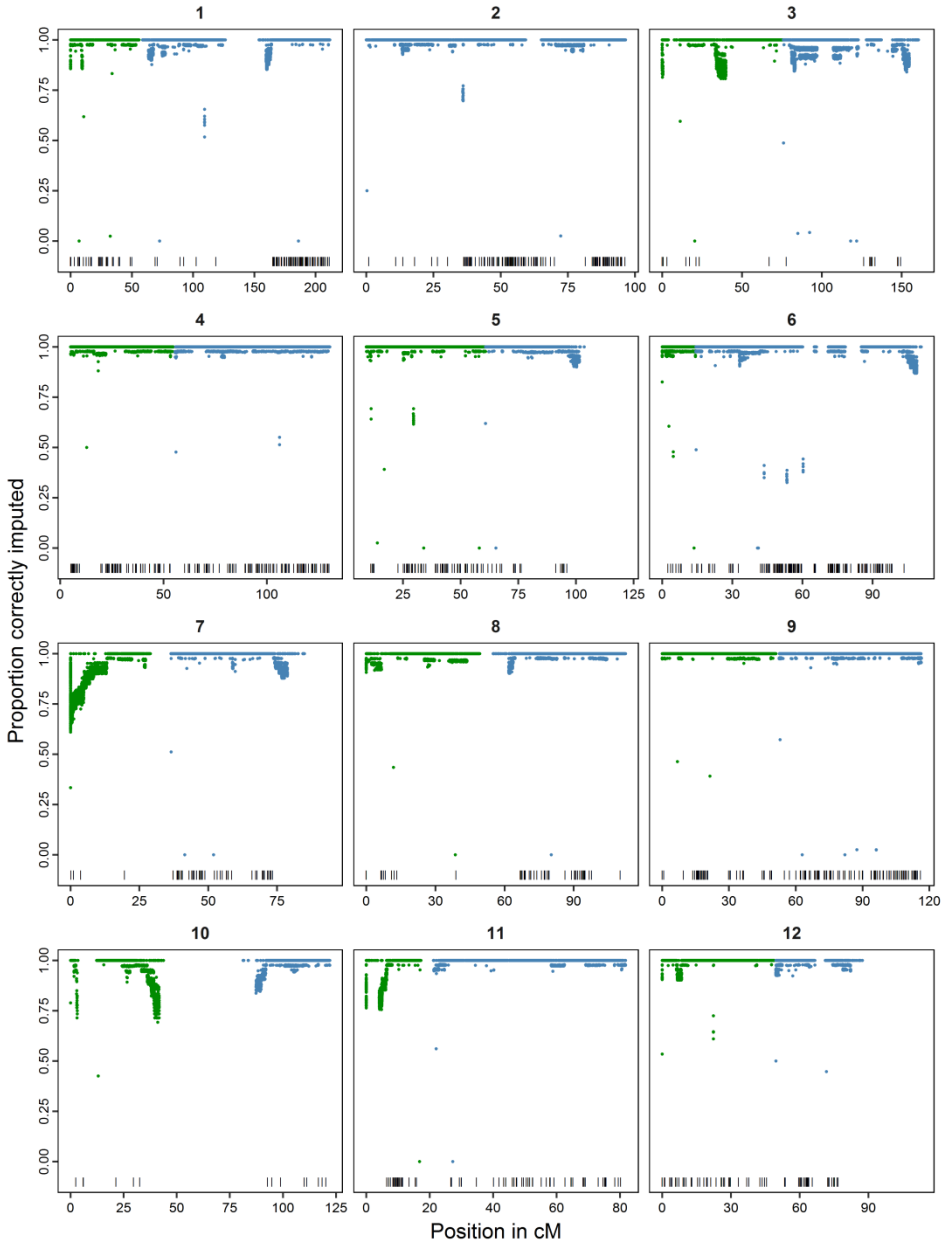

Figure 5.6 Proportion of genotypes imputed correctly per SNP versus the genetic position on the chromosomes. The colours represent the region on the chromosome: short arm (green) and long arm (blue). The black lines below the plots represent the positions of the SNPs on the SNP array panel

9749 SNPs for chromosome 7 and 11 respectively) before the first SNP on the SNP array panel. Of these SNPs, 117 SNPs (chromosome 7) and 7215 SNPs (chromosome 11) were not imputed or had a low proportion of genotypes imputed correctly. The average proportion of genotypes imputed correctly for the 4344 imputed SNPs in this region on chromosome 7 was 0.79 and 2145 SNPs had a proportion of genotypes imputed correctly below 0.75. For chromosome 11, the average proportion of genotypes imputed correctly for the 2534 imputed SNPs in this region was 0.92. Apart from the few poorly imputed regions, in general most SNPs were imputed correctly.

5.3.4 QTL mapping

For 49 out of the 51 sequenced lines measures were available for fruit weight and soluble solid content. Fruit weight measures ranged between 4.4 and 26.0 grams, and were on average 10.6 grams. Soluble solid content (measured in ° brix) was on average 6.9 and ranged between 4.8 and 9.4. QTL mapping results for SNP array panel, true sequence data, and imputed sequence data are presented below. These results were compared to previous found QTL (Grandillo and Tanksley, 1996; Tanksley *et al.*, 1996; Chen *et al.*, 1999).

SNP array panel. The final number of SNPs used for QTL analyses was 1,053 SNPs on the SNP array panel. For fruit weight two QTL were found on chromosome 9 and 11 (Table 5.3) using a linear regression approach. For soluble solid content four QTL were found on chromosome 1, 4, 7, and 12 (Table 5.4). These QTL were all nearby previously reported QTL (Table 5.3 and Table 5.4).

A common approach is to use interval mapping in addition to single marker linear regression. Therefore also interval mapping approach ("scanone" function in R/qtl (Broman *et al.*, 2003)) was performed using the SNP array data. With this approach two QTL were found for fruit weight on chromosome 10 and 12 (Table 5.3) and two QTL for soluble solid content on chromosome 7 and 12 (Table 5.4). In case of soluble solid content all the QTL were all nearby previous found QTL (Table 5.3 and Table 5.4).

True sequence data. Our main objective was to investigate the added power of the sequence data relative to a SNP array panel when performing QTL analyses. The final number of SNPs in the sequence data used for QTL analyses was 424,432. After linear regression with this true sequence data more QTL were found

compared to analyses with SNP array data. For fruit weight five QTL were found (Table 5.3 and for soluble solid content nine QTL were found (Table 5.4). All the QTL were near to previously found QTL (Table 5.3 and Table 5.4). The QTL found with SNP array data, were also found with sequence data.

Imputed sequence data. QTL analyses with imputed sequence data (357,849 SNPs) showed no additional QTL for fruit weight compared to SNP array data (Table 5.3). For soluble solid content another QTL was found on chromosome 3 compared to the SNP array data (Table 5.4). The QTL on chromosome 4 found with the SNP array data did not reach the threshold with imputed sequence data ($-log_{10}(p) = 2.05$).

Table 5.3 Chromosome, position on chromosome (in Mb and cM) and $-\boldsymbol{log_{10}(p)}$-value of QTL found for fruit weight. Three genotype datasets were tested using linear regression (LR) method: SNP array panel (SNP; 1053 SNPs); imputed sequence data (IMP; 357,849 SNPs); and true sequence data (SEQ; 424,432 SNPs). With the SNP array panel also an interval mapping (IM) approach was applied. In the last column previous found QTL around these locations are presented.

Data	Method	Chr.	Position (Mb)	Position (cM)	$-log_{10}(p)$	QTL
SNP	LR	9	3.8	33.5	2.77	*fw9.1* [(a)]; *FW9a* [(c)]
		11	50.0	43.3	2.87	*FW11a* [(c)]
SNP	IM	10	0.5	6.1	3.20 [(d)]	
		12	48.2	63.0	2.92 [(d)]	
IMP	LR	9	3.6	31.0	2.51	*fw9.1* [(a)]; *FW9a* [(c)]
		11	50.1	44.1	2.55	*FW11a* [(c)]
SEQ	LR	1	90.2	211.7	2.92	*fw1.2* [(b)]
		2	49.9	96.5	2.59	*fw2.2* [(a,b)]; *FW2a* [(c)]
		7	65.0	92.9	5.12	*fw7b* [(c)]
		9	3.8	32.9	2.93	*fw9.1* [(a)]; *FW9a* [(c)]
		11	50.2	45.5	2.92	*FW11a* [(c)]

[(a)] Tanksley *et al.* (1996); [(b)] Grandillo and Tanksley (1996); [(c)] Chen *et al.* (1999);
[(d)] = LOD-score

5.4 Discussion

Our objective was to investigate the added power of the sequence data relative to a SNP array panel when performing QTL analyses for fruit weight and soluble solid content in a tomato RIL population. The QTL analyses included 49 individuals with phenotypic data, SNP array panel data, and sequence data available. Despite the low number of individuals, the results indicate that sequence data were beneficial for QTL mapping, at least with respect to number of QTL found. First, the genotype data, including genotype imputation, will be discussed and thereafter the QTL analyses.

5.4.1 SNP distribution

The SNP array panel and sequence data differed in distribution of SNPs across the genome. For the sequence data 83.4% of the SNPs were located in the centromere regions, whereas for the SNP array panel 36.5% of the SNPs were located in the centromere regions. The SNP array panel was designed with knowledge on genetic linkage maps, typically capturing the locations of recombination hotspots, whereas the sequence data represent all physical positions along chromosomes that segregate in Moneymaker and PIMP compared to the Heinz reference genome. Only 37,348 SNPs out of the 2,787,027 SNPs called in the RIL population did also segregate in Moneymaker, which suggest that the other 2,749,679 SNPs in the RIL population came from differences in segregation between Heinz and the PIMP parent. A large number of SNP was found in the centromere regions. The PIMP genotype is different from the Heinz and Moneymaker and most differences will not be in the coding sequences because genes will be present in all three that make them tomatoes. However the non-coding sequences are very different and since they are mostly located in the centromere regions this might the reason why they are visible there.

5.4.2 Genotype imputation

In human, livestock, and some plant breeding studies genotype imputation is a frequently used approach to obtain high density SNP panel (or sequence data) of large number of individuals (Marchini and Howie, 2010; Hickey *et al.*, 2012; Druet *et al.*, 2014; van Binsbergen *et al.*, 2014). Instead of sequencing a whole-population, a cost effective approach could be to only sequence a part of the population or the parents and obtain sequence data of the whole population by imputation, as done

Table 5.4 Chromosome, position on chromosome (in Mb and cM) and $-log_{10}(p)$-value of QTL found for soluble solid content. Three genotype datasets were tested using linear regression (LR) method: SNP array panel (SNP; 1053 SNPs); imputed sequence data (IMP; 357,849 SNPs); and true sequence data (SEQ; 424,432 SNPs). With the SNP array panel also an interval mapping (IM) approach was applied. In the last column previous found QTL around these locations are presented

Data	Method	Chr.	Position (Mb)	Position (cM)	$-log_{10}(p)$	QTL
SNP	LR	1	3.2	34.3	3.41	*SSC1a* [(c)]
		4	62.8	122.0	2.62	*ssc4.1* [(a)]
		7	2.1	3.7	2.67	*SSC7a* [(c)]
		12	4.7	45.6	3.72	*ssc12.1* [(a)]; *SSC12a* [(c)]; *SSC12b* [(c)]
SNP	IM	7	33.2	19.6	4.03 [(d)]	*SSC7a* [(c)]
		12	46.8	45.6	3.33 [(d)]	*ssc12.1* [(a)]; *SSC12a* [(c)]; *SSC12b* [(c)]
IMP	LR	1	2.9	30.7	3.04	*SSC1a* [(c)]
		3	55.5	121.3	2.86	*SSC3b* [(c)]
		7	2.2	4.5	4.78	*SSC7a* [(c)]
		12	4.6	44.8	3.54	*ssc12.1* [(a)]; *SSC12a* [(c)]; *SSC12b* [(c)]
SEQ	LR	1	3.4	36.4	4.59	*SSC1a* [(c)]
		2	48.8	91.8	2.58	*ssc2.1* [(a)]
		3	46.4	81.6	4.29	*ssc3.2* [(a)]; *SSC3a* [(c)]
		4	62.8	122.5	2.56	*ssc4.1* [(a)]
		6	43.7	97.9	3.92	*ssc6.1* [(b)]; *SSC6a* [(c)]
		7	2.9	14.0	5.80	*SSC7a* [(c)]
		9	2.0	16.6	2.96	*SSC9a* [(c)]
		10	0.3	2.5	3.22	*SSC10a* [(c)]
		12	4.7	45.7	4.32	*ssc12.1* [(a)]; *SSC12a* [(c)]; *SSC12b* [(c)]

(a) Tanksley *et al.* (1996); (b) Grandillo and Tanksley (1996); (c) Chen *et al.* (1999);
(d) = LOD-score

in the present study. Results showed that on average >98% of the SNPs was correctly imputed. However, there were also regions were >10% of the SNPs did not have an assigned genotype (because the most likely genotype had a probability below 0.9) after imputation. In previous studies factors that influenced accuracy of imputation were distance to the first known SNP, and related to this, linkage disequilibrium between the SNPs (Hickey *et al.*, 2012; Bouwman and Veerkamp, 2014; van Binsbergen *et al.*, 2014). A larger distance between SNPs is closely related to lower linkage disequilibrium and to lower accuracy of imputation. In the present study, some of the poorly imputed regions contained relatively few SNPs on the SNP panel and therefore the distance between the imputed SNP and already known SNP apparently was too large. More SNPs on the SNP array panel could increase the accuracy of imputation.

The short arm of chromosome 7 showed poor imputation accuracy (Figure 5.6), because main sequence SNPs appeared before the first SNP on the SNP array panel at 0.12 cM. Only one flanking SNP on the array was present in this region, which makes accurate imputation difficult. Another reason might be the genetic map used in this study, as PlantImpute uses the genetic distance between markers in their model to predict transitions between states. In our study, the genetic positions of the sequenced SNPs were predicted by an interpolation based on the SNP array panel and the genetic position for 4488 out of the 4521 SNPs was predicted to be equal to 0 cM. In reality these genetic positions might differ, so if the true genetic positions were known, PlantImpute might perform better. Other reasons causing poor accuracy of imputation could be differences in expected allele frequency (Figure 5.4, differences in recombination pattern (Figure 5.5), sequencing errors or errors in the physical map. Imputation was assessed by the proportion of genotype imputed correct compared to the original true sequence data. As the original 'true' sequence data might suffer from high error rates (Nielsen *et al.*, 2011), the imputed genotypes might be considered as false, while they were in reality correct. This implies that the true proportion of SNPs imputed correctly might be higher than presented in this study.

Although PlantImpute performed well on accuracy of imputation, it was not computational efficient. The maximum amount of RAM ranged between ~3 GB for the short arm of chromosome 6 and 135 GB for the first part of the long arm of chromosome 3 (See Appendix; Table S5.1). The correlation between number of

SNPs on sequence data and the maximum amount of RAM was 0.86. However, without the regions that used ~64 GB RAM (the maximum allocated for those regions), this correlation increased to 0.98. The run time ranged between 2 and 33 hours and highly correlated with the number of SNPs on sequence (0.96). This performance was comparable to findings in another study where the memory consumption increased to more than 100 GB RAM, with increasing the number of SNPs to be imputed to 25,000 (Hickey *et al.*, 2015). So, both time and memory requirements increased dramatically with increasing numbers of SNPs on sequence data.

More computational efficient algorithms that are also based on Hidden Markov Models do exist, for example Beagle (Browning and Browning, 2007) and fastPHASE (Scheet and Stephens, 2006). These algorithms were primarily designed for populations that are outbred, e.g. human or livestock populations. Nevertheless, these algorithms have been used for inbred populations of crop species (e.g. Nazzicari *et al.*, 2016; Xavier *et al.*, 2016). In most of these studies these were used for imputing sporadic missing genotypes or augmentation of genotype-by-sequencing data. In such scenarios a lot of information about the phasing of haplotypes is available and those algorithms might perform well. However, in the present study much less information was available, as only parents had full sequence data. Beagle was tested here as well (with default settings, results not shown) and compared to PlantImpute it was much faster (max. 30 minutes per chromosome) and less memory intensive (max. 2 GB RAM per chromosome), but had a much lower accuracy. Since Beagle cannot take in to account the bi-parental population structure with only homozygous genotypes, it incorrectly imputed most genotypes as heterozygous.

5.4.3 QTL analyses

Two traits were studied: fruit weight and soluble solid content. Fruit weight is a highly polygenic trait and still little is known on the number and positions of underlying genes. In contrast, for soluble solid content many candidate genes and their function are known (Causse *et al.*, 2004; Stevens *et al.*, 2007; Pascual *et al.*, 2013). In Table S5.2 (Appendix) an overview is shown of candidate genes within 10 Mb of the QTL found in the present study. This table shows 15 candidate genes responsible for sugar transport, fructokinase, sucrose synthase, hexokinase,

invertase, and phosphoenolpyruvate carboxylase. Most QTL found in the present study were linked to a candidate gene.

The SNP panel used for QTL mapping in the present study contained 1,053 SNPs spread over the genome. With a total genome length of ~1,400 cM, on average one SNP every 1.3 cM was expected. This should be sufficient to locate recombination events and find QTL, and therefore increasing SNP density to whole-genome sequence data was expected to give no additional power with respect to QTL mapping. However, not all SNPs on the array were evenly spread over the genome. The average distance between two SNPs was 1.3 cM, but for 133 SNPs the distance to a neighbour SNP was > 2 cM, and for 25 SNPs this distance was even > 10 cM. For sequence data the average distance between two SNPs was 0.003 cM, for only 11 SNPs this distance was > 2 cM with a maximum distance between two neighbouring SNPs of 7.5 cM. The denser genome coverage with sequence data benefitted the QTL mapping results.

On chromosome 3 a large region was shown with many SNPs on sequence data, but almost no SNPs on the array. In this region an additional QTL for soluble solid content was located with (imputed) sequence data compared to the SNP array panel. When zooming in on this region, for many individuals two additional recombination events were shown in the area without SNPs on the array (Appendix; Figure S5.1). These regions, and QTL attached to these regions, can only be identified using the higher resolution from sequencing data.

