Studies on genomic prediction for carcass traits in Japanese Black cattle

SHINICHIRO OGAWA

2017

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	**: Description**
AI	: Artificial insemination
A matrix	: Additive relationship matrix
BLUP	: Best linear unbiased prediction
BMS	: Beef marbling standard
BTA	: Bos Taurus autosome
BTX	: Bos Taurus X chromosome
BV	: Breeding value
CW	: Carcass weight
GBLUP	: Genomic best linear unbiased prediction
GE	: Genomic evaluation
GEBV	: Genomic estimated breeding value
G matrix	: Genomic relationship matrix
GP	: Genomic prediction
GS	: Genomic selection
GWAS	: Genome-wide association study
LASSO	: Least Absolute Shrinkage and Selection Operator
LD	: Linkage disequilibrium
MABLUP	: Marker-assisted best linear unbiased prediction
MAF	: Minor allele frequency
MAS	: Marker-assisted selection
MS	: Marbling score

Ne	: Effective population size
PBV	: Predicted breeding value
QTL	: Quantitative trait locus
REML	: Restricted maximum likelihood
SD	: Standard deviation
SNP	: Single nucleotide polymorphism

CHAPTER ONE
General introduction

There are four Japanese domestic beef cattle (Wagyu) breeds: Japanese Brown, Japanese Shorthorn, Japanese Polled, and Japanese Black. The Japanese Black is the primary Wagyu breed, and is well known to excel in meat quality, especially in marbling. In Japan, native Japanese cattle were crossed with British and Continental breeds over a period of 10 years in the early 1900s; then, the four Wagyu breeds had been fixed through strict selection over many years under a completely closed breeding system, which excluded crossing among the them (Namikawa, 1992). Since the relaxation of beef import restrictions in Japan in 1991, beef quality traits, including the degree of marbling, have received more emphasis in domestic production of Japanese Black cattle. In the same year, genetic evaluation of carcass traits using a mixed model methodology, or best linear unbiased prediction (BLUP) (Henderson, 1973), was introduced using relevant field data collected at carcass markets located across Japan (Wagyu Registry Association, 2007). Subsequently, while there has been steady genetic improvement attained in carcass traits, because of intensive use of few sires with higher predicted breeding values (PBVs) for degree of marbling, it is known that there is a sharp decline occurred in effective population size (Ne) of this breed (Nomura et al., 2001).

In more recent years, on-farm progeny testing for carcass traits of young bulls has relied on the genetic evaluation system started in 1991. The model used in the evaluation is a single-trait animal model that includes additive genotypic values, namely breeding values (BVs), of animals as random variables, with incorporating an additive

relationship matrix (A matrix: Henderson, 1975; Quaas, 1976) among animals as the covariance structure, which is derived from pedigree information. This evaluation procedure is the so-called two-step procedure, which is based on a method developed by Ashida and Iwaisaki (1998, 1999) that consists of the average-information algorithm (Johnson and Thompson, 1995) of the restricted maximum likelihood (REML) method (Patterson and Thompson, 1971) for variance component estimation as the first step and the empirical BLUP as the second step. Assuming the infinitesimal model (Fisher, 1918; Bulmer, 1980), the whole genome is targeted in this genetic evaluation, but the whole genome itself is treated as an unobserved black box; therefore, effects of any of quantitative trait loci (QTLs) for the target trait are not explicitly considered in the model. Although an obvious genetic trend for each of carcass traits in recent years has been largely contributed by this genetic evaluation (Wagyu Registry Association, 2007), carcass record collecting from relatives, including progeny, for the genetic evaluation system is heavily cost- and time-consuming.

In the late twentieth-century, breakthroughs occurred in molecular biology and genetic engineering that established the technological basis for modern genomics and biotechnology. This facilitated development of the concepts and practices of QTL mapping and marker-assisted selection (MAS), and also motivated to develop the methodology to integrate molecular marker information into conventional genetic evaluation systems. Fernando and Grossman (1989) proposed a BLUP procedure of genetic merit using a gametic effect model that can fundamentally treat only one marker information. This kind of BLUP is called marker-assisted BLUP (MABLUP), as named by Saito and Iwaisaki (1997b). Goddard (1992) expanded this approach to treat two successive markers with the assumption that a QTL is flanked by them. Continued

effort was made to further develop and improve the MABLUP methodology: reduced animal model and combined-merit model (e.g., Cantet and Smith, 1991; Saito and Iwaisaki, 1996; Saito et al., 1998); total merit model (van Arendonk et al., 1994; Saito and Iwaisaki, 1997a, b); marker-haplotype model (Meuwissen and Goddard, 1996; Pagnacco and Jansen, 2001); chromosome segment model (Matsuda and Iwaisaki, 2001; Matsuda and Iwaisaki, 2002a, b); and mixed inheritance model (e.g., Meuwissen and Goddard, 1997; Almasy and Blangero, 1998). Generally, these models assume use of DNA polymorphisms that were previously detected as QTLs or markers of major genes that are thought to be strongly associated with traits based on the results of strict significance tests. For cattle, comprehensive microsatellite-based linkage maps were released in 1994 (Barendse et al., 1994; Bishop et al., 1994). In concept, successful MAS is expected to improve the rate of genetic gain (e.g., Soller, 1978; Soller and Beckmann, 1983; Smith and Simpson, 1986; Kashi et al., 1990; Meuwissen and van Arendonk, 1992). However, the impact of typical MAS on practical breeding programs has been smaller than initially envisaged, mainly because of the higher polygenic nature of economically important traits than expected, and there is limited availability of suitable DNA polymorphisms as markers for typical MAS (e.g., Dekkers, 2004; Bernardo, 2008). Today, there is a general consensus worldwide that most complex and quantitative traits are usually affected by a large number of small-effect genes (polygenes), and that accurate prediction of phenotypic and genotypic values requires concurrent consideration of a large number of genetic variants.

Meuwissen et al. (2001) is the ground-breaking article, proposing the idea of a new type of MAS that simultaneously treats all chromosome segments defined by adjacent markers such as genome-wide, high-density single nucleotide polymorphisms

(SNPs) with Bayesian procedures, regardless of whether each segment is significant in genome-wide association studies (GWASs). As stated in Gianola et al. (2009), the methods by Meuwissen et al. (2001) gained enormous attention in animal breeding, because 1) the procedures cope well with the so-called "small n, large p" situation, 2) marker-specific variances are allowed to vary at random over many loci, and 3) Bayesian methods have a natural way of taking into account uncertainty about all unknowns in a model and can be applied to almost any parametric statistical model when coupled with the power and flexibility of the Markov chain Monte Carlo method. Genome-wide, high-density markers are used with the expectation of tracing all underlying QTLs, or to explain all additive genetic variances for QTLs of a quantitative trait exploiting the status in linkage disequilibrium (LD) between QTLs and markers. Prediction of BV based on only genome-wide markers is referred to as genomic prediction or evaluation (GP or GE, respectively) of BV, and selection based on the result of GP is genomic selection (GS). GP consists of two steps; the effects of all markers used are estimated using individuals for which there was both phenotype and genotype data (denoted as training population), and the BVs of candidates for selection are predicted based on genotype data and the estimated marker effects.

Statistical approaches so far proposed for GP are classified into one of two types: the multiple-marker regression model approach and the conventional animal model approach with a relationship matrix among animals based on marker information, but not on pedigree information. Most currently popular models for the former are ones collectively dubbed as Bayesian alphabet models (e.g., Gianola et al., 2009), such as BayesA and BayesB (Meuwissen et al., 2001), BayesC(π) (Habier et al., 2011), and BayesR (Erbe et al., 2012). Ridge regression (Whittaker et al., 2000), least absolute

shrinkage and selection operator (LASSO: Tibshirani, 1996), and their Bayesian approaches (Park and Cassera, 2008; de los Campos et al., 2009) are also corresponded to the former type. These models differ in the prior distribution of marker effects adopted (e.g., de los Campos et al., 2013). For the latter, a relationship matrix among individuals based on genome-wide high-density markers can be referred to as a realized relationship matrix or genomic relationship matrix (G matrix: e.g., VanRaden 2008). This kind of approach was first proposed by Nejati-Javaremi et al. (1997). A commonly used procedure in this line is so-called genomic BLUP (GBLUP: Habier et al., 2007; VanRaden 2008), which is more applicable to the existing procedures of genetic evaluation systems. VanRaden (2008) stated that the G matrix can be obtained by at least three methods that he described, and the first of which has been widely used.

Sequencing of the cattle genome with a whole-genome shotgun approach began in December, 2003, and the first draft sequence (Btau_1.0), based on DNA taken from a Hereford dam L1 Dominette 01449, was released by the Baylor College of Medicine Human Genomes Sequencing Center (https://www.hgsc.bcm.edu/other-mammals/bovine-genome-project) in 2004. Efforts to improve the draft sequence have continued (Liu et al., 2009; Zimin et al., 2009). In parallel, taking together with the sequence of L1 Dominette 01449 as the reference bovine genome (Elsik et al., 2009), a large number of SNPs have also been generated from partial sequences of six breeds (Gibbs et al., 2009). Illumina BovineSNP50 BeadChip (Matukumalli et al., 2009) was the first high-density genotyping chip for an agricultural species, which was developed in December 2007, and included 54,001 bovine SNPs in the BovineSNP50 v.1 BeadChip that were selected in order to achieve even spacing across the bovine genome and possession of useful minor allele

frequencies (MAFs) in economically important cattle breeds, but not including the four Wagyu breeds. After developing the commercial SNP chip, GS following GP was introduced into routine genetic evaluation and selection of cattle, especially dairy cattle (e.g., VanRaden et al., 2009; Hayes et al., 2009); this is partly because of the possibility of reducing the costs of progeny testing schemes (Schaeffer, 2006). Very recently, Garcia-Ruiz et al. (2016) reported that, through the analysis of the US national dairy database, generation intervals have dramatically decreased over the past six years, especially in the sire-bull and sire-cow paths, and selection intensity for lowly heritable traits, such as somatic cell score, has considerably increased.

On the other hand, the adoption of GS in the beef industry has been slow compared with that in the dairy cattle industry (e.g., Garrick, 2011; Hayes et al., 2013; Van Eenennaam et al., 2014; Berry et al., 2016), even though the potential for GS to improve genetic gain in beef cattle is substantial mainly because reproduction, feed efficiency, meat quality, and carcass traits are key traits that contribute to profitability (Van Eenennaam et al., 2011; Pimentel et al., 2012). The GP accuracies for these traits in foreign beef breeds assessed using a within-breed training population are only low to moderate (e.g., Saatchi et al., 2011; Bolormaa et al., 2013; Boddhireddy et al., 2014; Boerner et al., 2014; Gunia et al., 2014; Neves et al., 2014; Lee et al., 2014; Chen et al., 2015; Fernandes Júnior et al., 2016); this is mainly due to smaller sizes of training populations. There are generally fewer sires with highly accurate PBVs in beef cattle than dairy cattle, because foreign beef industries usually do not use artificial insemination (AI) very heavily; therefore, it is much difficult to grow the size of within-breed training population for beef cattle. Additionally, unlike dairy cattle, there are a large number of important beef breeds, including crossbreds and even two

subspecies, *Bos taurus* and *Bos indicus*. To compensate for a small number of animals in within-breed training populations, combining data across countries and/or across breeds would be required. Additionally, higher-density marker panels would be required for GP using a multi-breed training population to allow accurate prediction as within-breed GP (de Roos et al., 2009; Kizilkaya et al., 2010), because LD between SNPs genotyped with the Illumina BovineSNP50 BeadChip and QTLs is not very consistent across breeds (de Roos et al., 2008). Unfortunately, the effective number of chromosome segments increases when a multi-breed training population is used, which indicates that an even larger training population is needed for accurate multi-breed GP (e.g., Goddard et al., 2010). Moreover, empirical studies have shown rather disappointing results; increase in prediction accuracy was smaller than expected and sometimes absent (e.g., Hayes et al., 2009; Pryce et al., 2011; Kachman et al., 2013; Khansefid et al., 2014), even when more than 600,000 markers genotyped with the Illumina BovineHD BeadChip were used (Harris et al., 2011; Erbe at al., 2012; Bolormaa et al., 2013). This is likely due to differences in LD phases between SNPs and QTLs across breeds (e.g., Hayes et al., 2013; Porto-Neto et al., 2015). Next challenge may be use of whole-genome sequencing data on many influential founder animals for multi-breed GP (Hayes et al., 2013), because the causative SNPs will be among the millions of polymorphisms that are genotyped. Consequently, GP accuracy will not rely on markers in LD with causative SNPs, but rather will conceptually include the causative SNPs themselves. In contrast to these situation for foreign beef breeds, there are no reports providing the information on GP accuracy for Japanese Black cattle.

Several factors have been reported to affect the results of GP, including training population size, Ne, the extent of LD, and marker density (e.g., Goddard and Hayes,

2009). A larger training population is generally preferable to accomplish more accurate GP; however, as is the case with other beef breeds, it seems to be not easy or difficult to provide a large-scale training population consisting of only sires with highly accurate PBVs for Japanese Black cattle. For the genetic evaluation of carcass traits in this breed, a possible alternative is to use fattened progenies with both phenotype and genotype data as a training population. A low Ne value was found in a Japanese Black cattle population using pedigree information, which was lower than those of most of the major western cattle breeds (Nomura et al., 2001). The lower Ne value could result in the observation of a higher extent of LD. Also, when higher-density SNP markers are used, there can be more chance to include SNPs in a higher degree of LD with QTLs or causative mutations themselves.