Based on the results, sequence data performed best with respect to number of QTL regions found for both traits. Unfortunately, sequence data are not yet available for all populations and most populations are still being genotyped using lower density marker panels. A common approach is to use interval mapping to estimate QTL positions (Lynch and Walsh, 1998). In case of multiple markers, an appropriate mapping function, e.g. Haldane (1919), is needed to predict recombination frequency between markers. This is a better approach for QTL mapping compared to linear regression if a low number of markers are available. However, interval mapping is not needed with sequence data (all genetic variation is already there), and might even work disadvantageous, as potential QTL might be missed due to smoothing across the genome. Interval mapping with SNP array data resulted in different QTL for fruit weight compared to linear regression. This might have to do with the low number of individuals used in the study and the highly

polygenic nature of fruit weight, which makes it more difficult to map QTL with high power. For soluble solid content the QTL found with interval mapping were also found with linear regression using SNP array data, true sequence data, or imputed sequence data.

Essentially, this interval mapping approach is a simple and fully regular type of marker genotype imputation based on a Haldane mapping function. However, with genotype imputation (as applied in the present study) extra high resolution information of the parents was included in the data. This extra information helped to better locate recombination events on the genome, e.g. as described before on chromosome 3. In this example, imputed sequence data performed second best with respect to the QTL found. The relative performance of imputed sequence data was closer to the performance of linear regression using SNP array panel than true sequence data. Based on these results imputation of sequence data of parents into the RIL population (assuming the RIL population is genotyped using a SNP array panel), and use this imputed data for QTL mapping is a good cost-effective alternative for QTL fine mapping if no true sequence data available.

5.4.4 Conclusions

With sequence data additional recombination events on the genome were found, compared to SNP array data. As a result, despite the low number of individuals, the sequence data was beneficial for QTL mapping, at least with respect to number of QTL found. When no true sequence data is available, a two-step approach by first imputing sequence data of parents into a RIL population, and then use this imputed data for QTL mapping is a good cost-effective alternative for QTL fine mapping.

5.5 Acknowledgements

This study was financially supported by Breed4Food (BO-22.04-011-001-ASG-LR), a public-private partnership in the domain of animal breeding and genomics. The authors want to acknowledge Jérémie Vandenplas and Carl Nettelblad for their help with running PlantImpute.

5.6 Appendix

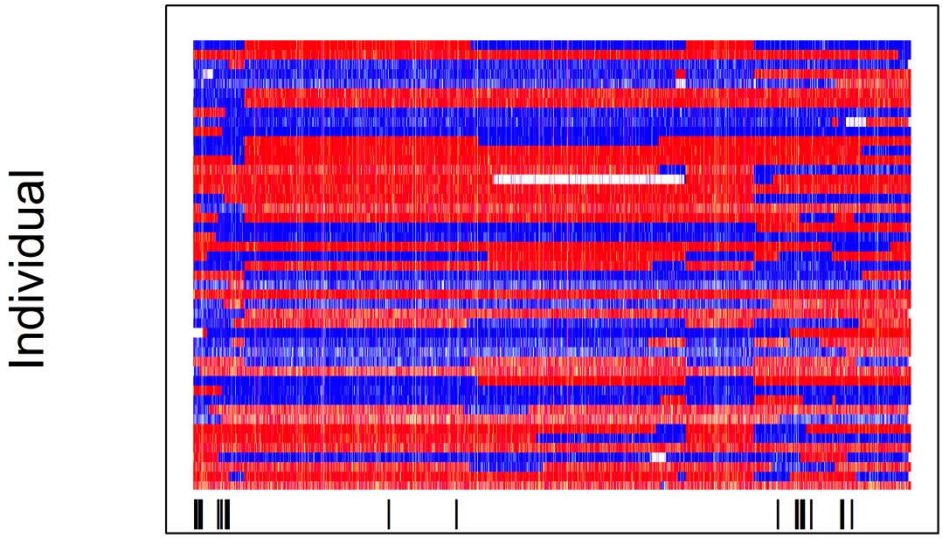

Figure S5.1 Genotype heatmap of chromosome 3 for true sequence data. Each row presents an individual. Each position on the x-axis is one SNP in the sequence data. Only the short and long arm of the chromosome is shown. Red indicates that the SNP genotype call was the same as *S. pimpinellifolium*, blue indicates that the SNP genotype call was the same as Moneymaker, and white indicates a missing genotype. The black lines below the plots represent the positions of the SNPs on the SNP array panel.

Table S5.1 Computational requirements PlantImpute per region (short arm (S) and long arm (L)) on chromosome. In the third and fourth column the number of SNPs on sequence data and the number of SNP in common between sequence data and SNP array panel are presented. The fifth column present the maximal amount of RAM allocated, and the sixth column present the maximum amount of RAM used. The last column present the total time needed.

Chr.	Region	# SNPs sequence	# SNPs array	RAM allocated (GB)	Max. RAM (GB)	Time (hh:mm)
1	S	12,482	30	64	54.5	05:59
1	L1	25,000	6	160	107.4	20:51
1	L2	24,130	81	64	64.3	24:44
2	L1	25,000	88	64	64.2	22:43
2	L2	16,627	67	64	64.3	10:26
3	S	23,517	11	64	64.3	28:58
3	L1	31,604	2	160	135.3	33:12
3	L2	14,035	11	64	61.0	07:36
4	S	12,231	64	64	53.4	05:56
4	L1	19,947	93	64	64.3	15:40
5	S	14,357	64	64	62.4	07:51
5	L1	12,586	17	64	54.9	06:53
6	S	7,002	9	64	3.1	02:06
6	L1	25,000	109	160	107.4	20:45
6	L2	10,055	36	64	44.2	04:29
7	S	8,988	4	64	39.7	03:13
7	L1	21,323	45	64	64.2	18:04
8	S	7,805	13	64	34.7	02:42
8	L1	19,482	42	64	64.3	14:49
9	S	11,388	38	64	49.8	09:44
9	L1	16,170	87	64	64.3	09:48

Table S5.1 continued…

Chr.	Region	# SNPs sequence	# SNPs array	RAM allocated (GB)	Max. RAM (GB)	Time (hh:mm)
10	S	8,885	6	64	39.2	06:15
10	L1	14,209	8	64	61.8	14:52
11	S	16,274	25	64	64.3	09:47
11	L1	11,483	47	64	50.2	05:26
12	S	11,517	44	64	50.4	05:20
12	L1	10,848	20	64	47.5	04:38
12	L2	8,486	11	64	37.6	03:50

Table S5.2 Candidate genes (Causse *et al.*, 2004; Stevens *et al.*, 2007; Pascual *et al.*, 2013) linked to QTL (within ~10 Mb) found in this study for soluble solid content and their gene activity related to soluble solid content. The last column contains the dataset(s) where the QTL was found using linear regression: SNP array panel (SNP; 1053 SNPs); true sequence data (SEQ; 424,432 SNPs); and imputed sequence data (IMP; 357,849 SNPs).

Chr.	Candidate gene	Activity	QTL found
1	*Solyc01g008240*	Sugar transport	SNP; SEQ; IMP
2	*Solyc02g081160*	Fructokinase	SEQ
	Solyc02g081300	Sucrose synthase	SEQ
	Solyc02g091490 (frk3)	Fructokinase	SEQ
	Solyc02g091830	Hexokinase	SEQ
3	*Solyc03g093400 - Solyc03g094170*	Sugar transport	SEQ; IMP
	Solyc03g093520	Fructokinase	SEQ; IMP
	Solyc03g121070 (hxk1)	Hexokinase	IMP
4	*Solyc04g08288*	Fructokinase	SNP; SEQ
6	*Solyc06g066440 (hxk2)*	Hexokinase	SEQ
	Solyc06g073190 (frk2)	Fructokinase	SEQ
9	*Solyc09g010080 (lin5)*	Invertase	SEQ
	Solyc09g010090 (lin7)	Invertase	SEQ
12	*Solyc12g008510 (hxk3)*	Hexokinase	SNP; SEQ; IMP
	Solyc12g014250 (ppc1)	Phosphoenolpyruvate carboxylase	SNP; SEQ; IMP

CHAPTER 6

IMPUTED WHOLE-GENOME SEQUENCE DATA INCREASES POWER OF GENOME WIDE ASSOCIATION STUDY IN TOMATO (*SOLANUM LYCOPERSICUM*)

RIANNE VAN BINSBERGEN[1, 2]
RICHARD FINKERS[3]
MARIO P.L. CALUS[2]
RICHARD G.F. VISSER[3]
ROEL F. VEERKAMP[2]
FRED A. VAN EEUWIJK[1]

[1] Wageningen University and Research, Biometris, P.O. Box 16, 6700 AA Wageningen, the Netherlands
[2] Wageningen University and Research, Animal Breeding and Genomics, P.O. Box 338, 6700 AH Wageningen, the Netherlands
[3] Wageningen University and Research, Plant Breeding, P.O. Box 386, 6700 AJ Wageningen, the Netherlands

Abstract

It is assumed that whole-genome sequence data includes the causal mutations . Thus, there is no dependency on linkage disequilibrium between the marker and the mutation, which is expected to increase the power of finding these mutations. Genotype imputation enables obtaining whole-genome sequence data of individuals genotyped at lower density. Our objectives were to investigate accuracy of imputation of sequence data and its added value relative to the SolCAP SNP array panel when performing a genome wide association study for soluble solid content in a panel of tomato accessions.

A panel of 196 accessions was genotyped with the SolCAP SNP array panel, and after edits 3,083 SNPs remained. For 38 accessions also whole-genome sequence data was available (containing 107,509 SNPs after edits including pruning for high linkage disequilibrium). Genotype imputation was performed using Beagle 4.1. Next to prediction of the dosage R^2 by Beagle, accuracy of imputation was measured empirically by leave-one-out cross validation for the 38 sequenced accessions. After imputation a genome-wide association study (GWAS) was performed, for 145 accessions with observations available for soluble solid content and imputed sequence data.

The Beagle dosage R^2 was 0.72 on average. For the 38 accessions with sequence data on average 83% of the alleles were imputed in concordance with the original sequence data after performing leave-one-out cross validation. The correlation between the original and imputed genotypes was on average 0.34. Despite the relatively poor accuracy of imputation more significant SNPs (>65 SNPs in 9 regions) were found in the GWAS using the imputed sequence compared to using the SNP array panel (no significant SNPs). Most of these regions were linked to previously identified QTL and candidate genes. These results show that even with a low number of sequenced accessions, and thus limited imputation accuracy, the power of a GWAS can still be improved.

Keywords: imputation, Beagle, SolCAP SNP array, accuracy, soluble solid content

6.1 Introduction

In the last decades available genomic information for individuals increased from a few molecular markers on the DNA to knowing all base-pairs on the genome (whole-genome sequence data). This genomic information can be used for prediction of trait phenotypes for individuals with only genotypic information available (genomic prediction; Meuwissen *et al.*, 2001). Next to genomic prediction, this genomic information could also be used to find causal mutations and/or genes for the trait of interest, for example to get a better understanding of the biological background of a trait (Chibon *et al.*, 2012). For genes with a large effect it is relatively straightforward to find the causal genes or mutations. Unfortunately, many traits of interest are controlled by many genes with predominantly small effects, which makes it more difficult to find the causal mutations. With marker panels these causal mutations are very unlikely to be included in the data. Hence, we rely on the association, i.e. linkage disequilibrium (LD), between the markers and the actual causal variants. Whole-genome sequence data includes all structural genetic variation of an individual, so should also include the causal variants. Using sequence data, therefore, should increase the power of finding the causal mutations, as this does not rely anymore on LD between a marker and the mutation (Meuwissen and Goddard, 2010).

Whole-genome sequencing of many individuals remains expensive, despite the rapidly decreasing costs (The National Human Genome Research Institute; https://www.genome.gov/sequencingcostsdata/). To further reduce costs, in for example human and cattle large consortia exist in which information of large numbers of sequenced individuals is shared (Daetwyler *et al.*, 2014; The 1000 Genomes Project Consortium, 2015). These sequenced genotypes can be used as reference for genotype imputation of individuals genotyped at lower density. Recently, also a number of tomato accessions (*Solanum lycopersicum*) and related wild species were sequenced (The 100 Tomato Genome Sequencing Consortium *et al.*, 2014). A number of these accessions are part of a larger collection of approximately 200 accessions which were genotyped using the SolCAP SNP array panel (Sim *et al.*, 2012) and had phenotypic data on soluble solid content available.

Our objectives were to investigate accuracy of genotype imputation and the added value of imputed sequence data relative to a SNP array panel when performing a genome wide association study in this panel of tomato accessions.

6.2 Materials and methods

The total dataset available for this study contained 231 accessions, of which 229 accessions had observations available on soluble solid content and shape (round or cherry) and 196 accessions were genotyped with the SolCAP SNP array panel (Sim *et al.*, 2012), including 133 round and 63 cherry accessions. From these 196 accessions, 38 accessions had also sequence data available, including 19 round and 19 cherry accessions. Below the genotypic data, phenotypic data, and the performed analyses are described in more detail.

6.2.1 Genotype data

A panel of 196 accessions was genotyped with the SolCAP SNP array panel (Sim *et al.*, 2012) containing 7,334 SNPs on chromosome 1-12. SNPs on the array that were also present in the sequence data (based on physical position) were kept. Whole-genome sequence data was available for 38 out of the 196 accessions (The 100 Tomato Genome Sequencing Consortium *et al.*, 2014). The raw data contained 39,693,648 variants over 12 chromosomes. Bi-allelic SNPs with a minor allele count ≥ 2 and with no missing data were kept. Per chromosome SNPs were pruned based on linkage disequilibrium, as described by Calus *et al.* (2016). In short: from SNPs with a squared correlation between genotypes > 0.90, only the "rightmost" SNP was retained. Next, overlapping SNPs between (pre-pruned) sequence data and SNP array that were deleted during pruning were added back to the sequence data.

Both the SNP array dataset and sequence dataset were phased separately with Beagle 4.1 (Browning and Browning, 2016) using a window of 400 SNPs with 40 SNPs overlap. Phasing was performed for each chromosome separate. The effective population size set to 20. The genetic map was predicted using the approach as described in Chapter 5: based on another SNP array dataset (Víquez-Zamora *et al.*, 2013) with genetic positions of the markers available each chromosome was divided into three regions the short- and the long arm (being recombination hotspots), and the centromere (cold-spot). For each region on every chromosome a linear regression from the genetic positions on the physical positions was performed to predict the genetic positions of the sequenced SNPs.

After phasing, chromosome strand and allele order in the SNP array dataset were adjusted to match the reference data by the conform-gt program (http://faculty.washington.edu/browning/conform-gt.html). Genotype calls of 38 accessions with both sequence data and SNP array panel data were compared to check the concordance between the two datasets.

6.2.2 Genotype imputation

Sequence data of 38 accessions was used as reference to impute sequence data for 158 target accessions with only SNP array data available. Imputation was performed using Beagle v4.1 (Browning and Browning, 2016). For imputation the same settings were used as for phasing, except genotype imputation was performed using a window of 4,000 SNPs with 400 SNPs overlap. Window size was increased to be sure at least one SNP on the SNP array was included in the window. To take imputation uncertainty into account in subsequent analyses, the allele dosage was used as imputed genotype, which equals $0 * P(AA) + 1 * P(AB) + 2 * P(BB)$.

As measure for accuracy of imputation Beagle predicts the dosage R^2 per SNP (range 0 – 1), which is the estimated squared correlation between the estimated allele dosage and the true allele dosage. Next to dosage R^2 a leave-one-out cross validation was performed to obtain an empirical measure for accuracy of imputation for the 38 accessions which had sequence data available. Instead of 38 reference accessions, 37 accessions were used as reference to impute the 38th accession. For this target accession it was assumed no sequence data was available. The 158 accessions with only SNP array data were also included in target dataset, as the software also makes use of haplotype information available in these accessions. This approach was repeated for all 38 accessions (leave-one-out cross validation). Imputation accuracy was measured as the proportion of alleles imputed correctly compared to the original true sequence data, where in both datasets the most likely genotype (0, 1, or 2) was used. Also the Pearson correlation was calculated between the true- and imputed sequence data, where the true sequence data was coded as 0, 1, or 2 and for the imputed sequence data the allele dosage was used.

6.2.3 Genome wide association study

For 229 unique accessions phenotypic information was available for soluble solid content (SSC; brix). These SSC data originated from six trials, performed at four locations in 2008, 2009, or 2010 (Table 6.1; https://www.eu-sol.wur.nl). Of these accessions, 145 accessions were included in further analyses, as they had one or more phenotypic values in at least two trials and had genotypic information available. The unfiltered and filtered distribution of the accessions and observations across the six experiments were shown in Table 6.1.

Table 6.1 Location, year, and the unfiltered number of accessions and number of records per trial. The last two columns represent the filtered number of accessions and number of records, which were the genotyped accessions with one or more records in at least two trials.

Location	**Year**	**Unfiltered number**		**Filtered number**	
		Accessions	*Records*	*Accessions*	*Records*
Hebrew University of Jerusalem, Akko, Israel	2008	206	207	138	139
Hebrew University of Jerusalem, Akko, Israel	2009	56	86	50	79
Hebrew University of Jerusalem, Akko, Israel	2010	96	108	78	90
De Ruiter Seeds	2008	18	21	17	20
Hazera Genetics, Israel	2008	68	73	55	58
Semillas Fitó, Spain	2009	37	37	34	34
		229	**532**	**145**	**420**

Per accession a best linear unbiased estimate (BLUE) was estimated using the lme4 package (version 1.1-12) in R (version 3.3.2) by fitting a linear model with accession as fixed effect and trial as random effect. These BLUEs were used as phenotype in the genome wide association study (GWAS). GWAS was performed for four marker datasets: SNP array panel and three (imputed) sequence data sets. The three sequence data sets included a set with all SNPs, a set with SNPs with dosage $R^2 \geq 0.25$, and a set with SNPs with dosage $R^2 \geq 0.75$. The GWAS was performed using default settings in the GAPIT package (version 2016.03.01; Lipka *et al.*, 2012) in R. Associations between SNPs and the phenotype were studied using a

compressed mixed linear model (CMLM; Zhang *et al.*, 2010). The kinship matrix was calculated using method 1 from VanRaden (2008). Accessions were clustered into groups based on the kinship estimates and for each group the average kinship among accessions was used. The optimal compression level was determined by fitting multiple mixed models and selecting the model with the largest log likelihood function value. An explorative liberal threshold of $-log_{10}(p) \geq 3$ was used to define SNPs associated with SSC.