One key obstacle to achieving successful GP and GS is genotyping cost, because accurate GP usually requires more individuals in a training population with higher-density marker information; however, the cost of high-density SNP chips is still relatively high. For this, Habier et al. (2009) proposed the use of equally spaced low-density SNP panels in GP; this would be an attractive alternative if low-density SNP panels are found to work relatively well. This type of approach may be further powerful when also conducting the genotype imputation with an independent population as a haplotype reference population. Beagle (Browning and Browning, 2007) is one of the most widely used software programs for genotype imputation. A localized haplotype cluster model with a hidden Markov model is employed in Beagle to identify the most likely haplotypes based on the genotypes of individuals, and relationship information among individuals is not explicitly utilized. Beagle is commonly used in bovine datasets (Hozé et al., 2013), and has shown relatively better performance in

many cases (e.g., Nicolazzi et al., 2013).

Following the historical, previous and current backgrounds mentioned above, there may be a room that GP (potentially followed by GS) becomes a promising approach for breeding applications in Japanese Black cattle. The objective of this thesis was to provide the scientific bases for application of GP in this breed. Some relevant studies were conducted using phenotypic records of two economically important carcass traits, or carcass weight (CW) and marbling score (MS), and SNP markers genotyped by the commercial chip in Japanese Black fattened steers. In chapter two, using SNP data genotyped with the Illumina BovineSNP50 BeadChip, the level of whole-genome LD among SNPs was assessed, and the effects of densities of equally spaced genome-wide SNPs on genetic variance explained and GP accuracy were investigated, assuming two statistical models. In chapter three, using the same data as chapter two, accuracy of genotype imputation from different low(er)-density SNP markers with Beagle was examined in detail. In chapter four, the influences of the imputed genotypes on variance component estimation and GP were examined by comparing the relevant estimates with those obtained using all available SNPs and only SNPs selected for low(er)-density marker subsets without imputation. In chapter five, employing both the training population and independent validation population with PBVs provided by the routine genetic evaluation, GP accuracy was assessed using data of evenly spaced SNP marker subsets of various densities genotyped with the Illumina BovineHD BeadChip, in which two approaches to calculate the G matrix were used. Finally, in chapter six, an overall discussion is provided.

CHAPTER TWO

Effects of single nucleotide polymorphism marker density on degree of genetic variance explained and genomic evaluation for carcass traits in Japanese Black beef cattle

2.1 Introduction

Most economically important traits in beef cattle, including carcass traits, are controlled by QTLs, which usually have relatively small individual effects. For such traits, GE and GS, as proposed by Meuwissen et al. (2001), is expected to chase the QTLs simultaneously using SNP markers, given that at least one SNP is in LD with each QTL. In concept, successful GS is expected to accelerate genetic improvement by reducing the generation interval and increasing the accuracy of genetic evaluation (Meuwissen et al., 2013).

The recent development of various SNP chips has enabled high-throughput genotyping and allowed animal breeders to study and conduct GE and GS. By simulating 50,000 genome-wide high-density biallelic markers like SNPs, VanRaden (2008) showed better performance of BLUP using G matrix, relative to that using A matrix based on pedigree information (Henderson, 1984). In dairy cattle, GS has already been adopted in some countries and is an effective method for increasing the rate of genetic improvement. In beef cattle, on the other hand, its adoption has been slower, because the accuracy of the genomic estimated breeding value (GEBV) is much lower because of less availability of sires with highly accurate results in progeny tests.

Habier et al. (2009) proposed the use of lower-density and equally spaced SNP panels for effective GE, irrespective of trait. If such SNPs can explain substantial

proportions of genetic variations in carcass traits and be almost as effective as higher-density panels in evaluating GEBVs, their lower cost would make them useful, especially in beef breeding females. Traits that are measured after slaughter, as well as those that are difficult or expensive to record, are also traits for which GS could substantially improve genetic gain. However, for carcass traits, including degree of marbling in beef cattle, the effects of differing densities of SNPs used to estimate genetic variance and GE have been poorly studied.

The accuracy of GE depends on the extent of LD between SNP markers and QTLs, the number of animals with phenotypes and genotypes in the training population, the heritability, and the distribution of QTL effects for the trait (Hayes et al., 2009a). The first of these factors is closely related to Ne, and the density of SNP markers used that can be under the control of animal breeders. In this chapter, effects of density of equally spaced genome-wide SNPs on genetic variance explained and GE were investigated for carcass traits in beef cattle, using Japanese Black data and assuming two statistical models.

2.2 Materials and Methods

2.2.1 Ethics statement

Animal care and use was according to the protocol approved by the Shirakawa Institute of Animal Genetics Animal Care and Use Committee, Nishigo, Japan (ACUCH21-1).

2.2.2 Phenotype data

Cold CW and MS of 872 Japanese Black fattened steers, whose ages ranged between 15.3–43.0 months, were used for the current analyses. These records were

collected from 2000–2009 at two large meat markets in Japan, namely Tokyo Metropolitan Central Wholesale Market and Osaka Municipal South Port Wholesale Market. MS is the degree of marbling, ranging from null (1) to very abundant (12), assessed on the ribeye of the carcass dissected at the sixth and the seventh rib section, according to the Japan carcass grading standards (Japan Meat Grading Association, 1988). The distributions of CW and MS are shown in Figure 2-1. The mean (± standard deviation (SD)) was 496.6 (± 48.0) kg for CW and 6.8 (± 3.5) for MS.

2.2.3 Genotype data

DNA samples were extracted from perirenal adipose tissues. Sample DNA was quantified and genotyped using the Illumina BovineSNP50 BeadChip (hereafter referred to as the 50K chip in this chapter). The 50K assay contains 54,001 SNPs with an average probe spacing of 51.5 kb and a median spacing of 37.3 kb. A total of 38,502 SNPs was included in the statistical analyses, based on the following criteria: MAFs and genotype call rates were larger than 0.01 and 0.95, respectively, were in Hardy-Weinberg equilibrium ($p > 0.001$) and had position information. As a few percent of genotype data was missing, missing genotype filling was conducted using Beagle 3.3.2 package (Browning and Browning, 2011).

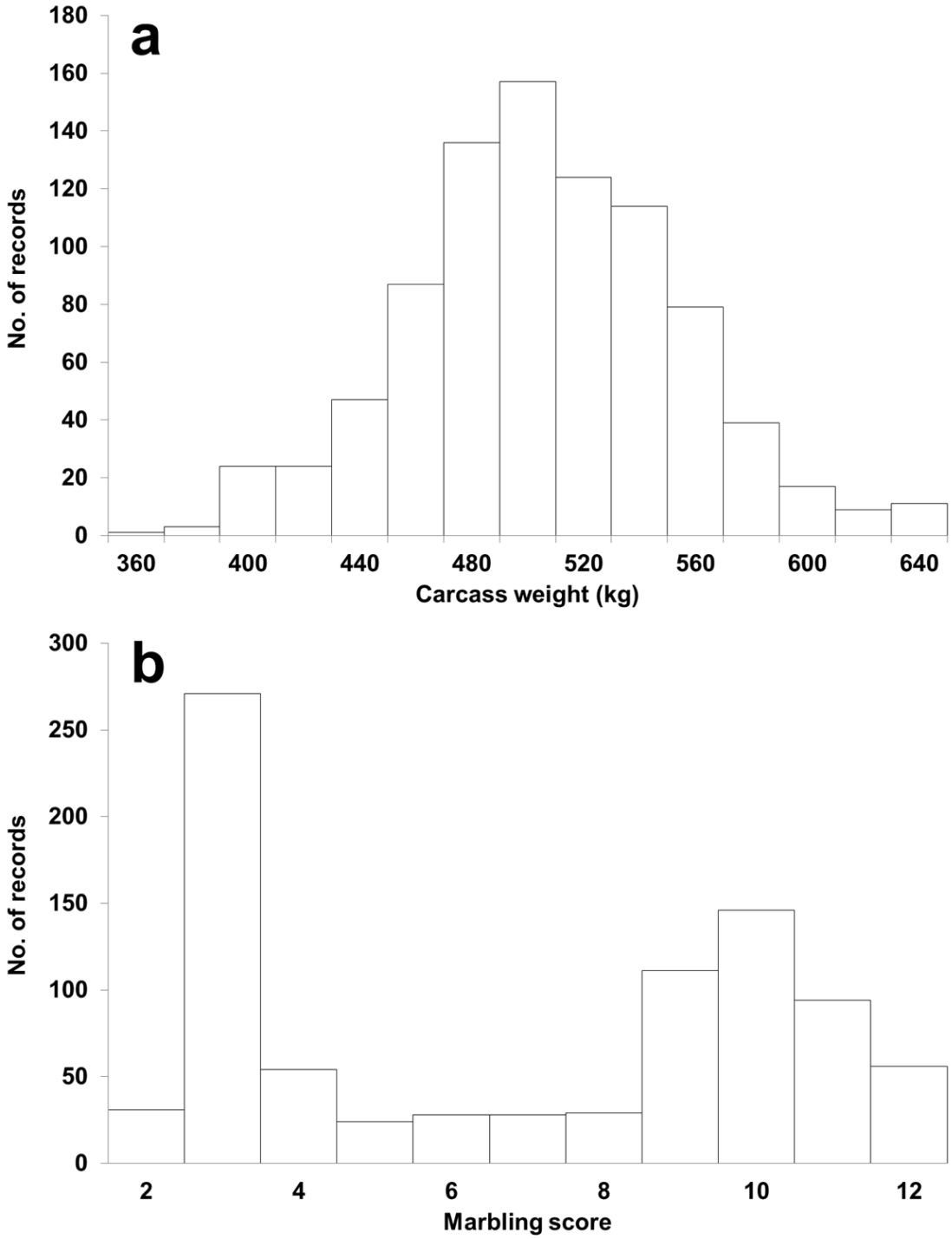

Figure 2-1. Distribution of carcass weights (a) and marbling scores (b).

2.2.4 Statistical analyses

Data were analysed using a following linear model (denoted as model 1 in this chapter):

$$\mathbf{y} = \mathbf{Xb} + \mathbf{g} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of records, $\mathbf{b}$ is the vector of fixed discrete effects of carcass market and year at slaughter and the continuous effects of the linear and quadratic covariates of month of age at slaughter, $\mathbf{g}$ is the vector of additive genetic effects being assumed to follow $N\left(\mathbf{0}, \mathbf{G}\sigma_g^2\right)$ with the genetic (polygenic) variance and the G matrix represented by σ_g^2 and $\mathbf{G}$, respectively, $\mathbf{e}$ is the vector of residuals assumed to follow $N\left(\mathbf{0}, \mathbf{I}\sigma_e^2\right)$ with the residual variance and the identity matrix denoted by σ_e^2 and $\mathbf{I}$, respectively, and $\mathbf{X}$ is incidence matrix.

Using the SNP genotype data, the G matrix was constructed according to VanRaden (2008) by:

$$\mathbf{G} = \frac{(\mathbf{M} - 2\mathbf{P})(\mathbf{M} - 2\mathbf{P})'}{2\sum_{i=1}^{n} p_i(1 - p_i)}$$

where $\mathbf{M}$ is the matrix whose row elements include the number of minor alleles in each animal at each SNP locus, $\mathbf{P}$ is the matrix whose row elements contain the MAF at each SNP locus, p_i is the MAF at the *i*th SNP locus, and n is the number of SNPs used.

In this chapter, 12 different G matrices were constructed and employed by selecting from 100 to 30,000 equally spaced SNPs in number or using all available SNPs. To make the G matrices always positive definite, $10^{-4}\mathbf{I}$ was added to $\mathbf{G}$ in construction. We note that pedigree information for the animals, consequently the A matrix, was not available in this chapter.

For each of the 12 sets of SNPs, including the set of all available SNPs, the extent of LD was measured by the squared correlation (r^2) of the alleles at two loci for all pairs of two adjacent SNPs on all chromosomes (Hill and Robertson, 1968). The mean and SD of the distance (denoted as d in this chapter) between two adjacent SNPs were also calculated. In addition, correlations between the diagonal, upper triangular, and all the elements of a given G matrix, and the corresponding elements of the matrix constructed using all available SNPs, were examined (denoted as r_D, r_N and r_A, respectively).

To assess the relationships between GEBVs ($\hat{\mathbf{g}}$) in each model, correlations were computed between GEBVs incorporating the G matrix constructed using all available SNPs, and those incorporating the G matrix using a given number of SNPs. Also, linear regressions were fit, where the dependent variables were GEBVs incorporating the G matrix constructed using a given number of SNPs, and the independent variable was GEBVs obtained using all available SNPs.

Additionally, choosing the two lower-density subsets, or those of 4,000 and 10,000 SNPs, we attempted to carry out the genotype imputation with Beagle 3.3.2 from those to all the 38,502 SNPs, in which as a reference, phased haplotype data of 494 animals not having records of both the traits whose data were collected at the same two markets as the 872 animals. Then, using the imputed data, the analyses with model 1 were also conducted.

Furthermore, the distribution of MS was obviously far from a normal distribution, as shown in Figure 2-1. Then, for this trait, a threshold model (denoted as model 2 in this chapter) was also fit, as follows:

$$\boldsymbol{\eta} = \mathbf{X}\mathbf{b} + \mathbf{g} + \mathbf{e}$$

where $\boldsymbol{\eta}$ is the vector of unobserved variables in the underlying scale, assuming that $\sigma_e^2 = 1$. Two sorts of analysis were conducted regarding the outward phenotype as either a binary trait in which the observed scores 2–6 and 7–12 were each classified into one class, or an ordered categorical trait using actually observed scores.

All the parameters in model 1 were estimated via the Bayesian framework using Gibbs sampling in BLR package (de los Campos et al., 2009) under R environment (R Development Core Team, 2011). A flat prior distribution was used for the nuisance parameters ($\mathbf{b}$), and multivariate normal distributions were employed as priors for the additive genetic effects. As prior distributions for σ_g^2 and σ_e^2, independent scaled inverted chi-square distributions were used with degree of belief and scale parameters of −2 and 0, respectively, assuming that there was no prior information. The BGLR package (de los Campos, 2013), or an improved version of the BLR software, was used to estimate the parameters in model 2. A single chain of 110,000 samples was run, and the first 10,000 samples were discarded as burn-in. Posterior summaries, or mean and standard deviation here, were computed using a thinning rate of 10.