Table 6.2 Number of variants per chromosome and in total in sequence data and on the SNP array panel. The filtered sequence data included bi-allelic SNPs with minor allele count ≥ 2, no missing data, and was pruned based on linkage disequilibrium. The filtered number of SNPs on the SNP array panel included SNPs that were also found in sequence data (based on physical position)

Chr.	Raw sequence data	Filtered sequence data	Raw SNP array panel	Filtered SNP array panel
1	4,605,493	10,722	517	145
2	2,474,705	8,724	828	297
3	3,369,511	10,310	647	250
4	3,494,330	11,849	893	291
5	3,461,673	9,902	720	357
6	2,485,097	7,889	760	362
7	3,644,532	7,785	441	173
8	3,199,368	8,027	362	131
9	3,766,826	8,297	422	221
10	3,239,434	7,105	441	154
11	2,818,102	8,662	955	527
12	3,134,577	8,314	348	197
All	39,693,648	107,586	7,334	3,105

6.3 Results

6.3.1 Genotype data

In total 19,568,707 SNPs out of 39,693,648 variants were kept on the sequence data after removing variants other than SNPs and removing SNPs with any missing calls or low minor allele count of < 2 (Table 6.2). The remaining SNPs were evenly spread across the genome. Pruning the sequence data based on linkage disequilibrium, reduced the number of SNPs drastically to 107,586 SNPs (Table 6.2). After pruning, the distribution of sequenced SNP across the genome was more like the SNP array panel, i.e. more SNPs at the start and end of the chromosomes compared to the centromere region. For SNP array data 3,105 SNPs out of 7,334 SNPs that were also present in the sequence data (based on physical position) were kept (Table 6.2). On average 2.2% of the genotype calls on the SNP array were missing and this ranged between 0.7% and 34.4% of missing genotype calls per accession.

6.3.2 Concordance sequence data and SNP array panel

Genotype calls of 38 accessions with both sequence data and SNP array panel data were compared to check the concordance of the 3,105 overlapping SNPs between the two datasets. The results are summarized in Table 6.3. On average 4.3% of the genotype calls on SNP array were missing for these 3,105 SNPs. Most of the missing calls were found on chromosome 5 (6.2 %), chromosome 7 (5.2 %), and chromosome 11 (5.0 %).

Table 6.3 Concordance of genotype calls for 38 accessions and 3,105 SNP positions in sequence data and on the SNP array panel. The last three rows represent the percentages of the non-missing genotype calls

	Average	Per accession		Per chromosome	
		Min.	Max.	Min.	Max.
Missing on SNP array	4.3%	0.8%	34.4%	2.8%	6.2%
Both alleles the same	93.5%	61.5%	98.2%	91.3%	95.2%
One allele different	4.7%	1.0%	36.1%	3.3%	6.5%
Both alleles different	1.8%	0.6%	12.1%	1.1%	2.9%

For the non-missing genotype calls on average 93.5% of the calls was the same in sequence data and on SNP array, in 4.7% of the cases one allele differed, and in 1.8% both alleles differed (opposite homozygote). Chromosome 4 showed the poorest concordance as in 91.3% of the cases both alleles were the same, and in 5.8% and 2.9% of the cases one or two alleles differed, respectively. Also chromosome 11 was an outlier, as in 6.5% of the cases one allele differed between genotype calls in sequence data and SNP array panel.

The concordance between SNP array and sequence data for the 19 cherry type accessions was poorer compared to the 19 round type accessions. For the cherry type accessions on average 90.4% of the genotype calls were in concordance, compared to 96.0% for the round type accessions. This difference was mainly caused due to differences in heterozygote versus homozygote calls: for cherry type accessions in 7.8% of the cases one allele differed (maximum was 36.1% and five accessions had above 10%), while for the round type accessions this was only 2.3% on average (one accession had 17.2%, the rest was below 5%).

6.3.3 Imputation accuracy

Accuracy of imputation was measured in two ways: the dosage R^2 per SNP predicted by Beagle software and an empirical measure by leave-one-out cross validation of the 38 sequenced accessions. On average the dosage R^2 was 0.72. In Figure 6.1 a histogram is shown with the distribution of SNPs over the different dosage R^2 classes. For subsequent analyses the dosage R^2 was used to make three sequence data sets: 1) a set with all 107,509 SNPs; 2) a set with 103,367 SNPs with dosage $R^2 \geq 0.25$; and 3) a set with 55,638 SNPs with dosage $R^2 \geq 0.75$.

To obtain an empirical measure for accuracy of genotype imputation, leave-one-out cross validation was performed for the 38 sequenced accessions. On average 83% of the alleles was imputed in concordance with the original sequence data and the Pearson correlation (calculated per SNP) was on average 0.34 and ranged from -1 to +1 per SNP. For 81 SNPs no correlation could be calculated, as no variation was present in the imputed sequence data (for all accessions the same genotype). The correlation between the squared empirical Pearson correlation and the dosage R^2 was 0.51.

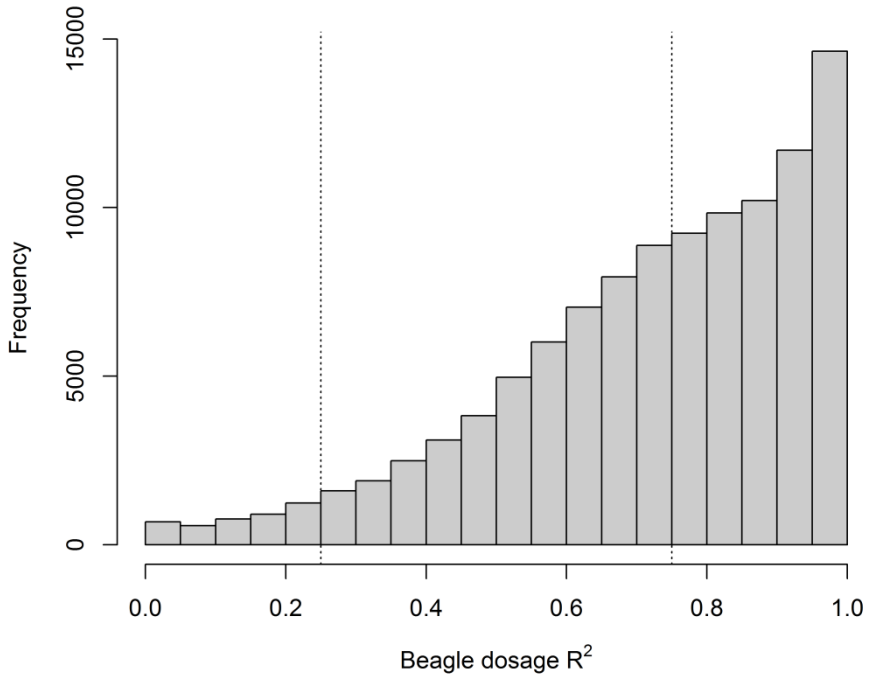

Figure 6.1 Histogram of Beagle dosage R^2 calculated for 107,509 SNPs. The dotted lines indicate a dosage R^2 of 0.25 and 0.75

A large variation in imputation accuracy exists across the genome although no clear region was found with much higher or lower accuracy of imputation accuracy compared to the average. Except for regions with lower SNP density on the array, which did show lower imputation accuracy compared to average. As shown in Table 6.4 the majority of sequenced SNPs (83,515 SNPs) was within 0.5 Mb of a SNP on the SNP array, and the average empirical correlation for those sequenced SNPs was 0.36. For SNPs in other distance classes the average empirical correlation ranged between 0.21 and 0.27 on average (Table 6.4). These regions with lower SNP density on the array were more common in the centromeres of the chromosomes, compared to the chromosomal arms. For SNPs on the chromosomal arms on average 81% (short arm) and 80% (long arm) of the alleles was imputed in concordance with the original sequence data and the Pearson correlation was on average 0.44 (short arm) and 0.34 (long arm), while for the SNPs in the centromere regions the proportion imputed in concordance with the original sequence data was 76% and the Pearson correlation 0.26 on average. These results are based on

all the sequenced SNPs, but similar patterns were observed for the other two sequence data sets when only SNPs with a dosage R^2 ≥0.25 or ≥0.75 were considered (results not shown).

Table 6.4 Average imputation accuracy (proportion alleles imputed in concordance with the original sequence data and Pearson correlation) for SNPs divided into classes based on distance to the closest SNP on the SNP array (in Mb). For each class also the number of SNPs is shown

Distance to closest SNP on SNP array (in Mb)	Number of SNPs	Imputation concordance	Number of SNPs with correlation	Pearson correlation
0.0 - 0.5	83,515	0.80	83,437	0.36
0.5 - 1.0	10,498	0.77	10,495	0.27
1.0 - 1.5	4,440	0.76	4,440	0.27
1.5 - 2.0	2,359	0.75	2,359	0.24
2.0 - 2.5	1,388	0.75	1,388	0.22
2.5 - 3.0	1,115	0.75	1,115	0.23
3.0 - 3.5	951	0.75	951	0.22
3.5 - 4.0	679	0.74	679	0.22
4.0 - 4.5	741	0.74	741	0.21
4.5 - 5.0	456	0.75	456	0.21
> 5.0	1,367	0.76	1,367	0.24

Next to imputation statistics per SNP, also imputation statistics per accession were calculated. These statistics were calculated per accession per chromosome (12 measures per accession) and in Figure 6.2 an overview is shown for the Pearson correlation. This figure shows that variation in imputation statistics exists between accessions. On average the cherry type accessions had lower Pearson correlation compared to the round type accessions (0.50 versus 0.56, respectively). The figure also shows that, although the average Pearson correlation increases with more strict selection of SNPs based on dosage R^2, the variation in Pearson correlation increases as well. Similar results were observed for proportion of alleles that was imputed in concordance with the original sequence data (results not shown).

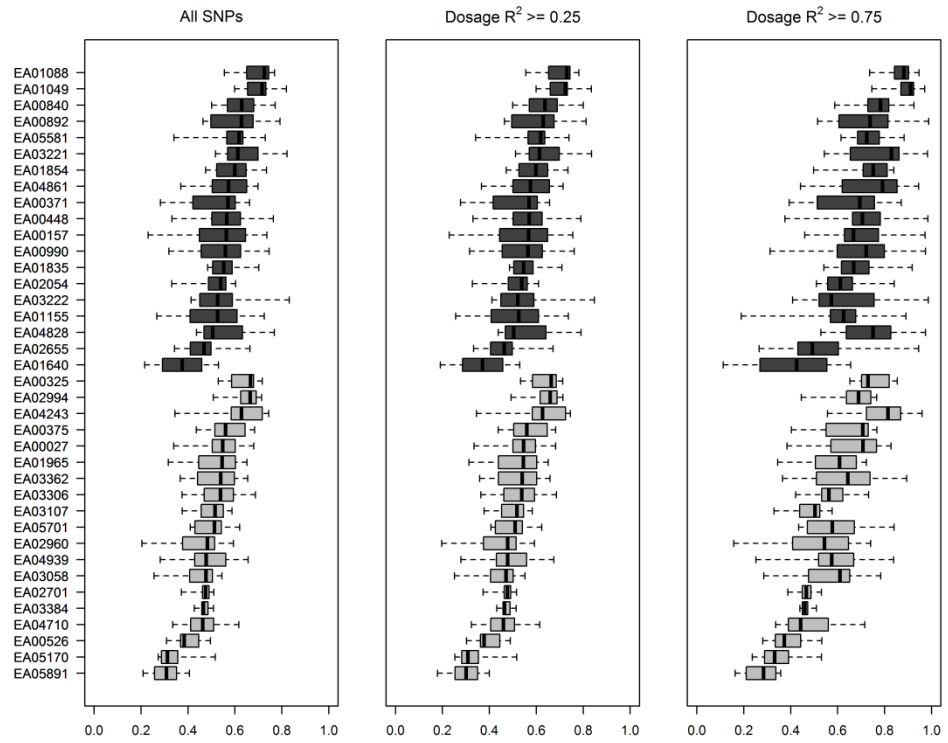

Figure 6.2 Boxplots of the Pearson correlation between true and imputed genotypes per accession. Each boxplot represents twelve correlations (one per chromosome) per accession. The dark grey boxplots are round type tomatoes; the light grey boxplots are cherry type tomatoes.

6.3.4 GWAS

In total 145 accessions were used in the GWAS (Table 6.1), including 97 round type accessions and 48 cherry type accessions. The BLUEs for soluble solid content ranged between 2.5 and 9.0 and were on average 5.3. The cherry type accessions had on average a higher BLUE (and larger range) compared to the round type accessions: 6.0 (3.3 – 9.0) versus 5.0 (2.5 – 6.7) respectively.

GWAS was performed for four genotype datasets: SNP array panel data (3,083 SNPs), sequence data including all SNPs (107,509 SNPs); sequence data including SNPs with dosage $R^2 \geq 0.25$ (103,367 SNPs); sequence data including SNPs with dosage $R^2 \geq 0.75$ (55,638 SNPs). With the SNP array panel data no significant ($-log_{10}(p) \geq 3$) SNPs associated with SSC were found (Figure 6.3), while

for the three sequence datasets respectively 66, 65, and 89 significant SNPs were found. These SNPs were found on chromosome 2 (centromere region, and two regions on the long arm), chromosome 3 (two regions on the long arm), chromosome 7 (long arm), and chromosome 9 (centromere region). With the first two sequence datasets also two regions with significant SNPs were found on chromosome 5 (short arm, and long arm). These were not found with sequence data including SNPs with dosage $R^2 \geq 0.75$, however with this dataset additional significant SNPs were identified on chromosome 4 (one SNP on short arm), and chromosome 11 (centromere region, and long arm). Most of these SNPs were linked to previous found QTL for soluble solid content: *SSC2a, SSC3a*, *SSC3b, ssc5.1*, *ssc5.3, SSC7a* (Tanksley *et al.*, 1996; Chen *et al.*, 1999) or candidate genes for fructokinase (Chr. 2 and 3), sucrose synthese (Chr. 3), sugar transport (Chr. 3), phosphoenolpyruvate carboxylase (Chr. 7), and invertase (Chr. 9) (Causse *et al.*, 2004; Stevens *et al.*, 2007; Pascual *et al.*, 2013).

6.4 Discussion

Our objectives were to investigate the accuracy of genotype imputation and the added value of (imputed) sequence data relative to a SNP array panel when performing genome wide association study for soluble solid content in a tomato association panel. The results showed that, despite the relatively poor accuracy of imputation, more significant SNPs were found using this imputed sequence data compared to the SNP array panel. So our study shows that even with a low number of sequenced accessions, the power of a GWAS can be improved.

6.4.1 Imputation accuracy

The Beagle dosage R^2 was 0.72 on average for the 38 accessions with sequence data and on average 83% of the alleles was imputed in concordance with the original sequence data after performing leave-one-out cross validation. The empirical Pearson correlation between the allele dosage and true genotype was on average 0.34 for these 38 accessions, which is rather low compared with imputation accuracies observed in human and animal populations (e.g. van Binsbergen *et al.*, 2014). Two issues that could have caused this low accuracy of imputation were the low number of sequence accessions and the poor concordance between genotype calls for SNP array and the same SNPs in the sequence data for these sequenced accessions.

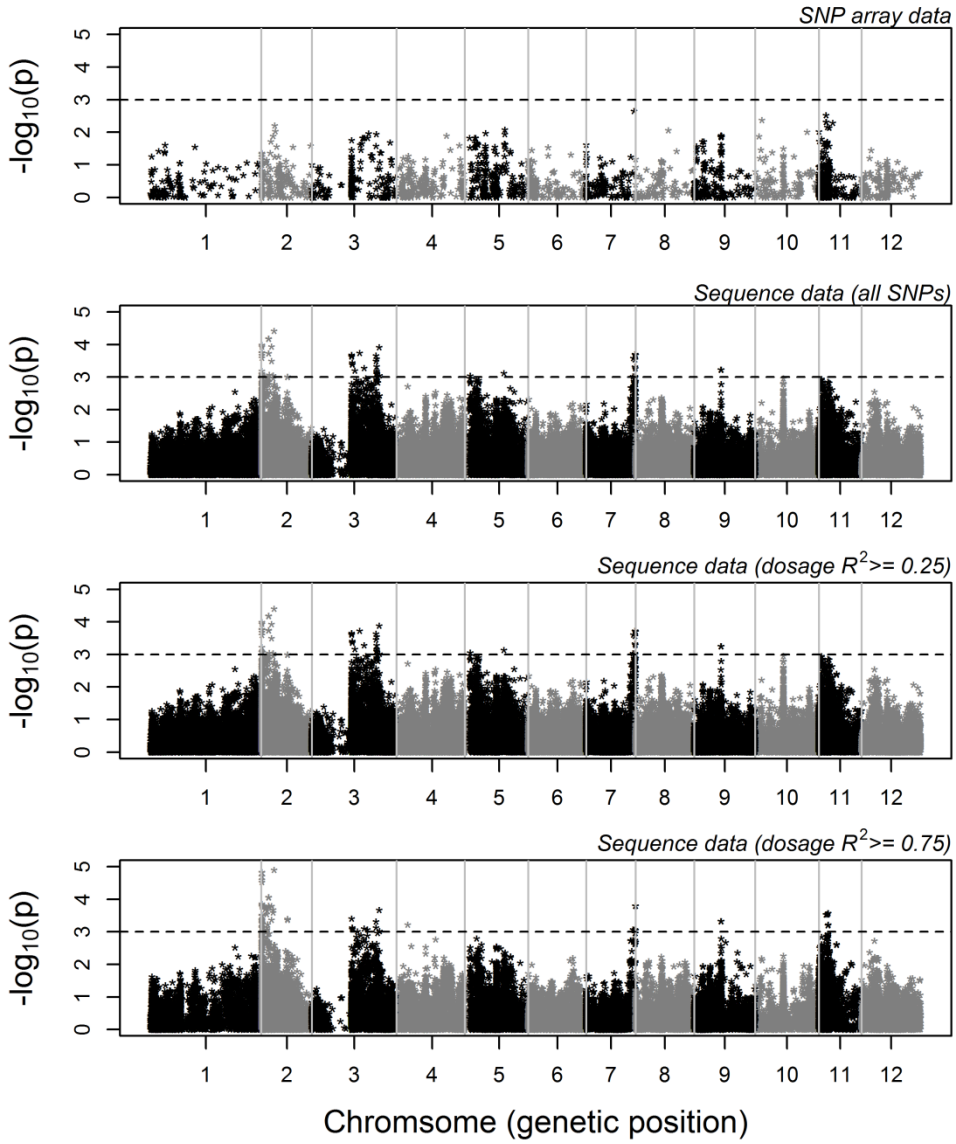

Figure 6.3 Genome wide association for soluble solid content for 145 tomato accessions. Four genotype datasets were compared: SNP array panel data (3,083 SNPs), sequence data including all SNPs (107,509 SNPs); sequence data including SNPs with dosage $R^2 \geq 0.25$ (103,367 SNPs); sequence data including SNPs with dosage $R^2 \geq 0.75$ (55,638 SNPs). The dashed line represents the significance threshold of $-\boldsymbol{log_{10}(p)} \geq \mathbf{3}$

The low number of reference individuals could influence the results in two ways. First, number of reference individuals is positively correlated with accuracy of imputation (e.g. Druet *et al.*, 2010). Increasing the number of reference individuals is especially beneficial for the low minor allele frequency SNPs (van Binsbergen *et al.*, 2014). Second, the empirical measures of accuracy in the present study were based on this low number of accessions, which yields a large standard error for the accuracies. For example, if for a SNP the opposite homozygote was imputed for a few accessions (e.g. 5 out of 38), the correlation between original and imputed genotype drops rapidly for this SNP. To increase accuracy of imputation and have a better accuracy estimate it would be beneficial to sequence more accessions.