2.3 Results and Discussion

2.3.1 Extent of linkage disequilibrium

For the extent of LD, summary statistics of the squared correlation (r^2) and the distance (d) for all pairs of two adjacent SNPs in each SNP set are presented in Table 2-1. Figure 2-2 depicts the changes in means of r^2 and d, together with all values of r^2. With all available SNPs, the means of r^2 and d were 0.204 and 0.07 Mb, respectively. When the number of SNPs used was decreased to 20,000, 10,000, 8,000, 6,000 and 4,000, the

average r^2 values became 0.144, 0.096, 0.086, 0.077 and 0.066, respectively, and the corresponding means of *d* were 0.13, 0.26, 0.33, 0.44 and 0.65 Mb in turn. With all the SNPs, the means of r^2 at ranges 0–0.1, 0.1–0.2, 0.2–0.5 and 0.5–1 Mb was 0.22, 0.13, 0.10 and 0.08, respectively, and 25.7, 13.9, 10.4 and 6.4% of the r^2 values exceeded 0.3, respectively.

Table 2-1. Extent of linkage disequilibrium and distance between two adjacent SNPs and correlations for elements of G matrices

No. of SNPs	r^2		*d* (Mb)		Correlation*		
selected	Mean	SD	Mean	SD	r_D	r_N	r_A
100	0.008	0.011	25.98	3.15	0.61	0.51	0.59
200	0.017	0.027	12.86	1.95	0.74	0.64	0.72
500	0.032	0.060	5.10	1.03	0.82	0.79	0.86
1,000	0.048	0.077	2.55	0.62	0.89	0.88	0.92
2,000	0.057	0.093	1.27	0.39	0.94	0.94	0.96
4,000	0.066	0.108	0.65	0.65	0.97	0.97	0.98
6,000	0.077	0.121	0.44	1.03	0.98	0.98	0.99
8,000	0.086	0.136	0.33	0.66	0.99	0.99	0.99
10,000	0.096	0.151	0.26	0.45	0.99	0.99	0.99
20,000	0.144	0.215	0.13	0.29	0.99	0.99	0.99
30,000	0.187	0.261	0.09	0.24	0.99	0.99	0.99
38,502	0.204	0.275	0.07	0.20	-	-	-

*Correlations between the diagonal (r_D), upper triangular (r_N) and all the elements (r_A) of two G matrices constructed using a given SNP subset and all available SNPs.

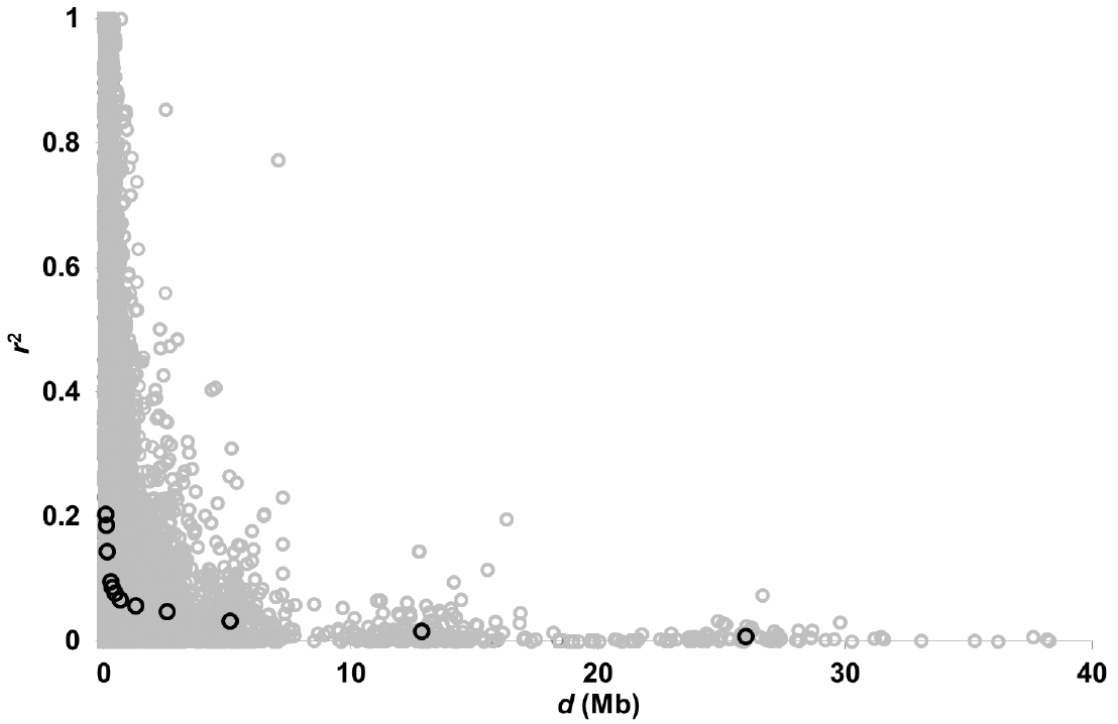

Figure 2-2. Change in mean of r^2 against mean of *d* (black circles), together with all values of r^2 (gray circles).

Investigating the overall average of r^2 using 2,670 SNPs for eight breeds, including Japanese Black cattle, McKay et al. (2007) reported that the average for 65 Japanese Black cattle was approximately 0.58 for all SNP pairs ≤ 1 kb apart, and 0.07 for all SNP pairs ≤ 2 Mb apart. In the current analysis, the mean r^2 for pairs of two adjacent SNPs, ≤ 1 kb–≤ 2 Mb apart, was 0.81–0.20, and most of the average r^2 values obtained using all the SNPs in each given distance range were higher than those reported for the eight breeds (data not shown). Furthermore, in Dutch and Australian Holstein-Friesian, Australian Angus and New Zealand Friesian and Jersey cattle, using about 3,000–7,000 SNPs, the average r^2 of 0.35 for inter-marker distances of 0–0.01 Mb declined to 0.22 for 0.02–0.04 Mb and 0.14 for 0.04–0.1 Mb (De Roos et al., 2008). As shown in Figure 2-1, a largely similar pattern of decreasing LD was observed with the current data for Japanese

Black cattle. However, it should be noted that most samples used in these previous studies were from a subpopulation, especially in representative dairy and beef breeds, or many small-scale families of the breed, including sires in some cases, which would be a factor responsible for constructed haplotype blocks in the population. In contrast to this, the samples used in this chapter were collected at two large-scale meat markets in Japan, to which fattened Japanese Black animals are sent from all over Japan. Therefore, samples used in this chapter are considered to reflect the effective size and LD extent of the national Japanese Black population.

Using the genotype data from 18,098 SNPs with MAFs greater than 10% for 25 AI sires of Brazilian Gyr dairy cattle, Silva et al. (2010) found that means of r^2 and d for two adjacent SNPs ranged from 0.24–0.17 and from 0.12–0.18 Mb, respectively, at the autosome-wide level. In this chapter of Japanese Black cattle, also with a relatively low Ne, the mean r^2 was nearly the same, but the mean d was about half that at the autosome-wide level (data not shown). Silva et al. (2010) also observed that at ranges 0–0.1, 0–0.2, 0–0.5 and 0–1 Mb, mean r^2 was 0.20, 0.18, 0.14 and 0.11, respectively, and that the proportion of SNP pairs exhibiting r^2 higher than 0.3 was 22.9, 19.7, 14.1 and 9.5% for the same ranges, respectively. In this chapter, for the same ranges, mean r^2 was 0.22, 0.21, 0.20 and 0.20, respectively, and the proportion of SNP pairs was 25.7, 24.1, 23.6 and 23.6%, respectively. From the current results, it is therefore likely that the extent of LD between more distant SNPs is relatively higher in Japanese Black cattle.

In addition, calculating the r^2 of all possible SNP pairs by chromosome, from more than 30,000 SNPs distributed genome-wide, the extent of LD and the structure of haplotype blocks were examined for 19 breeds, including Indicus, African and the composite cattle, in addition to some dairy and beef breeds (Villa-Angulo et al., 2009),

and for Angus, Charolais and crossbred beef cattle (Lu et al., 2012). Also, for Nellore cattle, whole genome LD was investigated using about 450,000 SNPs (Espigolan et al., 2013). Yan et al. (2009), using 632 maize lines genotyped for 1,229 SNP markers, demonstrated an increase in r^2 values between the markers, especially between closer SNP pairs, with an increasing MAF threshold and an increase, particularly between more distant pairs, with decreasing sample size. The MAF threshold we used was smaller than those in previous studies (Mckay et al., 2007; De Roos et al., 2008; Villa-Angulo et al., 2009; Silva et al., 2010; Lu et al., 2012; Espigolan et al., 2013), and sample size in this chapter was larger than those for most of used in these studies. In this chapter, we only calculated the r^2 of all pairs of two adjacent SNPs, avoiding a heavy computational burden. When our limited results were compared with the results of the previous studies, we found that, while the extent of whole genome LD in Zebu cattle, such as the Nellore, was relatively low, whole genome LD in the Japanese Black was likely to be higher than, or equal to, the whole genome LD in Angus, which was higher than in Charolais.

2.3.2 Change in the genomic relationship matrix

Table 2-1 also shows correlations between the diagonal (r_D), upper triangular (r_N) and all the elements (r_A) of a given G matrix, and the corresponding elements of the G matrix constructed using all available SNPs. The r_N was 0.51, 0.79, 0.88, 0.94, 0.97 and 0.99 using 100, 500, 1,000, 2,000, 4,000 and 8,000 SNPs, respectively. The changes in r_D, r_N and r_A with increasing SNP density are depicted in Figure 2-3. A correlation of 0.73 was observed between r_N and mean r^2, showing a very high linear relationship especially for SNP sets with smaller numbers. Analysing data from the 50K chip for 1,707 AI sires, along with the records of 698 steers of the Angus breed, Rolf et al. (2010) showed

that the average correlation of upper triangular elements between G matrices constructed from all available SNPs, and from its subset of SNPs selected randomly, reached nearly 0.8 using 1,000 SNPs, and exceeded 0.9 using 2,500 SNPs, suggesting that 2,500–10,000 SNPs distributed throughout the genome are required to robustly estimate a G matrix for feed efficiency traits with heritability ranging from 0.09–0.14. The changing patterns in Figure 2-3 are similar to those of Rolf et al. (2010), although we used a scheme of equally spaced selection in the number of SNPs. Compared with (Rolf et al., 2010), however, the correlation (r_N) in this chapter reached 0.9 using a lower number of SNPs and 0.99 using 8,000 SNPs, which would, at least in part, be due to a smaller Ne of the Japanese Black population.

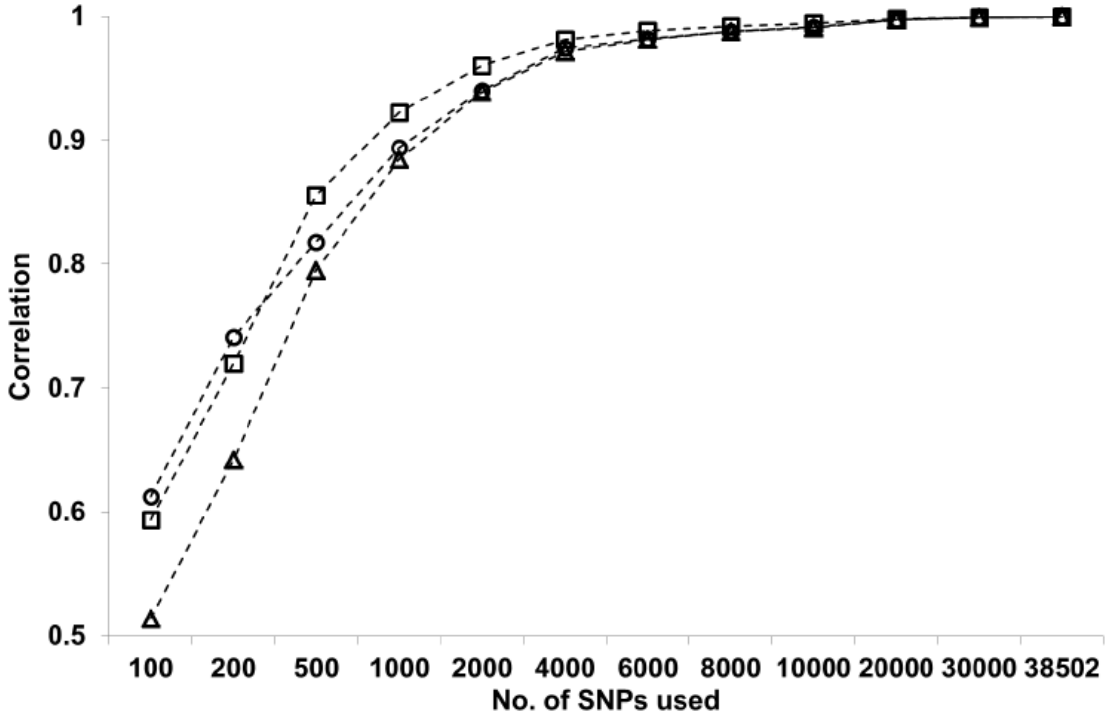

Figure 2-3. Changes in r_D, r_N and r_A with increasing density of SNPs used to construct G matrix. Circles: r_D; triangles: r_N; squares: r_A.