Next to number of reference accessions, the poor concordance between genotype calls for SNP array and the same SNPs in the sequence data had probably a major impact on the imputation accuracy. Similar as for the concordance, also the imputation accuracy of the cherry type accessions was lower compared to the round type accessions. The two round type accessions with the poorest imputation accuracy (Figure 6.2), had also the poorest concordance between genotype calls for SNP array and the sequence data from the round shape accession (< 90% concordance). For the cherry type accessions this was less clear, although two out of the three accessions in the bottom of Figure 6.2 had also a poor concordance (< 75%).

Several reasons may explain the observed differences in genotype calls between the SNP array and sequence data. For the SNP array and for sequencing not the same seeds were used, and a small proportion of the genotypes may differ between seeds, especially for more heterogenic accessions. Another reason might be that most cherry type accessions were more distant related to the other accessions. In case of the cherry type accessions, the lower concordance was due to differences in heterozygote versus homozygote calls (i.e. on the array a homozygote call and for the same SNP in the sequence data a heterozygote call or the other way around). Most of the round type accessions were commercial cultivars, which were closely related. The cherry accessions seemed to be a bit more exotic and heterogenic, which was also shown by Víquez-Zamora *et al.* (2013). The SNP array was originally developed on the basis of transcriptome sequencing a set of eight accessions, mapped against the processing tomato cultivar Heinz 1706 (Sim *et al.*, 2012). Heinz is more closely related to the round accessions in this

study. As genetic distances increase, more SNPs are expected, which will especially affect the cherry accessions in this study. Additional SNPs could be in the region of the probe sequences, influencing binding efficiency and affecting the final measurable probe intensity and thus the genotypic call.

6.4.2 GWAS

Despite the relatively poor accuracy of imputation using this imputed sequence data more significant SNPs were found compared to the SNP array panel, albeit that the identified associations were not very strong (the lowest p-value was $1.2 * 10^{-5}$). When applying more strict procedures, like Bonferroni-correction or the false discovery rate procedure (Benjamini and Hochberg, 1995) no significant SNPs were left. In spite of our relatively liberal threshold, most of the found SNPs were linked to previously found QTL or candidate genes, except for the SNPs found on the short arm of chromosome 11 using sequence data including SNPs with dosage R2 ≥ 0.75. These SNPs also showed an increase in $-log_{10}(p)$-value in the other three genotype datasets, but did not reach the significance threshold. A further study with accurate whole-genome data is needed to better fine-map this region.

Except for one region, more significant SNPs found using imputed sequence data could be verified with previous found QTL and candidate genes. Although the accuracy of imputation seemed relatively low, genotype imputation is useful as long as the squared imputation accuracy (r^2) is higher than the (expected) linkage disequilibrium (measured as r^2) between SNPs on the array and the SNPs in the sequence data, which means additional information is added by imputation. Given that using the imputed sequence data improved the results of the GWAS, apparently the linkage disequilibrium between the array and sequence SNPs in the individuals used for the GWAS is on average lower than the average squared imputation accuracy, which equals 0.12. In conclusion, our results suggest that even with a low number of sequenced accessions, the power of a GWAS can be improved.

6.5 Acknowledgements

This study was financially supported by Breed4Food (BO-22.04-011-001-ASG-LR), a public-private partnership in the domain of animal breeding and genomics. The SOLCAP marker data and generation of the Field data was supported by the European Commission through the 6th framework program. Contract number: FOOD-CT-2006-016214.

CHAPTER 7

GENERAL DISCUSSION

7.1 Introduction

The expectations of applying whole-genome sequence data for quantitative trait loci (QTL) detection and genomic prediction were high. Due to less (or no) dependency on linkage disequilibrium between a marker and QTL compared to SNP array panels, it was expected that power of QTL detection would increase with whole-genome sequence data compared to these SNP array panels. Moreover, accuracy of genomic prediction was expected to increase as well with whole-genome sequence data compared to SNP array panels, especially prediction across generations or predictions in multi-breed context would be more accurate.

In the last decade increasing amounts of DNA sequence data were generated for plants and animals. For an increasing number of species whole-genome sequence data of multiple individuals are available, for example Arabidopsis (Alonso-Blanco *et al.*, 2016); tomato (The 100 Tomato Genome Sequencing Consortium *et al.*, 2014); wheat and barley (WHEALBI project; www.whealbi.eu); cattle (**Chapter 2**); and chicken (Heidaritabar *et al.*, 2016). It is expected that more sequence data will be available in the coming years, especially if the decrease in costs will continue at the same rate as described in **Chapter 1** (Figure 1.2). The hypothesis of this thesis was that the use of whole-genome sequence data could increase the accuracy of genomic prediction and the power of QTL mapping. This hypothesis was tested using real whole-genome sequence data in dairy cattle (**Chapter 2**, **Chapter 3**, and **Chapter 4**) and tomato (*Solanum Lycopersicum;* **Chapter 5** and **Chapter 6**). In this last chapter the results from the previous chapters are discussed within the context of the current status for dairy cattle breeding and tomato breeding with respect to the use of whole-genome sequence data. Especially dairy cattle breeding research has progressed since **Chapter 2**, **Chapter 3**, and **Chapter 4** were written and published. Thereafter future perspectives with respect to critical steps for the successful application of whole-genome sequence data in plant and animal breeding will be discussed: genotype imputation, QTL detection, and genomic prediction.

7.2 Current status

7.2.1 Dairy cattle breeding

In 2009 the genome sequence of the cattle genome (*Bos taurus*) was published (The Bovine Genome Sequencing and Analysis Consortium *et al.*, 2009) and soon the commercial BovineSNP50 array became available (Matukumalli *et al.*, 2009), containing 54,001 SNPs across the genome. With this array, also the proposed genome wide prediction method of Meuwissen *et al.* (2001) could be applied on real data. This appeared to be very successful within the Holstein Friesian breed, and has been applied in many Holstein Friesian breeding programs across the world (Hayes *et al.*, 2009; Boichard *et al.*, 2016). However, in the case of numerical small breeds, the accuracy of genomic prediction is still (too) low, as a large reference population is a prerequisite (e.g. Meuwissen *et al.*, 2001; VanRaden *et al.*, 2009; Lin *et al.*, 2014). Numerical small breeds were expected to benefit from combining breeds in the reference population. However, LD between QTL and SNP differs across populations (de Roos *et al.*, 2008; Wientjes *et al.*, 2015) which hinders the combination of breeds in a reference population. The expectation is that including the functional mutation for the trait in the data, this dependency on consistency of LD is removed. With this inclusion power and precision of QTL detection and the persistency of genomic prediction across multiple populations or generations could be increased.

Based on these expectations the aim of the 1000 bull genomes consortium was to sequence key ancestors from the modern cattle population. The results of the first phase of the project were presented in **Chapter 2**. A total of 234 individuals were sequenced and 28.3 million variants were identified. Utilising these whole-genome sequence data, rapid identification of a few genetic defects was demonstrated. For genomic prediction, the goal was to apply genotype imputation to obtain a large reference set of individuals with (imputed) sequence variants (**Chapter 2** and **Chapter 3)**. For the common variants accuracy of imputation was reasonably high, however for low minor allele frequency variants the accuracy of imputation dropped rapidly. Also in regions with less SNP coverage on the SNP array the accuracy of imputation dropped rapidly. The lower imputation accuracy negatively affects accuracy in subsequent analyses, like genomic prediction or QTL detection.

After publication of the results of the first phase of the 1000 bull genomes project in 2014 (**Chapter 2**), more partners got involved in the project and more individuals got sequenced. At the moment of writing, more than 2500 individuals are sequenced and approximately 35 million variants have been identified. Next to an increase in individuals, also many studies followed that used the obtained sequence data (e.g. Bouwman and Veerkamp, 2014; MacLeod *et al.*, 2014; Brøndum *et al.*, 2015; Eynard *et al.*, 2015). Depending on the imputation method used and the design of reference population a reasonable good accuracy of imputation can be reached and QTL detection results become even more promising. However, in contrast to the expectations, increasing marker density to whole-genome sequence data did not improve genomic prediction or only a little (**Chapter 4**). At this stage, whole-genome sequence data within cattle populations appears mostly of interest for researchers and has not been applied in practice for genomic prediction yet.

7.2.2 Tomato breeding

Tomato is an important crop and used as model species for fruit traits, plant growth, plant architecture and disease resistance. Several genetic maps were constructed, mostly based on populations derived from crosses between *S. Lycopersicum* and wild relatives (Shirasawa *et al.*, 2010). Many QTL and candidate genes were detected using these mapping populations (e.g. Tanksley *et al.*, 1996; Chen *et al.*, 1999; Causse *et al.*, 2004). Three years after publication of the cattle genome, the genome of *Solanum Lycopersicum* was published, by sequencing the inbred tomato cultivar 'Heinz 1706' (The Tomato Genome Consortium, 2012). With the availability of the reference genome and high density SNP markers (Sim *et al.*, 2012) genes can be physically located on the genome and discovery of new genes and QTL is accelerated (Causse and Grandillo, 2016). Availability of a reference genome made also genome wide association studies easier. At the moment the number of genome wide association studies is rather limited, but results are promising (e.g. Ranc *et al.*, 2012; Shirasawa *et al.*, 2013; Xu *et al.*, 2013; Sauvage *et al.*, 2014). Two years after publication of the reference genome, a set of 84 accessions was sequenced (The 100 Tomato Genome Sequencing Consortium *et al.*, 2014). To my knowledge no other studies investigated genotype imputation and QTL detection using tomato sequence data as done in **Chapter 5** and **Chapter 6**. Based on our results QTL detection using imputed sequence data in tomato is

promising and might be applied more in the future. Critical steps for the successful application of whole-genome sequence data with respect to genotype imputation and QTL detection will discussed hereafter.

7.3 Genotype imputation

Genotype imputation is used to impute random missing genotypes, or to combine different datasets of individuals genotyped with different SNP arrays, or to increase marker density from a few hundred or few thousand to higher density marker panels. This section includes this last application: increase marker density from SNP marker panels to whole-genome sequence data.

A first issue to be considered when imputing sequence data is that in general SNPs on a SNP array were selected based on segregation in multiple breeds and based on their allele frequency in those breeds,. This was done to explain a large part of the genetic variance. This causes that for most SNP arrays that the number of SNPs per frequency class is evenly distributed. In reality, so with whole-genome sequence data, a more U-shaped distribution is shown, with more SNPs in the low minor allele frequency classes. This discrepancy in the frequency distribution of SNPs could be an issue with respect to accuracy of imputation into whole-genome sequence data. As shown in **Chapter 3** imputation of the SNP with a lower minor allele frequency (MAF) was more difficult. It is hypothesized that causal mutations have a low MAF (Druet *et al.*, 2014) and inclusion of these causal mutations is expected to be the main benefit of sequence data. Consequently, low accuracy of imputation of low MAF SNPs will decrease the potential benefit of imputed sequence data.

Increasing the number of sequenced individuals will increase accuracy of imputation of low minor allele frequency SNPs (**Chapter 3**). More individuals will increase the number of occurrences of the minor allele in the population. When no individuals are available from the same breed, individuals of other breeds could be added to the reference population. An additional benefit of adding other breeds to the reference population is the potential increase in allele frequency of the low minor allele SNPs. This is because allele frequency distribution is not the same across breeds, i.e. some alleles might occur frequent in one breed, but are rare in other breeds. This property could be exploited by using a multi-breed reference population for genotype imputation. Bouwman and Veerkamp (2014) and Brøndum

et al. (2014) showed that the use of an multi-breed reference population is beneficial if only a few sequenced individuals are available of one breed.

Next to number of reference individuals, it is also important that the reference group include (some) individuals related to the imputed individuals, for example the founders of the population. Bi-parental populations have only two founders, sequencing these founders is enough to obtain high accuracy of imputation (**Chapter 5**). In this situation it is important to use the right imputation algorithm as discussed later. The closely related individuals will most likely have similar haplotypes as the imputed individual and therefore can help the imputation algorithm and increase imputation accuracy.

In both, **Chapter 3** and **Chapter 5**, genotype imputation was tested by masking SNPs on the sequence data except for the positions that were present on the SNP array panel. Then sequence data was imputed into this "SNP array panel"-data. In this situation the genotype calls of SNPs in both SNP array panel and sequence data were exactly the same. In practice, different datasets are used when genotype imputation is applied, which can cause differences in genotype calls per individual, as shown in **Chapter 6**. For example, due to genotyping of different seeds of one supposedly inbred line with SNP array panel and with the sequence data. When such a line is not fully homozygous some different genotype calls can occur between samples. But also the SNP array panel or the sequence data can give wrong genotype calls. In case of **Chapter 3** and **Chapter 5** the same genotype dataset was used for sequence data and SNP array panel (masked sequence data), therefore the reported accuracies might therefore be overestimated in comparison to what will be seen in practice.

In **Chapter 1** four types of imputation algorithms were described: 1) naïve approaches; 2) general statistical approaches; 3) family-based approaches; and 4) population-based approaches. In case of imputation of whole-genome sequence data into lower density SNP array, the last two methods are the better choice. For the cattle studies and the tomato accession panel (**Chapter 2**, **Chapter 3**, **Chapter 4**, and **Chapter 6**) no (complete) pedigree information was available, so we chose to use a population-based approach: Beagle (Browning and Browning, 2009). Beagle is used often in human and animal studies (e.g. Marchini and Howie, 2010; Brøndum *et al.*, 2014; Heidaritabar *et al.*, 2016), and some plant studies (e.g. Nazzicari *et al.*, 2016), and has a reasonably good accuracy. However, it is claimed

that other methods perform better with regard to low minor allele frequency SNPs and with a low number of reference individuals (Browning, 2008; Pei *et al.*, 2008; Nothnagel *et al.*, 2009). The advantage of Beagle is that it uses haplotype clustering strategies, which reduces computer time and memory compared to IMPUTE (Marchini *et al.*, 2007) or MaCH (Li *et al.*, 2010), which is an advantage when handling large whole-genome sequence datasets (Browning and Browning, 2009; Nothnagel *et al.*, 2009).

The RIL population in **Chapter 5** had a very clear structure, so family-based approaches (e.g. PlantImpute; Hickey *et al.*, 2015) were more suitable. We also tested Beagle, but it was not able to consider the specific characteristics of the data, and mainly imputed heterozygote genotypes, where homozygous genotypes were expected. At the moment, not many algorithms exists that can handle genotype imputation of sequence data into lower density marker panels for bi-parental populations. PlantImpute gave accurate imputation results, but had large computational requirements. A newer version of the algorithm is under development and will be less computational demanding (C. Nettelblad; personal communication).

Nowadays, multiple imputation algorithms do exist and in many situations reasonable imputation accuracy can be reached. The question arises: how much effort should be put into further improving genotype imputation? As shown in **Chapter 1** (Figure 1.2) the costs of sequencing are decreasing rapidly, so if this decrease continues soon all individuals could be sequenced and genotype imputation of sequence data into SNP array data is no longer needed, at least when DNA is still available for individuals of interest. In that situation, genotype imputation only might be needed for imputation of sporadic missing genotypes, as this could be a prerequisite in subsequent analyses. On the other hand, genotype information of many individuals is available at the moment, why should we spend money on all these individuals and genotype them again at high density (e.g. whole-genome sequence data)? A more cost saving strategy is to sequence a core set of individuals (e.g. the founders) and apply genotype imputation for the other individuals. By clever choosing individuals to be sequenced and development of the right imputation algorithm (that will differ depending on the population structure), imputation will remain important to obtain accurate sequence data for all individuals in a cost effective way.

Another issue regarding sequence data that has not been discussed in this thesis so far, is the sequencing coverage (average number of reads per position) of the individuals. In all chapters the individuals were sequenced at moderate to high coverage (7x coverage or higher). The advantage of high coverage sequence data is the lower number of error rates, compared to low coverage sequencing. For example, coverage of 30x results in a genotype accuracy of more than 99% (Bentley *et al.*, 2008). With low coverage sequence data detecting heterozygote genotype calls is more difficult (Li *et al.*, 2011), and it has been suggested that at least 4x coverage is needed to detect heterozygote genotype calls (Bentley *et al.*, 2008). The drawback of sequencing at higher coverage is the increase in costs per individual. With a fixed amount of money available sequencing at higher coverage implies less individuals to be sequenced, which could be a drawback for detection of low MAF SNPs (<0.5%), as the minor allele might not be segregation in the sequenced individuals. Detection of low MAF SNPs (0.2 – 0.5%) increased when sequencing many individuals at low coverage given the same amount of money and sequencing effort (Li *et al.*, 2011).