2.3.3 Genetic variance explained

Results of variance component estimation for CW and MS, using all available SNPs or subsets, by the conventional linear model (model 1) are presented in Tables 2-2 and 2-3, respectively. Figure 2-4 depicts the changes in proportions of estimated genetic variances for both traits. Genetic and residual variances, or σ_g^2 and σ_e^2, estimated with the G matrix using all available SNPs, were 1096.3 and 928.1 kg^2 for CW and 8.30 and 3.81 $score^2$ for MS, respectively, which resulted in heritability estimates of 0.54 and 0.68, respectively. These estimates of heritability were similar to those previously estimated in the Japanese Black population using pedigree information (Oyama, 2011), although we note that our estimate for MS might be somewhat overestimated because of the distribution of the records used. Heritability of human height was estimated to be 0.45 using 565,040 autosomal SNPs from over 10,000 unrelated individuals (Yang et al., 2011), which is lower than the estimates of 0.8–0.9 reported in previous family and twin studies (Vissher et al., 2008). Ne of humans was estimated to be 10,000 (Takahata et al., 1993), and therefore, for human polygenic traits like height, many more SNPs for a much higher LD level with causative variations would be needed to capture the total genetic variation. In contrast, Ne of cattle breeds would be much smaller, usually 100 or lower. In the case of Japanese Black cattle, where the Ne is only about 30 (Nomura et al., 2001), it is likely that a large part of the genetic variance for the carcass traits studied here could be captured by using all available SNPs within the 50K chip.

Table 2-2. Variance components estimated with model 1 for carcass weight

No. of SNPs selected	σ_e^2 (kg^2)*[1]	σ_g^2 (kg^2)*[1]	σ_p^2 (kg^2)*[1]	σ_g^2/σ_p^2
100	1798.9 (193.8) ± 90.7	289.9 (26.4) ± 78.1	2088.8 (103.2) ± 112.9	0.14 ± 0.03
200	1737.7 (187.2) ± 91.1	322.4 (29.4) ± 76.1	2060.0 (101.8) ± 107.7	0.16 ± 0.03
500	1447.0 (155.9) ± 86.0	586.8 (53.5) ± 102.5	2033.7 (100.5) ± 109.3	0.29 ± 0.04
1,000	1290.3 (139.0) ± 90.1	745.1 (68.0) ± 121.9	2035.4 (100.5) ± 111.3	0.36 ± 0.05
2,000	1168.5 (125.9) ± 96.6	844.9 (77.1) ± 135.8	2013.4 (99.5) ± 110.6	0.42 ± 0.05
4,000	1025.2 (110.5) ± 107.7	1008.6 (92.0) ± 160.7	2033.8 (100.5) ± 114.7	0.49 ± 0.06
6,000	1028.9 (110.9) ± 107.5	980.0 (89.4) ± 155.5	2009.0 (99.2) ± 111.2	0.49 ± 0.06
8,000	992.1 (106.9) ± 113.0	1032.1 (94.1) ± 165.5	2024.2 (100.0) ± 113.3	0.51 ± 0.06
10,000	956.4 (103.1) ± 112.5	1065.1 (97.2) ± 166.7	2021.5 (99.9) ± 113.5	0.53 ± 0.06
20,000	895.6 (96.5) ± 117.1	1137.0 (103.7) ± 176.7	2032.6 (100.4) ± 115.5	0.56 ± 0.07
30,000	915.8 (98.7) ± 117.2	1112.6 (101.5) ± 174.2	2028.5 (100.2) ± 114.3	0.55 ± 0.07
38,502	928.1 (100) ± 117.6	1096.3 (100) ± 173.5	2024.4 (100) ± 113.7	0.54 ± 0.07

Imp1*2	867.2	(93.4)	±	119.7	1166.6	(106.4)	±	180.3	2033.8	(100.5)	±	115.6	0.57	±	0.07
Imp2*2	931.1	(100.3)	±	118.0	1093.9	(99.8)	±	173.5	2025.0	(100.0)	±	113.7	0.54	±	0.07

*1Values in parentheses represent the percentage relative to the estimate obtained with model 1 incorporating the G matrix constructed using all available SNPs.

*2Imp1 and imp2: 38,502 SNP genotypes imputed from 4,000 and 10,000 SNPs, respectively.

Table 2-3. Variance components estimated with model 1 for marbling score

No. of SNPs selected	σ_e^2(score^2)*1				σ_g^2(score^2)*1				σ_p^2(score^2)*1				σ_g^2/σ_p^2		
100	10.68	(280.4)	±	0.55	0.93	(11.3)	±	0.38	11.62	(95.9)	±	0.60	0.08	±	0.03
200	10.81	(283.9)	±	0.57	0.72	(8.7)	±	0.35	11.54	(95.3)	±	0.58	0.06	±	0.03
500	9.27	(243.4)	±	0.55	2.35	(28.3)	±	0.55	11.63	(96.0)	±	0.61	0.20	±	0.04
1,000	8.01	(210.4)	±	0.57	3.80	(45.7)	±	0.73	11.81	(97.5)	±	0.64	0.32	±	0.05
2,000	6.64	(174.2)	±	0.57	5.18	(62.4)	±	0.82	11.82	(97.6)	±	0.65	0.44	±	0.05
4,000	5.40	(141.8)	±	0.59	6.49	(78.2)	±	0.92	11.89	(98.2)	±	0.67	0.54	±	0.06

6,000	4.85	(127.2)	±	0.60	7.07	(85.2)	±	0.97	11.92	(98.4)	±	0.68	0.59	±	0.06
8,000	4.86	(127.7)	±	0.62	7.07	(85.1)	±	0.98	11.93	(98.5)	±	0.68	0.59	±	0.06
10,000	4.36	(114.5)	±	0.63	7.65	(92.2)	±	1.03	12.01	(99.2)	±	0.69	0.63	±	0.06
20,000	3.74	(98.1)	±	0.64	8.36	(100.7)	±	1.08	12.10	(99.9)	±	0.71	0.69	±	0.06
30,000	3.77	(98.8)	±	0.66	8.37	(100.8)	±	1.10	12.13	(100.2)	±	0.71	0.69	±	0.06
38,502	3.81	(100)	±	0.66	8.30	(100)	±	1.09	12.11	(100)	±	0.71	0.69	±	0.06
Imp1[*2]	4.11	(107.9)	±	0.68	8.01	(96.6)	±	1.09	12.12	(100.1)	±	0.70	0.66	±	0.06
Imp2[*2]	3.91	(102.7)	±	0.66	8.19	(98.7)	±	1.09	12.10	(99.9)	±	0.70	0.67	±	0.06

*[1]Values in parentheses represent the percentage relative to the estimate obtained with model 1 incorporating the G matrix constructed using all available SNPs.

*[2]Imp1 and imp2: 38,502 SNP genotypes imputed from 4,000 and 10,000 SNPs, respectively.