In this thesis the approach was taken to use high coverage sequence data for a limited number of key individuals in the population, and then to impute this sequence for other individuals genotyped with SNP array. An alternative strategy to obtain sequence data for a whole population is to sequence many individuals at low coverage, and use genotype imputation to obtain higher quality sequence data by using information of individuals that carry the same haplotype (Li *et al.*, 2011; Hickey, 2013). The optimal design to obtain sequence data for a population likely depends on the population and the application of the sequence data, for some applications it might be better so sequence a few individuals very deep (e.g. when studying Mendelian disorders). For QTL detection and genomic prediction, where large number of individuals is needed, low-coverage sequencing could be a good strategy. Accurate detection of heterozygote genotype calls and low MAF SNPs is however important, and therefore at least 4x coverage is recommended.

7.4 QTL detection

The results from **Chapter 5** and **Chapter 6** showed an increase in number of QTL found with whole-genome sequence data in tomato, which could be verified in previous studies (Grandillo and Tanksley, 1996; Tanksley *et al.*, 1996; Chen *et al.*, 1999; Causse *et al.*, 2004; Stevens *et al.*, 2007; Pascual *et al.*, 2013). In human they found more potentially functional variants using imputed sequence data including over 14 million variants, compared to using 2.4 million array based SNPs from HapMap2 (The 1000 Genomes Project Consortium, 2015). And also in cattle it was demonstrated that with imputed whole-genome sequence data variants associated with milk production and curly coat in cattle could be identified (**Chapter 2**). Another study (van den Berg *et al.*, 2016a) showed that using imputed whole-genome sequence data in a multi-breed population of five French and Danish dairy cattle breeds improves power and precision of genome wide association studies (GWAS) for production traits. However, in that scenario it is assumed that QTL segregate across breeds. They concluded that composition of the population, e.g. number of breeds and number of individuals per breed, were important to improve the power and precision (van den Berg *et al.*, 2016a).

As mentioned in **Chapter 1** LD decays increasingly faster when moving from bi-parental populations, to multi-parent populations, to association panels and outbred populations. In an association panel or (natural) outbred populations the fastest LD decay can be found. With long LD decay power of QTL detection is high and a limited number of markers are needed. Therefore, the main advantage of whole-genome sequence data was expected within populations with fast LD decay. In contrast to our expectations, it was shown that whole-genome sequence data did also increase power of QTL detection in RIL populations (**Chapter 5**; Spindel *et al.*, 2013). The current SNP panels with on average more than one SNP per centiMorgan should be sufficient to locate recombination events and find QTL (Lynch and Walsh, 1998; Hu and Xu, 2008; Takuno *et al.*, 2012). However, as shown in **Chapter 5** with whole-genome sequence data additional recombination events on the genome were found, compared to SNP array data. Although the distance between two SNPs on the SNP array data was on average 1.3 cM, the SNPs were not evenly spread across the genome and some neighbouring SNPs were more than 10 cM apart. For example, in **Chapter 5** on chromosome 3 a large region was shown with many SNPs on sequence data, but almost no SNPs on the array. In this

region two additional recombination events were shown and an additional QTL for soluble solid content was located with (imputed) sequence data compared to the SNP array panel. The denser genome coverage across the whole genome (on average 0.003 cM between two SNPs) with sequence data benefitted the QTL mapping results.

Next to the studied population, also the method is important to increase power and precision of QTL detection. The previous described studies used a single-SNP method. In **Chapter 4** a method was used that simultaneously fitted all SNPs. Although imputed sequence data did explain the same amount of genetic variation, the Manhattan plots in Figure 4.2, Figure 4.3, and Figure 4.4 did not show a clear peak in potential QTL regions. With sequence data the effects of the QTL were spread across multiple SNPs that were in high LD. The method included variable selection, however the small SNP effects might explain a very large part of the variance and, thus the larger QTL remained unclear by the model. Also Calus *et al.* (2016) showed that within the same Holstein population their Bayesian variable selection model failed to find GWAS peaks when using all imputed whole-genome sequence data. When fitting so many SNP (which are all conditional on each other) in the same model the relatively long-range LD in the population hinders precise the QTL detection. However, when the number of imputed variants was reduced, and those were selected based on association with the trait, then the peaks eventually were found (Calus *et al.*, 2016). Similar, when single SNP testing was applied (Veerkamp *et al.*, 2016) in the same data, clearer QTL peaks were identified when using sequence data in comparison with SNP array data. Other studies showed that a multi-breed population results in more precise QTL mapping than a single breed population (Raven *et al.*, 2014; Kemper *et al.*, 2015). This suggests that a large large population of sequenced individuals from multiple breeds or populations is optimal for QTL detection, and should be exploited in future applications.

Power of QTL detection in RIL populations was increased using whole-genome sequence data, indicating that the LD properties in the RIL population deviated from those expected. As expected, power and precision of QTL detection in GWAS can be increased by using whole-genome sequence data, but the size of this increase depends on the studied population. A large population of sequenced individuals from multiple breeds or populations is advised.

7.5 Genomic prediction

As for QTL detection, the expectations of applying whole-genome sequence data for genomic prediction were high. In **Chapter 4** genomic prediction with whole-genome sequence was investigated in real Holstein Friesian cattle data using an approach which is used commonly in dairy cattle; i.e. using a training set with closely related individuals and two types of standard genomic prediction methods: genomic-enabled best linear unbiased prediction (GBLUP) and a Bayesian method. With GBLUP it was assumed a priori that all SNPs have small effects, while with the Bayesian method it was assumed a priori that most SNPs will have a small effect, and a few will have a larger effect. However, in contrast to the expectations, the results showed no advantage of using imputed sequence compared to using the high density SNP array panel for genomic prediction. Next to causal mutations, with whole-genome sequence data also a lot of other variants are added to the prediction model which increase noise in prediction. Instead of using all SNPs for prediction, we first have to perform some pre-selection of SNPs.

Pre-selection can be performed by first performing a GWAS and select the SNPs with the lowest p-values for genomic prediction. Veerkamp *et al.* (2016) investigated this approach, but did not find a clear benefit of using whole-genome sequence data for genomic prediction using the same data as used in **Chapter 4**. This dataset contains only Holstein Friesian animals, which could make detection of causal variants difficult due to long range LD, caused by strong family relationships in the population. In contrast to these results, Brøndum *et al.* (2015) found some increase in genomic prediction reliability when adding QTL variants derived from GWAS to a 54k SNP panel compared to using the 54k SNP panel alone. For the GWAS whole-genome sequence data was available for three Nordic dairy breeds, and genomic prediction was performed for two out of these three breeds and for a French Holstein population. The largest increase in genomic prediction reliability was shown for this French Holstein population, which was not included in the GWAS. Also van den Berg *et al.* (2016b) showed an increase in prediction reliability when adding QTL variants selected from a multi-breed GWAS using sequence data to a 50k SNP panel.

Although pre-selection of QTL variants for genomic prediction using a multi-breed population is promising, there are still some issues to be solved. First, pre-selection of QTL variants using GWAS appeared to work better for traits which are

controlled by a few larger genes, high polygenic traits did not gain a lot of prediction reliability (Brøndum *et al.*, 2015). Furthermore, the optimal set of QTL variants differed per population, which makes it very difficult to apply in routine genomic evaluations (van den Berg *et al.*, 2016b).

Instead of using GWAS for pre-selection, a better approach might be to incorporate biological information in the genomic prediction procedure (Perez-Enciso *et al.*, 2015; MacLeod *et al.*, 2016). Examples of biological information are known QTLs, candidate genes or specific genomic sites like non-synonymous coding sites. The advantage of this type of biological information is that information comes from a range of independent sources, and can be uniform across populations, which makes it easier to apply in routine genome evaluations. MacLeod *et al.* (2016) investigated the use of prior biological information in a Bayesian mixture model (BayesRC) for dairy cattle. In their study they assumed three variant classes with different *a priori* assumptions about the additive genetic variance: a class with non-synonymous coding variants located in set of genes associated with milk productions; a class with other variants in this set of genes; and a class including all other variants. In the best case scenario they found a modest increase in accuracy (correlation between predicted genomic values and corrected phenotypes increased with 0.02) of genomic prediction for milk production traits compared to a BayesR without this prior information. At the moment still a lot of this biological information is unknown for many animal and plant species. To increase accuracy of prediction, we need good prior information, and therefore a stronger collaboration with molecular geneticist is recommended. As with the ENCODE project in human (Skipper *et al.*, 2012), we need to unravel biological background of traits and functional annotations of DNA. The Functional Annotation of Animal Genomes (FAANG) project is a good example of this (The FAANG Consortium *et al.*, 2015).

The challenges for the future with respect to genomic prediction using whole-genome sequence data are to obtain reliable sequence data of many individuals; unravel biological background of traits and functional annotations of DNA; and develop statistical models that incorporate this biological information and that can efficiently handle large datasets. Added value of whole-genome sequence data in genomic prediction will be limited, until these issues are largely solved.

LIST OF REFERENCES

Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork *et al.* 2010. A method and server for predicting damaging missense mutations. Nature Methods 7: 248-249.

Agerholm, J. S., C. Bendixen, O. Andersen, and J. Arnbjerg. 2001. Complex vertebral malformation in Holstein calves. Journal of Veterinary Diagnostic Investigation 13: 283-289.

Agerholm, J. S., J. Arnbjerg, and O. Andersen. 2004. Familial chondrodysplasia in Holstein calves. Journal of Veterinary Diagnostic Investigation 16: 293-298.

Alonso-Blanco, C., J. Andrade, C. Becker, F. Bemm, J. Bergelson, Karsten M. Borgwardt *et al.* 2016. 1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell 166: 1-11.

Bai, Y., and P. Lindhout. 2007. Domestication and breeding of tomatoes: What have we gained and what can we gain in the future? Annals of Botany 100: 1085-1094.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57: 289-300.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown *et al.* 2008a. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53-59.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown *et al.* 2008b. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53-59.

Bionaz, M., and J. J. Loor. 2008. ACSL1, AGPAT6, FABP3, LPIN1, and SLC27A6 are the most abundant isoforms in bovine mammary tissue and their expression is affected by stage of lactation. The Journal of Nutrition 138: 1019-1024.

Boichard, D., L. Maignel, and E. Verrier. 1997. The value of using probabilities of gene origin to measure genetic variability in a population. Genetics Selection Evolution 29: 5-23.

Boichard, D., and M. Brochard. 2012. New phenotypes for new breeding goals in dairy cattle. Animal 6: 544-550.

Boichard, D., V. Ducrocq, P. Croiseau, and S. Fritz. 2016. Genomic selection in domestic animals: Principles, applications and perspectives. Comptes Rendus Biologies 339: 274-277.

Bonaventure, J., L. Cohen-Solal, P. Ritvaniemi, L. Van Maldergem, N. Kadhom, A. L. Delezoide *et al.* 1995. Substitution of aspartic acid for glycine at position 310 in type II collagen produces achondrogenesis II, and substitution of serine at position 805 produces hypochondrogenesis: analysis of genotype-phenotype relationships. Biochemical Journal 307: 823-830.

Bouwman, A. C., and R. F. Veerkamp. 2014. Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. BMC Genetics 15: 105.

Breiman, L. 2001. Random Forests. Machine Learning 45: 5-32.

Broman, K. W., H. Wu, Ś. Sen, and G. A. Churchill. 2003. R/qtl: QTL mapping in experimental crosses. Bioinformatics 19: 889-890.

Brøndum, R. F., B. Guldbrandtsen, G. Sahana, M. Lund, and G. Su. 2014. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics 15: 728.

Brøndum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and M. S. Lund. 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. Journal of Dairy Science 98: 4107-4116.

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. The American Journal of Human Genetics 84: 210-223.

Browning, B. L., and S. R. Browning. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194: 459-471.

Browning, Brian L., and Sharon R. Browning. 2016. Genotype imputation with millions of reference samples. The American Journal of Human Genetics 98: 116-126.

Browning, S. R., and B. L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. The American Journal of Human Genetics 81: 1084-1097.

Browning, S. R. 2008. Missing data imputation and haplotype phase inference for genome-wide association studies. Human Genetics 124: 439-450.

Calus, M., C. Schrooten, and R. Veerkamp. 2014. Genomic prediction of breeding values using previously estimated SNP variances. Genetics Selection Evolution 46: 52.

Calus, M. P. L. 2014. Right-hand-side updating for fast computing of genomic breeding values. Genetics Selection Evolution 46: 24.

Calus, M. P. L., A. C. Bouwman, C. Schrooten, and R. F. Veerkamp. 2016. Efficient genomic prediction based on whole-genome sequence data using split-and-merge Bayesian variable selection. Genetics Selection Evolution 48: 1-19.

Causse, M., P. Duffe, M. C. Gomez, M. Buret, R. Damidaux, D. Zamir *et al.* 2004. A genetic map of candidate genes and QTLs involved in tomato fruit size and composition. Journal of Experimental Botany 55: 1671-1685.

Causse, M., and S. Grandillo. 2016. Gene mapping in tomato. In: M. Causse, J. Giovannoni, M. Bouzayen and M. Zouine (eds.) The Tomato Genome. p 23-37. Springer Nature, Berlin.

Chen, F. Q., M. R. Foolad, J. Hyman, D. A. St. Clair, and R. B. Beelaman. 1999. Mapping of QTLs for lycopene and other fruit traits in a *Lycopersicon esculentum* × *L. pimpinellifolium* cross and comparison of QTLs across tomato species. Molecular Breeding 5: 283-299.

Chen, L., C. Li, M. Sargolzaei, and F. Schenkel. 2014. Impact of genotype imputation on the performance of GBLUP and Bayesian methods for genomic prediction. PLoS ONE 9: e101544.

Chibon, P.-Y., H. Schoof, R. G. F. Visser, and R. Finkers. 2012. Marker2sequence, mine your QTL regions for candidate genes. Bioinformatics 28: 1921-1922.

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

Cleveland, M. A., and J. M. Hickey. 2013. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. Journal of Animal Science 91: 3583-3592.

Cole, J. B., P. M. VanRaden, J. R. O'Connell, C. P. Van Tassell, T. S. Sonstegard, R. D. Schnabel *et al.* 2009. Distribution and location of genetic effects for dairy traits. Journal of Dairy Science 92: 2931-2946.

Collard, B. C. Y., M. Z. Z. Jahufer, J. B. Brouwer, and E. C. K. Pang. 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142: 169-196.

Crossa, J., Y. Beyene, S. Kassa, P. Pérez, J. M. Hickey, C. Chen *et al.* 2013. Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3: Genes|Genomes|Genetics 3: 1903-1926.

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., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R. F. Brøndum *et al.* 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics 46: 858–865.

Dalton, R. 2009. No bull: genes for better milk. Nature 457: 369.

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

de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle. Genetics 179: 1503-1512.

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., and F. Hospital. 2002. The use of molecular genetics in the improvement of agricultural populations. Nature Reviews Genetics 3: 22-32.

Dekkers, J. C. M. 2004. Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. Journal of Animal Sciences 82 (E. Suppl.): E313-E328.

Delaneau, O., J.-F. Zagury, and J. Marchini. 2013. Improved whole-chromosome phasing for disease and population genetic studies. Nature Methods 10: 5-6.

Druet, T., and M. Georges. 2010. A Hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics 184: 789-798.

Druet, T., C. Schrooten, and A. P. W. de Roos. 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. Journal of Dairy Science 93: 5443-5454.

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.

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

Ertl, J., C. Edel, R. Emmerling, H. Pausch, R. Fries, and K. U. Götz. 2014. On the limited increase in validation reliability using high-density genotypes in genomic best linear unbiased prediction: Observations from Fleckvieh cattle. Journal of Dairy Science 97: 487-496.

Eynard, S. E., J. J. Windig, G. Leroy, R. van Binsbergen, and M. P. Calus. 2015. The effect of rare alleles on estimated genomic relationships from whole genome sequence data. BMC Genetics 16: 1-12.

Falconer, D. S., and T. F. C. Mackay. 1996. Introduction to quantitative genetics, 4th edition. Longman, Harlow, UK.

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

Finlay, E. K., C. Gaillard, S. M. F. Vahidi, S. Z. Mirhoseini, H. Jianlin, X. B. Qi *et al.* 2007. Bayesian inference of population expansions in domestic bovines. Biology letters 3: 449-452.

Freeman, L., L. Aragon-Alcaide, and A. Strunnikov. 2000. The condensin complex governs chromosome condensation and mitotic transmission of rDNA. Journal of Cell Biology 149: 811-824.

Fridley, B. L., G. Jenkins, M. E. Deyo-Svendsen, S. Hebbring, and R. Freimuth. 2010. Utilizing genotype imputation for the augmentation of sequence data. PLoS ONE 5: e11018.

Gasparin, G., M. Miyata, L. L. Coutinho, M. L. Martinez, R. L. Teodoro, J. Furlong *et al.* 2007. Mapping of quantitative trait loci controlling tick [Riphicephalus (Boophilus) microplus] resistance on bovine chromosomes 5, 7 and 14. Animal Genetics 38: 453-459.

Gautier, M., T. Faraut, K. Moazami-Goudarzi, V. Navratil, M. Foglio, C. Grohs *et al.* 2007. Genetic and haplotypic structure in 14 European and African cattle breeds. Genetics 177: 1059-1070.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson. 2014. ASReml User Guide Release 4.0, VSN International Ltd, Hemel Hempstead, HP1 1ES, UK www.vsni.co.uk.

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

Goddard, M. E., B. J. Hayes, and T. H. E. Meuwissen. 2010. Genomic selection in livestock populations. Genetics Research 92: 413-421.

Godfrey, M., D. R. Keene, E. Blank, H. Hori, L. Y. Sakai, L. A. Sherwin, and D. W. Hollister. 1988. Type II achondrogenesis-hypochondrogenesis: morphologic and immunohistopathologic studies. The American Journal of Human Genetics 43: 894-903.

Grandillo, S., and S. D. Tanksley. 1996. QTL analysis of horticultural traits differentiating the cultivated tomato from the closely related species *Lycopersicon pimpinellifolium*. Theoretical and Applied Genetics 92: 935-951.

Grant, J. R., A. S. Arantes, X. Liao, and P. Stothard. 2011. In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. Bioinformatics 27: 2300-2301.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi *et al.* 2002. Positional candidate coning 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 Research 12: 222-231.

Grisart, B., F. Farnir, L. Karim, N. Cambisano, J.-J. Kim, A. Kvasz *et al.* 2004. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proceedings of the National Academy of Sciences 101: 2398-2403.

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. Genetics Selection Evolution 42: 5.

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

Haldane, J. B. S. 1919. The probable errors of calculated linkage values, and the most accurate method of determining gametic from certain zygotic series. Journal of Genetics 8: 291-297.

Hamblin, M. T., E. S. Buckler, and J. L. Jannink. 2011. Population genetics of genomics-based crop improvement methods. Trends in Genetics 27: 98-106.

Hao, K., E. Chudin, J. McElwee, and E. Schadt. 2009. Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. BMC Genetics 10: 27.

Hayes, B. J., P. M. Visscher, H. C. McPartlan, and M. E. Goddard. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Research 13: 635-643.

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

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009b. Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science 92: 433-443.

Hayes, B. J., P. J. Bowman, H. D. Daetwyler, J. W. Kijas, and J. H. J. van der Werf. 2012. Accuracy of genotype imputation in sheep breeds. Animal Genetics 43: 72-80.

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 in Genetics 29: 206-214.

Hayes, B. J., I. M. Macleod, H. D. Daetwyler, P. J. Bowman, A. J. Chamberlain, C. J. Vander Jagt *et al.* 2014. Genomic prediction from whole genome sequence in livestock: the 1000 bull genomes project. In: 10th World Congress of Genetics Applied to Livestock Production, Vancouver, Canada

Heffner, E. L., M. E. Sorrells, and J.-L. Jannink. 2009. Genomic selection for crop improvement. Crop Science 49: 1-12.

Hegarty, R. S., D. Alcock, D. L. Robinson, J. P. Goopy, and P. E. Vercoe. 2010. Nutritional and flock management options to reduce methane output and methane per unit product from sheep enterprises. Animal Production Science 50: 1026-1033.

Heidaritabar, M., M. P. L. Calus, H. J. Megens, A. Vereijken, M. A. M. Groenen, and J. W. M. Bastiaansen. 2016. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. Journal of Animal Breeding and Genetics 133: 167-179.

Hickey, J. M., J. Crossa, R. Babu, and G. de los Campos. 2012a. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Science 52: 654-663.

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

Hickey, J. M. 2013. Sequencing millions of animals for genomic selection 2.0. Journal of Animal Breeding and Genetics 130: 331-332.

Hickey, J. M., G. Gorjanc, R. K. Varshney, and C. Nettelblad. 2015. Imputation of single nucleotide polymorphism genotypes in biparental, backcross, and topcross populations with a Hidden Markov Model. Crop Science 55: 1934-1946.

Howie, B., J. Marchini, and M. Stephens. 2011. Genotype imputation with thousands of genomes. G3: Genes, Genomes, Genetics 1: 457-470.

Howie, B., C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics 44: 955-959.

Howie, B. N., P. Donnelly, and J. Marchini. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics 5: e1000529.

Hu, Z., and S. Xu. 2008. A simple method for calculating the statistical power for detecting a QTL located in a marker interval. Heredity 101: 48-52.

Huang, B. E., K. L. Verbyla, A. P. Verbyla, C. Raghavan, V. K. Singh, P. Gaur *et al.* 2015. MAGIC populations in crops: current status and future prospects. Theoretical and Applied Genetics 128: 999-1017.

Huang, L., C. Wang, and N. A. Rosenberg. 2009. The relationship between imputation error and statistical power in genetic association studies in diverse populations. The American Journal of Human Genetics 85: 692-698.

Hudson, D. F., P. Vagnarelli, R. Gassmann, and W. C. Earnshaw. 2003. Condensin is required for nonhistone protein assembly and structural integrity of vertebrate mitotic chromosomes. Developmental cell 5: 323-336.

Hudson, D. F., K. M. Marshall, and W. C. Earnshaw. 2009. Condensin: architect of mitotic chromosomes. Chromosome Research 17: 131-144.

Iwata, H., and J.-L. Jannink. 2010. Marker genotype imputation in a low-marker-density panel with a high-marker-density reference panel: accuracy evaluation in barley breeding lines. Crop Science 50: 1269-1278.

Jannink, J.-L., A. J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: from theory to practice. Briefings in Functional Genomics 9: 166-177.

Jansen, S., B. Aigner, H. Pausch, M. Wysocki, S. Eck, A. Benet-Pages *et al.* 2013. Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. BMC Genomics 14: 446.

Jiang, C., and Z.-B. Zeng. 1997. Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Genetica 101: 47-58.

Johnson, K. A., and R. S. Goody. 2011. The original Michaelis constant: translation of the 1913 Michaelis–Menten paper. Biochemistry 50: 8264-8269.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S.-y. Kong, N. B. Freimer *et al.* 2010. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics 42: 348-354.

Kemper, K., C. Reich, P. Bowman, C. vander Jagt, A. Chamberlain, B. Mason *et al.* 2015. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-

breed population leads to greater accuracy of across-breed genomic predictions. Genetics Selection Evolution 47: 29.

Khatkar, M., G. Moser, B. Hayes, and H. Raadsma. 2012. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. BMC Genomics 13: 538.

Kim, E. S., and B. W. Kirkpatrick. 2009. Linkage disequilibrium in the North American Holstein population. Animal Genetics 40: 279-288.

Knol, E. F., B. Nielsen, and P. W. Knap. 2016. Genomic selection in commercial pig breeding. Animal Frontiers 6.

Körkkö, J., D. H. Cohn, L. Ala-Kokko, D. Krakow, and D. J. Prockop. 2000. Widely distributed mutations in the COL2A1 gene produce achondrogenesis type II/hypochondrogenesis. American journal of medical genetics 92: 95-100.

Kosambi, D. D. 1944. The estimation of map distance from recombination values. Annals of Eugenics 12: 172-175.

Kumar, P., S. Henikoff, and P. C. Ng. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols 4: 1073-1081.

Lander, E. S., and D. Botstein. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185-199.

Leu, S.-Y., and P. K. Sen. 2014. Non-homogeneous poisson process model for genetic crossover interference. Communications in Statistics - Theory and Methods 43: 44-71.

Lewontin, R. C. 1995. The detection of linkage disequilibrium in molecular sequence data. Genetics 140: 377-388.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer *et al.* 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

Li, L., Y. Li, S. R. Browning, B. L. Browning, A. J. Slater, X. Kong *et al.* 2011a. Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. PLoS ONE 6: e24945.

Li, Y., C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic Epidemiology 34: 816-834.

Li, Y., C. Sidore, H. M. Kang, M. Boehnke, and G. R. Abecasis. 2011b. Low-coverage sequencing: Implications for design of complex trait association studies. Genome Research 21: 940-951.

Lin, Z., B. J. Hayes, and H. D. Daetwyler. 2014. Genomic selection in crops, trees and forages: a review. Crop and Pasture Science 65: 1177-1191.

Lipka, A. E., F. Tian, Q. Wang, J. Peiffer, M. Li, P. J. Bradbury *et al.* 2012. GAPIT: genome association and prediction integrated tool. Bioinformatics 28: 2397-2399.

Lynch, M., and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Inc.

MacEachern, S., B. Hayes, J. McEwan, and M. Goddard. 2009. An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (Bos taurus) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in domestic cattle. BMC Genomics 10: 181.

Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, D. Zhu *et al.* 2012. The Drosophila melanogaster Genetic Reference Panel. Nature 482: 173-178.

MacLeod, I., P. Bowman, C. Vander Jagt, M. Haile-Mariam, K. Kemper, A. Chamberlain *et al.* 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics 17: 144.

MacLeod, I. M., T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard. 2009. A novel predictor of multilocus haplotype homozygosity: comparison with existing predictors. Genetics Research 91: 413-426.

Macleod, I. M., B. J. Hayes, and M. E. Goddard. 2013. Will sequence SNP data improve the accuracy of genomic prediction in te presence of long term selection? Proceedings of the Association for the Advancement of Animal Breeding and Genetics 20: 215-219.

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

Macleod, I. M., B. J. Hayes, C. J. Vander Jagt, K. E. Kemper, M. Haile-Mariam, P. J. Bowman *et al.* 2014b. A Bayesian analysis to exploit imputed sequence variants for QTL discovery. In: 10th World Congress of Genetics Applied to Livestock Production, Vancouver, Canada

Marchini, J., B. Howie, S. Myers, G. McVean, and P. Donnelly. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics 39: 906-913.

Marchini, J., and B. Howie. 2010. Genotype imputation for genome-wide association studies. Nature Reviews Genetics 11: 499-511.

Martinez, M. L., M. A. Machado, C. S. Nascimento, M. V. G. B. Silva, R. L. Teodoro, J. Furlono *et al.* 2006. Association of BoLA-DRB3.2 alleles with tick (Boophilus microplus) resistance in cattle. Genetics and Molecular Research 5: 513-524.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton *et al.* 2009. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. PLoS ONE 4: e5350.

Mayor, P. J., and R. Bernardo. 2009. Genomewide selection and marker-assisted recurrent selection in doubled haploid versus F2 populations. Crop Science 49: 1719-1725.

Menda, N., S. R. Strickler, and L. A. Mueller. 2013. Advances in tomato research in the post-genome era. Plant Biotechnology 30: 243-256.

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

Miller, S. 2013. Sharp upper limit for r2 as a measure of linkage disequilibrium in multiple marker maps. In: Gordon Research Conference "Quantitative Genetics and Genomics", Galveston

Moore, G. M. 1965. Cramming more components onto integrated circuits. Electronics 38: 114-117.

Mortler, G. R., D. J. Wilkin, W. R. Wllcox, D. L. Rimoin, R. S. Lachman, D. R. Eyre, and D. H. Cohn. 1995. A radiographic, morphologic, biochemical and molecular analysis of a case of achondrogenesis type II resulting from substitution for a glycine residue (Gly691[rarr]Arg) in the type II collagen trimer. Human Molecular Genetics 4: 285-288.

Mueller, J. C. 2004. Linkage disequilibrium for different scales and applications. Briefings in Bioinformatics 5: 355-364.

Muir, W. M. 2007. 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 124: 342-355.

Mulder, H. A., M. P. L. Calus, T. Druet, and C. Schrooten. 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. Journal of Dairy Science 95: 876-889.

Nazzicari, N., F. Biscarini, P. Cozzi, E. C. Brummer, and P. Annicchiarico. 2016. Marker imputation efficiency for genotyping-by-sequencing data in rice (*Oryza sativa*) and alfalfa (*Medicago sativa*). Molecular Breeding 36: 1-16.

NCBI Resource Coordinators. 2013. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 41: D8-d20.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song. 2011. Genotype and SNP calling from next-generation sequencing data. Nature Review Genetics 12: 443-451.

Nothnagel, M., D. Ellinghaus, S. Schreiber, M. Krawczak, and A. Franke. 2009. A comprehensive evaluation of SNP genotype imputation. Human Genetics 125: 163-171.

Ober, U., J. F. Ayroles, E. A. Stone, S. Richards, D. Zhu, R. A. Gibbs *et al.* 2012. Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. PLoS Genetics 8: e1002685.

Pasaniuc, B., N. Rohland, P. J. McLaren, K. Garimella, N. Zaitlen, H. Li *et al.* 2012. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nature Genetics 44: 631-635.

Pascual, L., J. Xu, B. Biais, M. Maucourt, P. Ballias, S. Bernillon *et al.* 2013. Deciphering genetic diversity and inheritance of tomato fruit weight and composition through a systems biology approach. Journal of Experimental Botany 64: 5737-5752.

Pausch, H., B. Aigner, R. Emmerling, C. Edel, K.-U. Götz, and R. Fries. 2013. Imputation of high-density genotypes in the Fleckvieh cattle population. Genetics Selection Evolution 45: 1-10.

Pei, Y. F., J. Li, L. Zhang, C. J. Papasian, and H. W. Deng. 2008. Analyses and comparison of accuracy of different genotype imputation methods. PLoS ONE 3: e3551.

Perez-Enciso, M., J. Rincon, and A. Legarra. 2015. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genetics Selection Evolution 47: 43.

Porter, R., M. Gandhi, N. Wilson, P. Wood, W. McLean, and E. Lane. 2004. Functional analysis of keratin components in the mouse hair follicle inner root sheath. British Journal of Dermatology 150: 195-204.

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

Qanbari, S., E. C. G. Pimentel, J. Tetens, G. Thaller, P. Lichtner, A. R. Sharifi, and H. Simianer. 2010. The pattern of linkage disequilibrium in German Holstein cattle. Animal Genetics 41: 346-356.

Ranc, N., S. Muños, J. Xu, M.-C. Le Paslier, A. Chauveau, R. Bounon *et al.* 2012. Genome-wide association mapping in tomato (*Solanum lycopersicum*) is possible using genome admixture of *Solanum lycopersicum* var. *cerasiforme*. G3: Genes|Genomes|Genetics 2: 853-864.

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.

Rozen, S., and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. Bioinformatics methods and protocols 132: 365-386.

Runkel, F., M. Klaften, K. Koch, V. Böhnert, H. Büssow, H. Fuchs *et al.* 2006. Morphologic and molecular characterization of two novel *Krt71* (*Krt2-6g*) mutations: *Krt71*^{rco12} and *Krt71*^{rco13}. Mammalian Genome 17: 1172-1182.

Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel. 2011. FImpute - An efficient imputation algorithm for dairy cattle populations. Journal of Animal Sciences 89: S421.

Sargolzaei, M., J. Chesnais, and F. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15: 478.

Sauvage, C., V. Segura, G. Bauchet, R. Stevens, P. T. Do, Z. Nikoloski *et al.* 2014. Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. Plant Physiology 165: 1120-1132.

Scheet, P., and M. Stephens. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. The American Journal of Human Genetics 78: 629-644.

Schrooten, C., R. Dassonneville, V. Ducrocq, R. Brondum, M. Lund, J. Chen *et al.* 2014. Error rate for imputation from the Illumina BovineSNP50 chip to the Illumina BovineHD chip. Genetics Selection Evolution 46: 10.

Segelke, D., J. Chen, Z. Liu, F. Reinhardt, G. Thaller, and R. Reents. 2012. Reliability of genomic prediction for German Holsteins using imputed genotypes from low-density chips. Journal of Dairy Science 95: 5403-5411.

Shirasawa, K., S. Isobe, H. Hirakawa, E. Asamizu, H. Fukuoka, D. Just *et al.* 2010. SNP Discovery and Linkage Map Construction in Cultivated Tomato. DNA Research 17: 381-391.

Shirasawa, K., H. Fukuoka, H. Matsunaga, Y. Kobayashi, I. Kobayashi, H. Hirakawa *et al.* 2013. Genome-wide association studies using single nucleotide polymorphism markers developed by re-sequencing of the genomes of cultivated tomato. DNA Research 20: 593-603.

Shuster, D. E., M. E. Kehrli, M. R. Ackermann, and R. O. Gilbert. 1992. Identification and prevalence of a genetic defect that causes leukocyte adhesion deficiency in Holstein cattle. Proceedings of the National Academy of Sciences 89: 9225-9229.

Siddiqui, N. U., P. E. Stronghill, R. E. Dengler, C. A. Hasenkampf, and C. D. Riggs. 2003. Mutations in Arabidopsis condensin genes disrupt embryogenesis, meristem organization and segregation of homologous chromosomes during meiosis. Development 130: 3283-3295.

Sikorska, K., E. Lesaffre, P. F. Groenen, and P. H. Eilers. 2013. GWAS on your notebook: fast semi-parallel linear and logistic regression for genome-wide association studies. BMC Bioinformatics 14: 1-11.

Sim, S.-C., G. Durstewitz, J. Plieske, R. Wieseke, M. W. Ganal, A. Van Deynze *et al.* 2012. Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. PLoS ONE 7: e40563.

Skipper, M., R. Dhand, and P. Campbell. 2012. Presenting ENCODE. Nature 489: 45-45.

Spencer, C. C. A., Z. Su, P. Donnelly, and J. Marchini. 2009. Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. PLoS Genetics 5: e1000477.

Spindel, J., M. Wright, C. Chen, J. Cobb, J. Gage, S. Harrington *et al.* 2013. Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. Theoretical and Applied Genetics 126: 2699-2716.

Stekhoven, D. J., and P. Bühlmann. 2012. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 28: 112-118.

Stevens, R., M. Buret, P. Duffé, C. Garchery, P. Baldet, C. Rothan, and M. Causse. 2007. Candidate genes and quantitative trait loci affecting fruit ascorbic acid content in three tomato populations. Plant Physiology 143: 1943-1953.

Stray, J. E., N. J. Crisona, B. P. Belotserkovskii, J. E. Lindsley, and N. R. Cozzarelli. 2005. The Saccharomyces cerevisiae Smc2/4 condensin compacts DNA into (+) chiral structures without net supercoiling. The Journal of Biological Chemistry 280: 34723-34734.

Strunnikov, A. V., E. Hogan, and D. Koshland. 1995. SMC2, a Saccharomyces cerevisiae gene essential for chromosome segregation and condensation, defines a subgroup within the SMC family. Genes & Development 9: 587-599.

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

Sung, Y. J., L. Wang, T. Rankinen, C. Bouchard, and D. C. Rao. 2012. Performance of genotype imputations using data from the 1000 Genomes project. Human Heredity 73: 18-25.

Sved, J. A. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theoretical Population Biology 2: 125-141.

Takuno, S., R. Terauchi, and H. Innan. 2012. The power of QTL mapping with RILs. PLoS ONE 7: e46545.

Tanakaa, S., I. Miurab, A. Yoshikic, Y. Katoa, H. Yokoyamab, A. Shinogib *et al.* 2007. Mutations in the helix termination motif of mouse type I IRS keratin genes impair the assembly of keratin intermediate filament. Genomics 90: 703-711.

Tanksley, S. D., S. Grandillo, T. M. Fulton, D. Zamir, Y. Eshed, V. Petiard *et al.* 1996. Advanced backcross QTL analysis in a cross between an elite processing line of tomato and its wild relative *L. pimpinellifolium*. Theoretical and Applied Genetics 92: 213-224.

The 100 Tomato Genome Sequencing Consortium, S. Aflitos, E. Schijlen, H. de Jong, D. de Ridder, S. Smit *et al.* 2014. Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. The Plant Journal 80: 136-148.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature 526: 68-74.

The Bovine Genome Sequencing and Analysis Consortium, C. G. Elsik, R. L. Tellam, and K. C. Worley. 2009. The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. Science 324: 522-528.

The Bovine HapMap Consortium. 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science 324: 528-532.

The FAANG Consortium, L. Andersson, A. L. Archibald, C. D. Bottema, R. Brauning, S. C. Burgess *et al.* 2015. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biology 16: 57.

The Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485: 635-641.

Thompson, J. D., T. J. Gibson, and D. G. Higgins. 2002. Multiple sequence alignment using ClustalW and ClustalX. Current Protocols in Bioinformatics.

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani *et al.* 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17: 520-525.

UniProt Consortium. 2011. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Research 39: D214-D219.

Vagnarelli, P., D. F. Hudson, S. A. Ribeiro, L. Trinkle-Mulcahy, J. M. Spence, F. Lai *et al.* 2006. Condensin and Repo-Man-PP1 co-operate in the regulation of chromosome architecture during mitosis. Nature Cell Biology 8: 1133-1142.

van Binsbergen, R., M. C. A. M. Bink, M. P. L. Calus, F. A. van Eeuwijk, B. J. Hayes, I. Hulsegge, and R. F. Veerkamp. 2014a. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution 46: 41.

Van Binsbergen, R., M. P. L. Calus, M. C. A. M. Bink, C. Schrooten, F. A. Van Eeuwijk, and R. F. Veerkamp. 2014b. Genomic prediction with 12.5 million SNPs for 5503 Holstein Friesian bulls. In: 10th World Congress of Genetics Applied to Livestock Production, Vancouver, Canada

van den Berg, I., D. Boichard, and M. S. Lund. 2016a. Comparing power and precision of within-breed and multibreed genome-wide association studies of production traits using whole-genome sequence data for 5 French and Danish dairy cattle breeds. Journal of Dairy Science 99: 8932-8945.

van den Berg, I., D. Boichard, and M. S. Lund. 2016b. Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. Genetics Selection Evolution 48: 83.

VanRaden, P. M., and G. R. Wiggans. 1991. Derivation, calculation, and use of national animal model information. Journal of Dairy Science 74: 2737-2746.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science 91: 4414-4423.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92: 16-24.

VanRaden, P. M., J. R. O. O'Connell, G. R. Wiggans, and K. A. Weigel. 2011a. Genomic evaluations with many more genotypes. Genetics Selection Evolution 43: 10.

VanRaden, P. M., K. M. Olson, D. J. Null, and J. L. Hutchison. 2011b. Harmful recessive effects on fertility detected by absence of homozygous haplotypes. Journal of Dairy Science 94: 6153-6161.

VanRaden, P. M., D. J. Null, M. Sargolzaei, G. R. Wiggans, M. E. Tooker, J. B. Cole *et al.* 2013. Genomic imputation and evaluation using high-density Holstein genotypes. Journal of Dairy Science 96: 668-678.

VanRaden, P. M., C. Sun, and J. R. O'Connell. 2015. Fast imputation using medium or low-coverage sequence data. BMC Genetics 16: 1-12.

Veerkamp, R. F., A. C. Bouwman, C. Schrooten, and M. P. L. Calus. 2016. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein–Friesian cattle. Genetics Selection Evolution 48: 95.

Verbyla, K. L., B. J. Hayes, P. J. Bowman, and M. E. Goddard. 2009. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. Genetics Research 91: 307-311.

Víquez-Zamora, M., B. Vosman, H. v. d. Geest, A. Bovy, R. G. Visser, R. Finkers, and A. W. v. Heusden. 2013. Tomato breeding in the genomics era: insights from a SNP array. BMC Genomics 14: 1-13.

Víquez-Zamora, M., M. Caro, R. Finkers, Y. Tikunov, A. Bovy, R. Visser *et al.* 2014. Mapping in the era of sequencing: high density genotyping and its application for mapping TYLCV resistance in *Solanum pimpinellifolium*. BMC Genomics 15: 1152.

Vissing, H., M. D'Alessio, B. Lee, F. Ramirez, M. Godfrey, and D. W. Hollister. 1989. Glycine to serine substitution in the triple helical domain of pro-[alpha]1 (II) collagen results in a lethal perinatal form of short-limbed dwarfism. The Journal of Biological Chemistry 264: 18265-18267.

Voorrips, R., W. Verkerke, R. Finkers, R. Jongerius, and J. Kanne. 2000. Inheritance of taste components in tomato. Acta Physiologiae Plantarum 22: 259-261.

Wang, X., C. Wurmser, H. Pausch, S. Jung, F. Reinhardt, J. Tetens *et al.* 2012. Identification and dissection of four major QTL affecting milk fat content in the German Holstein-Friesian population. PLoS ONE 7: e40711.

Weckx, S., J. Del-Favero, R. Rademakers, L. Claes, M. Cruts, P. D. Jonghe *et al.* 2005. novoSNP, a novel computational tool for sequence variation discovery. Genome Research 15: 436-442.

Weigel, K. A., G. de los Campos, A. I. Vazquez, G. J. M. Rosa, D. Gianola, and C. P. Van Tassell. 2010. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. Journal of Dairy Science 93: 5423-5435.

Wientjes, Y. C. J., R. F. Veerkamp, and M. P. L. Calus. 2015. Using selection index theory to estimate consistency of multi-locus linkage disequilibrium across populations. BMC Genetics 16: 87.

Wimmer, V., C. Lehermeier, T. Albrecht, H.-J. Auinger, Y. Wang, and C.-C. Schön. 2013. Genome-wide prediction of traits with different genetic architecture through efficient variable selection. Genetics 195: 573-587.

Winter, A., W. Krämer, F. A. O. Werner, S. Kollers, S. Kata, G. Durstewitz *et al.* 2002. Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA:diacylglycerol acyltransferase (DGAT1) with variation at a quantitative trait locus for milk fat content. Proceedings of the National Academy of Sciences 99: 9300-9305.

Xavier, A., W. M. Muir, and K. M. Rainey. 2016. Impact of imputation methods on the amount of genetic variation captured by a single-nucleotide polymorphism panel in soybeans. BMC Bioinformatics 17: 1-9.

Xu, J., N. Ranc, S. Muños, S. Rolland, J.-P. Bouchet, N. Desplat *et al.* 2013. Phenotypic diversity and association mapping for fruit quality traits in cultivated tomato and related species. Theoretical and Applied Genetics 126: 567-581.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt *et al.* 2010. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42: 565-569.

Young, N. D., D. Zamir, M. W. Ganal, and S. D. Tanksley. 1988. Use of isogenic lines and simultaneous probing to identify DNA markers tightly linked to the tm-2a gene in tomato. Genetics 120: 579-585.

Zhang, Z., and T. Druet. 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. Journal of Dairy Science 93: 5487-5494.

Zhang, Z., E. Ersoz, C.-Q. Lai, R. J. Todhunter, H. K. Tiwari, M. A. Gore *et al.* 2010. Mixed linear model approach adapted for genome-wide association studies. Nature Genetics 42: 355-360.

Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu *et al.* 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biology 10: R42.

SUMMARY

The rapid decrease in costs of DNA sequencing implies that whole-genome sequence data will be widely available in the coming few years. Whole-genome sequence data includes all base-pairs on the genome that show variation in the sequenced population. Consequently it is assumed that the causal mutations (e.g. quantitative trait loci; QTL) are included in the data. Compared to currently used marker array panels which only represent a small number of variations (e.g. single nucleotide polymorphisms; SNPs), whole-genome sequence allows testing a given trait directly for association with the QTL, which might lead to discovery of new QTL or higher accuracy in genomic predictions compared to SNP panel genotypes. The main aim of this thesis was to investigate the benefits of using whole-genome sequence data in breeding of animals and plants compared to currently available marker panels. The thesis focusses on genotype imputation, QTL detection, and genomic prediction. First the benefits of using whole-genome sequence data were studied in cattle with respect to genotype imputation and genomic prediction. Thereafter the use of (imputed) whole-genome sequence data for QTL detection in tomato (*S. Lycopersicum*) was studied. Finally, the results from the previous chapters are discussed and some future perspectives are given with respect to successful application of whole-genome sequence data.

Chapter 2 describes the work of a large consortium to collect whole-genome sequence data for 234 individuals. In total 28.3 million variants were identified. One goal of this collaboration was to build a dataset of sequenced individuals that can be used as reference for imputation of sequence data in cattle populations. Accuracy of imputation of the sequence variants into larger data sets genotyped with SNP arrays was assessed using fivefold cross-validation, as well as potential benefits of whole-genome sequence data with respect to genome wide association studies were demonstrated.

In **Chapter 3** the accuracy of imputation of genotypes from two commercially used SNP panels to whole-genome sequence data in Holstein Friesian dairy cattle was studied in more detail for chromosome 1. Accuracy of imputation to whole-genome sequence data (1,737,471 SNPs) was generally high for imputation from the BovineHD beadchip (40,492 SNPs), but was low from the BovineSNP50 beadchip (3,132 SNPs)**.** Stepwise imputation from the BovineSNP50 to the BovineHD beadchip and then to sequence data substantially improved accuracy of imputation. Factors that decreased accuracy of imputation to sequence

data were low linkage disequilibrium between an imputed SNP and the SNP on the lower density panel, low minor allele frequency of the imputed SNP and small size of the reference group.

Accordingly, in **Chapter 4** whole-genome sequence data (12,590,056 SNPs) was imputed into BovineHD genotypes (631,428 SNPs) of 5503 Holstein Friesian bulls. Reliability of genomic prediction based on this imputed whole-genome sequence data was compared to genomic prediction using BovineHD genotypes. For genomic prediction two different methods were tested: with the first method it was assumed a priori that the variance of SNP effects is equal (GBLUP); while with the second method it was assumed a priori that the effects of many SNPs is very small or zero, and for only a few SNPs it is large (Bayes-SSVS). The results showed no advantage of using whole-genome sequence data compared to BovineHD genotypes for genomic prediction. Apparently, the approaches used currently for genomic prediction in dairy cattle were not optimal to capitalize on the potential provide by whole-genome sequence data. A training set with a larger number of individuals that are distantly related to each other and genomic prediction models that incorporate biological information on the SNPs or that apply stricter SNP pre-selection should be considered.

In **Chapter 5** we investigated whether the use of whole-genome sequence data (430,431 SNPs) would reveal additional QTL, relative to a SNP array panel (1,053 SNPs), when performing QTL analyses in a tomato population of recombinant inbred lines (*S. lycopersicum* x *S. pimpinellifolium*). Based on standard theory for linkage disequilibrium decay in bi-parental mapping populations, increasing marker density of marker panels with few thousand SNPs to whole-genome sequence data was not expected to provide additional power to detect QTLs. However, in contrast to these expectations more QTLs were found when using the sequence data, indicating that the linkage disequilibrium properties in the RIL population deviated from those expected.

Chapter 6 describes a genome wide association study with imputed whole-genomes sequence data (107,509 SNPs) for 145 accessions (*Solanum lycopersicum* and related wild species). Despite the relatively poor accuracy of imputation (correlation between the original and imputed genotypes was on average 0.34) more significant SNPs (>65 SNPs in 9 regions) were found using the imputed sequence compared to using the SNP array panel (no significant SNPs). Most of

these regions were linked to previously identified QTL and candidate genes. These results show that even with a low number of sequenced accessions (n = 38), and thus limited imputation accuracy, the power of a GWAS in an association panel can still be improved.

Finally, in **Chapter 7** the current status for dairy cattle breeding and tomato breeding with respect to the use of whole-genome sequence data is shortly described. Especially in cattle breeding a lot has happened the last few years, since the collection of whole-genome sequence data of many individuals (**Chapter 2**). In the last part of this chapter future perspectives with respect to successful application of whole-genome sequence data in plant and animal breeding were discussed: genotype imputation, QTL detection, and genomic prediction. By clever choosing individuals to be sequenced and development of the right imputation algorithm (that considers and possibly takes advantage of the specific characteristics of the population structure), genotype imputation will remain important to obtain accurate sequence data for all individuals in a cost effective way. Applying this imputed sequence data increases power of QTL detection in RIL populations, association panels or outbred populations, but the size of the power increase depends on the studied population. The challenges for the future with respect to genomic prediction using whole-genome sequence data are to unravel the biological background of traits and functional annotations of DNA; and develop statistical models that incorporate this biological information and that can efficiently handle large datasets. Added value of whole-genome sequence data in genomic prediction will be limited, until these issues are largely solved.

SAMENVATTING

De kosten voor het verkrijgen van de volledige DNA sequentie (het complete genoom) van een individu dalen snel. Deze prijsdaling betekent dat in de komende jaren (veel) meer DNA sequenties beschikbaar komen. Een groot deel van de DNA sequentie is gelijk voor alle individuen en een ander deel varieert in een populatie. Het deel dat varieert zijn onder andere SNPs. Sommige van deze SNPs zijn verantwoordelijk voor verschillen in eigenschappen tussen individuen en worden ook wel QTL genoemd. De huidige DNA analyses zijn gebaseerd op enkele honderden of duizenden SNPs en vertegenwoordigen maar een klein gedeelte van de totale variatie in het DNA. Omdat de volledige DNA sequentie van een individu alle SNPs bevat, wordt aangenomen dat deze ook alle QTL bevat. Dit maakt het mogelijk om nieuwe QTL op te sporen of om een hogere betrouwbaarheid te realiseren in het voorspellen van kenmerken in vergelijking met de huidige DNA analyses. In dit proefschrift zijn de voordelen van het gebruik van de volledige DNA sequentie ten opzichte van de huidige DNA analyses onderzocht in veefokkerij en plantveredeling. Hierbij ligt de focus op genotype imputatie, QTL detectie en het voorspellen van kenmerken met behulp van DNA informatie. Als eerste zijn deze voordelen onderzocht in melkvee en daarna in tomaten. Het laatste hoofdstuk bevat enkele perspectieven met betrekking tot het toepassen van DNA sequentie data in de toekomst.

Hoofdstuk 2 beschrijft het werk van een groot consortium dat volledige DNA sequenties van 234 dieren heeft verzameld. In totaal zijn 28,3 miljoen baseparen gevonden op het DNA die varieerden tussen de dieren. Eén van de doelen van dit consortium is het vormen van een dataset met volledige DNA sequenties van een groot aantal dieren. Dit dataset kan gebruikt worden als referentie voor het imputeren (schatten) van de volledige DNA sequentie van andere dieren in de rundveepopulatie. De betrouwbaarheid van imputatie varieerde over het genoom, met name SNPs met een lage frequentie waren moeilijk te imputeren. Naast imputatie is de potentie van het gebruik van volledige DNA sequentie in associatie studies gedemonstreerd.

In **Hoofdstuk 3** is imputatie van de volledige DNA sequentie vanuit twee commerciële SNP panels in meer detail bestudeerd voor chromosoom 1 bij Holstein Friesian melkvee. De betrouwbaarheid van imputatie van 40,492 SNPs (BovineHD panel) naar volledige DNA sequentie (1,737,471 SNPs) was hoog, in tegenstelling tot imputatie van 3,132 SNPs (BovineSNP50 panel) naar volledige

DNA sequentie. Een stijging in betrouwbaarheid was zichtbaar met imputatie van BovineSNP50 via BovineHD naar volledige DNA sequentie. Factoren die imputatie betrouwbaarheid beïnvloeden zijn de relatie tussen de SNPs, de frequentie van de geïmputeerde SNP en de grootte van de referentiegroep (individuen met volledige DNA sequentie beschikbaar).

In een vervolg studie (**Hoofstuk 4**) hebben we geïmputeerde DNA sequentie data (12,590,056 SNPs) van 5503 Holstein Friesian stieren gebruikt voor het voorspellen van enkele kenmerken. De betrouwbaarheid van deze voorspellingen is vergeleken met de betrouwbaarheid van voorspellingen waarbij gebruik gemaakt worden van 631,428 SNPs (BovineHD panel). Daarnaast zijn twee verschillende methodes gebruikt: een methode die op voorhand er vanuit gaat dat de variantie in SNP effecten gelijk is (GBLUP); en een methode die er vanuit gaat dat de meeste SNPs een klein effect hebben en een paar SNPs een groot effect (Bayes-SSVS). De resultaten lieten geen voordeel zien bij het gebruik van de volledige DNA sequentie. Waarschijnlijk zijn de gebruikte methodes niet in staat de potentie van volledige DNA sequentie te gebruiken. Minder gerelateerde individuen en andere methodes die gebruik kunnen maken van biologische informatie of een strengere SNP selectie procedure zijn het overwegen waard.

In **Hoofdstuk 5** hebben we onderzocht met behulp van de volledige DNA sequentie (430,431 SNPs) in een tomaten inteelt lijn (*S. lycopersicum* x *S. pimpinellifolium*) extra QTL gevonden worden in vergelijking met een SNP panel van 1,053 SNPs. Op basis van de huidige theorieën met betrekking tot relaties tussen SNPs in inteelt lijnen was verwacht dat dit geen extra QTL zou opleveren. In tegenstelling tot deze verwachtingen werden juist wel extra QTL gevonden. Dit resultaat impliceert dat de huidige aannames mogelijk onjuist zijn.

Naast een tomaten inteelt lijn, hebben we in **Hoofdstuk 6** gekeken naar het gebruik van volledige DNA sequentie in een associatie panel met 145 accessies (*Solanum lycopersicum* en aanverwante wilde soorten). Voor deze accessies hadden we geïmputeerde DNA sequentie data (107,509 SNPs) beschikbaar. Ondanks de beperkte imputatie betrouwbaarheid (de gemiddelde correlatie tussen de originele en geïmputeerde SNPs was 0.34) vonden we meer SNPs met een associatie met het kenmerk (>65 SNPs in 9 regio's) in vergelijking met wanneer we een SNP panel gebruikten (geen significante SNPs). De meeste van de gevonden regio's zijn gelinkt aan QTL gevonden in vorige studies en kandidaat genen. Deze resultaten

laten zien dat het gebruik van volledige DNA sequentie data de kracht van een associatie studie kan verhogen, zelfs met een klein aantal individuen met originele DNA sequentie data (n = 38) en daardoor een beperkte imputatie betrouwbaarheid.

In het laatste hoofdstuk (**Hoofstuk 7**) is kort de huidige status in melkveefokkerij en tomaten veredeling beschreven met betrekking tot het gebruik van volledige DNA sequentie data. Met name in de veefokkerij is veel gebeurd de afgelopen jaren sinds de start van het verzamelen van DNA sequenties (**Hoofdstuk 2**). Het laatste gedeelte van het hoofdstuk bevat enkele toekomst perspectieven met betrekking tot het toepassen van DNA sequentie data in veefokkerij en plantveredeling. De focus ligt hierbij op genotype imputatie, QTL detectie en het voorspellen van kenmerken met behulp van DNA informatie. Imputatie blijft belangrijk om op een relatief goedkope wijze betrouwbare DNA sequenties van veel individuen te verkrijgen. Hierbij is het belangrijk om goed te kijken voor welke individuen de volledige DNA sequentie wordt verkregen en is de ontwikkeling van goede imputatie methoden van belang (omgang met specifieke populatie structuur karakteristieken). Afhankelijk van de populatie, zal deze geïmputeerde DNA sequentie data de kracht van QTL detectie in inteelt lijnen, associaties panels en natuurlijke populaties verhogen. De uitdagingen in de toekomst met betrekking tot voorspellen van kenmerken met behulp van volledige DNA sequentie data liggen in (1) het ontrafelen van de biologische achtergrond van kenmerken; (2) ontwikkeling van statistische methoden die deze biologische informatie meenemen en efficiënt kunnen omgaan met de grote hoeveelheid data. Zolang deze uitdagingen niet grotendeels zijn opgelost, zal de toegevoegde waarde voor het voorspellen van kenmerken met behulp van volledige DNA sequentie data beperkt blijven.

CURRICULUM VITAE

- About the author
- Over de auteur
- List of publications

About the author

Rianne van Binsbergen was born on the 8th of September 1988 in Lochem, the Netherlands. In 2006, she graduated from high school Staring College in Lochem. In the same year, she started her study Animal Sciences at Wageningen University. Rianne followed the research master variant, and specialized in Animal Health and Management and Animal Breeding and Genetics. For her specialization in Animal Health and Management, she performed a minor thesis about the effect of antimicrobial treatment of recently acquired subclinical mastitis on occurrence of clinical mastitis in Dutch dairy cows. For her specialization in Animal Breeding and Genetics Rianne performed her major thesis at Wageningen UR Livestock Research. She studied SNP effects underlying the makeup of the genetic correlations and covariances between milk production traits in dairy cattle. In 2011, Rianne went to CRV Ambreed in New Zealand to perform a minor thesis. In New Zealand, she investigated the error rate for imputation from lower density marker panels to a high density marker panel in a multi-breed New Zealand dairy cattle dataset. After finishing her MSc in 2012, she started her PhD at Wageningen University & Research. The project was a collaboration between the Animal Breeding and Genomics and Biometris. The results of this research are described in this thesis. Next to her PhD, Rianne became in 2012 owner of a dairy farm in Lochem together with her parents, brother, and partner.

Over de auteur

Rianne van Binsbergen is op 8 september 1988 geboren te Lochem. In 2006 heeft ze haar VWO diploma behaald op het Staring College in Lochem. In hetzelfde jaar is ze gestart met de studie Dierwetenschappen aan Wageningen Universiteit. Rianne heeft de 'research master variant' gevolgd met de specialisaties Animal Health Management en Fokkerij en Genetica. Voor de specialisatie Animal Health Management heeft ze een klein afstudeervak gedaan over het effect van het optreden van klinische mastitis na antibioticabehandelingen bij subklinische mastitis bij Nederlands melkvee. Voor de specialisatie Fokkerij en Genetica heeft Rianne een groot afstudeervak gedaan bij Wageningen UR Livestock Research. Hier heeft ze gekeken naar de onderliggende SNP effecten in de opbouw van genetische correlaties en covarianties tussen melkproductie-eigenschappen in melkvee. In 2011 heeft Rianne een kleine afstudeervak gedaan bij CRV Ambreed in Nieuw Zeeland. Hier heeft ze genotype imputatie bestudeerd over meerdere rassen in een Nieuw Zeelandse melkveedataset. Na het afronden van haar MSc in 2012, is Rianne gestart met haar PhD bij Wageningen University & Research. Haar PhD project was een samenwerking tussen Animal Breeding and Genomics en Biometris. De resultaten van dit onderzoek zijn beschreven in dit proefschrift. Naast haar PhD, is Rianne per 2012 mede-eigenaar van een melkveebedrijf in Lochem, samen met haar ouders, broer en partner.

Peer reviewed publications

van Binsbergen, R., M. P. L. Calus, M. C. A. M. Bink, F. A. van Eeuwijk, C. Schrooten, and R. F. Veerkamp. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution 47: 71.

Eynard, S. E., J. J. Windig, G. Leroy, **R. van Binsbergen**, and M. P. L. Calus. 2015. The effect of rare alleles on estimated genomic relationships from whole genome sequence data. BMC Genetics 16: 1-12.

van Binsbergen, R., M. C. A. M. Bink, M. P. L. Calus, F. A. van Eeuwijk, B. J. Hayes, I. Hulsegge, and R. F. Veerkamp. 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution 46: 41.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, **R. van Binsbergen**, R. F. Brøndum *et al.* 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics 46: 858–865.

van Binsbergen, R., R. F. Veerkamp, and M. P. L. Calus. 2012. Makeup of the genetic correlation between milk production traits using genome-wide single nucleotide polymorphism information. Journal of Dairy Science 95: 2132-2143.

Conference proceedings, abstracts and presentations

van Binsbergen, R., M. C. A. M. Bink, H.J. Finkers, M. P. L. Calus, R. F. Veerkamp, and F. A. van Eeuwijk. 2016. Utilizing low-coverage sequence data in tomato recombinant inbred lines (*S. lycopersicum* x *S. pimpinellifolium*). 20th Eucarpia General Congress, Zurich, Switzerland.

Veerkamp, R. F., **R. van Binsbergen**, M. P. L. Calus, C. Schrooten, and A. C. Bouwman. 2015. Comparing genomic prediction and GWAS with sequence information vs HD or 50k SNP chips. 66th Annual meeting of the European Association of Animal Production, Warsaw, Poland.

van Binsbergen, R., M. P. L. Calus, M. C. A. M. Bink, F. A. van Eeuwijk, and R. F. Veerkamp. 2015. Accuracy of imputation and prediction in dairy cattle using whole-genome sequence data. International Biometric Society Channel Meeting, Nijmegen, the Netherlands.

van Binsbergen, R., M. P. L. Calus, M. C. A. M. Bink, F. A. van Eeuwijk, and R. F. Veerkamp. 2015. Is the use of whole-genome sequence data in dairy cattle the future? WIAS Science Day, Wageningen, the Netherlands.

van Binsbergen, R., M. P. L. Calus, M. C. A. M. Bink, C. Schrooten, F. A. van Eeuwijk, and R. F. Veerkamp. 2014. Genomic prediction with 12.5 million SNPs for 5503 Holstein Friesian bulls. 10th World Congress on Genetics Applied to Livestock Production, Vancouver, Canada.

Eynard, S. E., J. J. Windig, G. Leroy, E. Verrier, S.J. Hiemstra, **R. van Binsbergen**, and M. P. L. Calus. 2014. The use of whole genome sequence data to estimate genetic relationships including rare alleles information. 10th World Congress on Genetics Applied to Livestock Production, Vancouver, Canada.

van Binsbergen, R., M. P. L. Calus, M. C. A. M. Bink, C. Schrooten, F. A. van Eeuwijk, and R. F. Veerkamp. 2014. Added value of whole-genome sequence data to genomic predictions in dairy cattle. XXVII International Biometric Conference, Florence, Italy.

van Binsbergen, R., M. C. A. M. Bink, M. P. L. Calus, F. A. van Eeuwijk, B. J. Hayes, I. Hulsegge, and R. F. Veerkamp. 2013. Genotype imputation accuracy in Holstein Friesian cattle in case of whole-genome sequence data. 64th Annual meeting of the European Association of Animal Production, Nantes, France.

van Binsbergen, R., M. C. A. M. Bink, M. P. L. Calus, F. A. van Eeuwijk, and R. F. Veerkamp. 2013. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. International Plant & Animal Genome XXI, San Diego, CA, USA.

Schrooten, C., **R. van Binsbergen**, P. Beatson, and H. Bovenhuis. 2012. Imputation from lower density marker panels to BovineHD in a multi-breed dataset. 63th Annual meeting of the European Association of Animal Production, Bratislava, Slovakia.

Veerkamp, R. F., **R. van Binsbergen**, and M. P. L. Calus. 2012. The genetic correlation between traits using genome-wide SNP information. International Plant & Animal Genome XX, San Diego, CA, USA.

TRAINING AND SUPERVISION PLAN

Training and Supervision Plan

The Basic Package (3 ECTS)	**Year**
WIAS Introduction Course	2013
Ethics and Philosophy in Life Sciences	2013
Scientific Exposure (19 ECTS)	
International conferences (10 ECTS)	
Annual meeting of the European Association of Animal Production (EAAP), Bratislava, Slovakia	2012
International Plant & Animal Genome (PAG) XXI, San Diego, CA, USA	2013
Interbull meeting, Nantes, France	2013
Annual meeting of the European Association of Animal Production (EAAP), Nantes, France	2013
XXVII International Biometric Conference (IBC), Florence, Italy	2014
10th World Congress on Genetics Applied to Livestock Production (WCGALP), Vancouver, Canada	2014
International Biometric Society (IBS) Channel Meeting 2015, Nijmegen, the Netherlands	2015
20th Eucarpia General Congress, Zurich, Switzerland	2016
Seminars and workshops (2 ECTS)	
Fokkerij en Genetica Connection, Vught	2012
WIAS Science Day 2013, Wageningen	2013
Genomic selection for novel traits, Wageningen	2013
WIAS Science Day 2014, Wageningen	2014
Fokkerij en Genetica Connection, Ellecom	2014
WIAS Science Day 2015, Wageningen	2015
Genomic prediction using multiple breeds, Wageningen	2016

Presentations (7 ECTS)

Poster presentation at PAG, San Diego, CA, USA	2013
Oral presentation at EAAP, Nantes, France	2013
Oral presentation at IBC, Florence, Italy	2014
Poster presentation at WCGALP, Vancouver, Canada	2014
Oral presentation at WIAS Science Day, Wageningen	2015
Poster presentation at IBS Channel Meeting, Nijmegen	2015
Oral presentation at Eucarpia General Congress, Zurich, Switzerland	2016

In-Depth Studies (9 ECTS)

Disciplinary and interdisciplinary courses (7 ECTS)

Post-Graduate Training School: 'Identity By Descent (IBD) approaches to genomic analyses of genetic traits', Wageningen	2012
Advanced methods and algorithms in animal breeding with focus on genomic selection, Wageningen	2012
Sequence Data Analysis Training school, SDAC, Wageningen	2012
Genetic analysis using ASReml4.0, Wageningen	2014
Introduction to theory and implementation of genomic selection, Wageningen	2014

PhD students' discussion groups (2 ECTS)

Quantitative Genetics Discussion Group	2012-2016
Biometris StatGen meetings	2012-2016

Professional Skills Support Courses (4 ECTS)

PhD Competence assessment	2012
Course Techniques for Writing and Presenting a Scientific Paper	2013
Effective behaviour in your professional surroundings	2014
WGS PhD Workshop Carousel	2015
Presenting with Impact	2016
Last Stretch of the PhD Programme	2016

Research Skills Training (7 ECTS)	
ABGC Course: Getting started in AS-Reml	2011
Preparing own PhD research proposal	2012
Introduction to R for Statistical Analysis	2013
Didactic Skills Training (3 ECTS)	
Lecture "Imputation of genotype data" during course "Introduction to theory and implementation of genomic selection"	2014
Assist practical Modern Statistics for the Life Sciences	2014
Assist practical Statistics 1	2016
Management Skills Training (2 ECTS)	
Organisation Quantitative Genetics Discussion Group	2013-2014
Organisation Biometris PhD Day	2016
Education and Training Total	**45 ECTS**

DANKWOORD

Na vijf jaar is het zover, mijn proefschrift is klaar en ik mag dit dankwoord schrijven. Ook al staat alleen mijn naam voor op dit boekwerk, het schrijven van een proefschrift doe je niet alleen. Daarom wil ik iedereen bedanken die mij op welke manier dan ook de afgelopen vijf jaar heeft geholpen, waaronder een paar personen in het bijzonder.

Allereest natuurlijk mijn begeleiders tijdens de afgelopen jaren. Net als voor mij, was de samenwerking tussen dierfokkerij en statistische genetica/plant veredeling voor hen ook een uitdaging. Roel, bedankt dat je na mijn MSc thesis nog ruim vier jaar met mij wilde samenwerken en bedankt voor je hulp met het stellen van deadlines. Mario, wat was het fijn om jou in mijn begeleidingscommissie te hebben, je hebt of maakt bijna altijd tijd om vragen te beantwoorden of stukken na te kijken. Fred, bedankt voor je enthousiasme en kennis op het gebied van statistische genetica en plant veredeling. Marco, bedankt voor je begeleiding tijdens de eerste paar jaar. Met mijn achtergrond in dierfokkerij was jouw kennis over zowel dierfokkerij als plantveredeling erg welkom. Als laatste wil ik Richard bedanken voor de begeleiding tijdens de laatste paar jaar en de hulp met de tomatendata. Ondanks de drukke agenda, maakte je wel tijd voor mij.

Mijn PhD was een samenwerking tussen verschillende groepen binnen Wageningen University & Research. Naast een uitdaging op inhoudelijk gebied, leverde dit ook een bijzonder contract en daarmee een uitdaging op administratief gebied op. Daarom wil ik graag de secretaresses van Wageningen Livestock Research Genomica in Lelystad en Wageningen, de secretaresses bij Animal Breeding and Genomics, de secretaresses bij Biometris en Lucia heel erg bedanken voor hun hulp.

De eerste twee jaar van mijn PhD periode heb ik doorgebracht bij Wageningen Livestock Research in het Triton gebouw. Myrthe en Krista bedankt dat jullie met mij het kantoor wilden delen en voor jullie gezelligheid. Heel bijzonder dat ieder gesprek met jullie uiteindelijk weer over paarden gaat. Also thanks to Marcin and Sonia for sharing the office with me. Marcin, sorry that I occupied your desk. Sonia thanks for being a great co-genius! Ook de andere collega's binnen Triton bedankt voor de leuke tijd daar!

Na twee jaar werd het tijd om te verhuizen: de ABG collega's naar Radix Oost en ik naar Biometris in Radix West. De collega's bij Biometris wil ik bedanken voor de gezellige koffiepauzes en jullie interesse in ons melkveebedrijf. Jullie zijn altijd welkom om een keer te komen kijken in Lochem! Also thanks to the other PhD candidates at Biometris for nice lunch breaks.

In het bijzonder wil ik Yvonne, Yvette en Aniek bedanken, voor jullie gezelligheid gedurende mijn hele PhD periode, ondanks de 'grote' afstand tussen onze kantoren een deel van de tijd. Yvonne, wij kennen elkaar ondertussen al ruim 10 (!) jaar. Bijzonder om te zien hoe we beiden zijn veranderd gedurende onze studententijd en de periode daarna. Wie had 10 jaar geleden bedacht dat jij zo'n wereldreiziger zou worden? Yvette, bedankt voor de gezelligheid en adviezen. Je bent toch wel een beetje "mama Yvette", ook al mag ik je zo niet noemen. Aniek, bedankt voor de inhoudelijke discussies en gezelligheid. Daarnaast is het erg bijzonder dat jullie familie-uitbreiding bijna gelijk loopt met die van ons. Yvonne en Yvette, leuk dat jullie tijdens de verdediging ook bij mij op het podium willen zitten!

Niet alleen op het werk, maar ook buiten het werk heb ik veel hulp gehad. Allereest wil ik alle vrienden bedanken voor de gezellige afleiding die jullie mij hebben gegeven. Michiel bedankt voor de hulp met de verbouwing, en Anne bedankt voor de gezellige maandagochtenden in de auto. Daarnaast wil ik Rike heel erg bedanken voor haar steun en hulp de laatste jaren, ondanks dat dit soms zwaar was voor haar. Helaas, kan ze het afronden van mijn PhD niet meemaken. Ook mijn ouders wil ik heel erg bedanken voor hun steun en voor het opvangen van Bennie en Emma. Het is heel fijn dat jullie altijd klaarstaan voor ons.

Als laatste wil ik Emma en Bennie bedanken. Emma, ik ben heel blij dat jij in mijn leven bent gekomen en in de afgelopen twee jaar heb ik al heel veel van je geleerd (onder andere "Nee!" zeggen). Ik hoop dat je een trotse en lieve grote zus mag worden. En Bennie… bedankt voor alles!

Rianne

Colophon

The research described in this thesis was funded equally by Biometris (Wageningen University & Research), and the Breed4Food consortium (BO-22.04-011-001-ASG-LR), a public-private partnership in the domain of animal breeding and genomics.

The cattle data used in this thesis were provided by the 1000 bull genomes consortium (Chapter 2-4) and CRV B.V., Arnhem, the Netherlands (Chapter 4).

The cover of this thesis was designed by Rianne van Binsbergen

This thesis was printed by Digiforce, Vianen, the Netherlands

PROPOSITIONS

1. In contrast to expectations, whole-genome sequence data do not improve genomic prediction in Holstein Friesian cattle. (this thesis)

2. In contrast to expectations, whole-genome sequence data increase power of QTL mapping in recombinant inbred lines. (this thesis)

3. Multidisciplinary research improves disciplinary research.

4. Grazing of cows improves wellbeing of Dutch citizens more than wellbeing of cows.

5. Joining coffee breaks with colleagues should be mandatory.

6. Unequal treatment of people is unavoidable.

Propositions belonging to the thesis entitled:
"Prospects of whole-genome sequence data in animal and plant breeding"

Rianne van Binsbergen
Wageningen, 5 July 2017