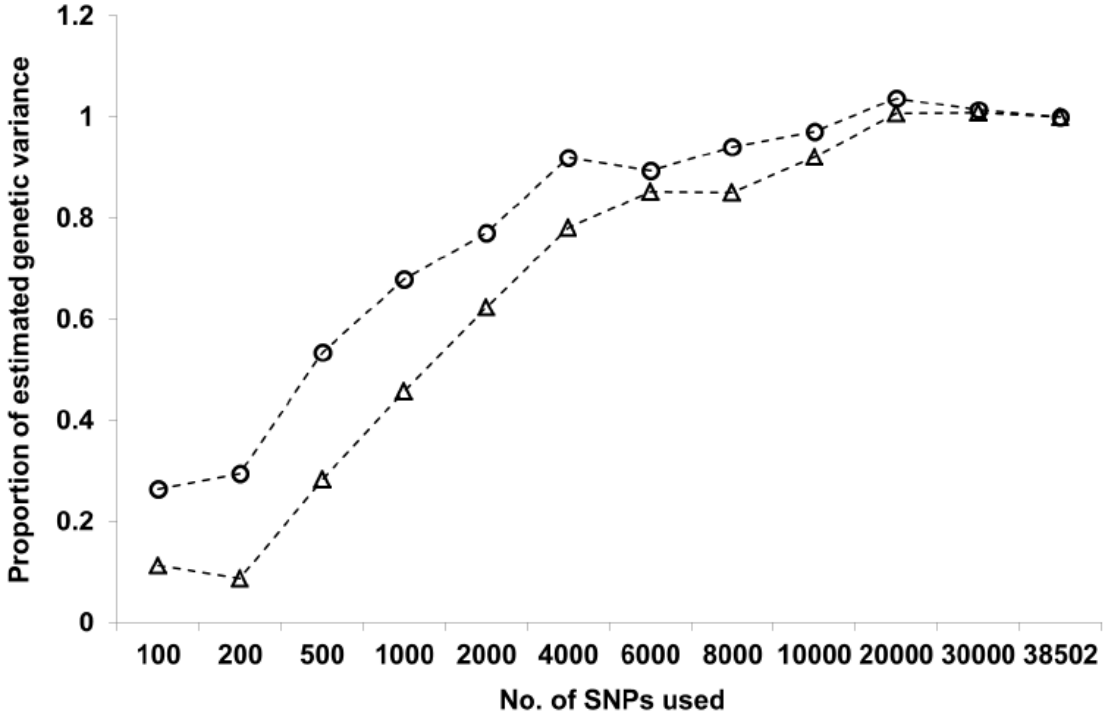

Figure 2-4. Changes in proportions of estimated genetic variances in model 1, with increasing density of SNPs used to construct G matrix. Circles: carcass weight; triangle: marbling score.

As expected, as the number of SNPs used became higher, estimated residual and genetic variances gradually decreased and increased, respectively. This is mainly because the higher the SNP marker density, the higher the LD levels between SNP markers and true QTL regions. For instance, in the case of CW, correlations between mean r^2 and the estimates of σ_e^2 and σ_g^2 in model 1 were −0.79 and 0.80, respectively. For both carcass traits, considering standard errors, a largely constant value of phenotypic variance (σ_p^2) was obtained, even with the different numbers of SNPs used. It was also observed that the value of genetic variance per SNP became larger when fewer SNPs were used (data not shown), which would be partly due to the additional variance explained by the correlated effect of SNPs around those used

to construct the G matrix. However, the proportion of genetic variance explained by SNPs decreased slightly with an increase from 4,000 to 6,000 and from 6,000 to 8,000 SNPs, for CW and MS, respectively (Figure 2-4). This could be interpreted partly as a reflection of the genetic background and architecture, or the distribution of real QTL regions and their effects relevant to each trait, in Japanese Black cattle.

For CW, approximately 90 and 97% of the genetic variance estimated with the G matrix using all available SNPs was obtained using 4,000–6,000 and 10,000 SNPs, respectively. For MS, the proportion of the genetic variance accounted for by a given number of SNPs was consistently low when compared with CW, particularly when a relatively small number of SNPs were used. This finding may indicate that the degree of marbling is controlled by only QTLs with relatively small effects, compared with the CW. In fact, three QTLs for CW, called *CW-1*, *-2* and *-3*, have been identified in GWAS, in which their allele substitution effects were relatively large (Mizoshita et al., 2004; Takasuga et al., 2007; Setoguchi et al., 2009; Nishimura et al., 2012), whereas no such QTLs have been detected for the degree of marbling until now. Using 10,000 SNPs, however, as much as 92% of genetic variance in MS was accounted for in this chapter.

2.3.4 Accuracy of genomic estimated breeding value

Correlations and linear regressions on GEBVs obtained with the different densities of SNPs used are shown in Table 2-4. When 4,000 and 10,000 SNPs were used in model 1, the correlations between the GEBVs and those obtained using all available SNPs were both 0.99 for CW and 0.98 and 0.99 for MS, respectively, with the corresponding linear regression coefficients of 0.94 and 0.98 for the former trait

and 0.82 and 0.94 for the latter trait. This showed a trend of underestimation of GEBVs with a lower number of SNPs used, particularly for the latter trait. The different levels of underestimation of GEBVs could be because of different genetic architectures of the two traits. As stated previously, while three QTLs with relatively large effects on CW in Japanese Black cattle have been found (Mizoshita et al., 2004; Takasuga et al., 2007; Setoguchi et al., 2009; Nishimura et al., 2012), no such QTLs have been found for degree of marbling. Thus, considering the results of the estimated genetic variances, the lower underestimation of GEBVs observed for CW relative to MS might reflect the observation that relatively larger effects of SNPs linked to the CW QTLs could be better captured, even with a lower number of SNPs.

2.3.5 Use of imputed genotype information

Accuracy of imputation, expressed as the percentage of correctly predicted genotypes, was 93.4 ± 2.5 and 97.4 ± 1.2% (average ± SD) for 38,502 genotypes imputed from 4,000 and 10,000 SNPs, respectively. Variance components estimated using the imputed genotype data are shown in Tables 2-2 and 2-3. Scatter plots of the GEBVs obtained for carcass weight and marbling score against those obtained using all the available SNPs without imputation are shown in Figure 2-5. Use of the imputed data resulted in a similar level of estimated variances as the level obtained using all the SNPs without imputation. Correlations between the GEBVs obtained with imputation and those obtained from all the SNPs without imputation were higher than 0.99 for both the traits. Imputation of SNP genotypes from low density to high density is now a standard procedure for using low-density marker panels in GS schemes

(Habier et al., 2009; Khatkar et al., 2012). Our results using the imputed SNP information support the use of the imputation from low-density marker panels.

Table 2-4. Correlation between and linear regression of GEBVs obtained with model 1 using a given SNP set and all available SNPs

No. of SNPs selected	Correlation coefficient		Regression coefficient	
	Carcass weight	Marbling score	Carcass weight	Marbling score
100	0.64	0.48	0.34	0.12
200	0.71	0.53	0.40	0.10
500	0.86	0.75	0.64	0.32
1,000	0.92	0.87	0.75	0.48
2,000	0.96	0.94	0.84	0.66
4,000	0.99	0.98	0.94	0.82
6,000	0.99	0.99	0.94	0.88
8,000	0.99	0.99	0.96	0.88
10,000	0.99	0.99	0.98	0.94
20,000	0.99	0.99	1.02	1.01
30,000	0.99	0.99	1.01	1.00
Imp1*	0.99	0.99	1.03	0.96
Imp2*	0.99	0.99	1.00	0.99

*Imp1 and imp2: 38,502 SNP genotypes imputed from 4,000 and 10,000 SNPs, respectively.

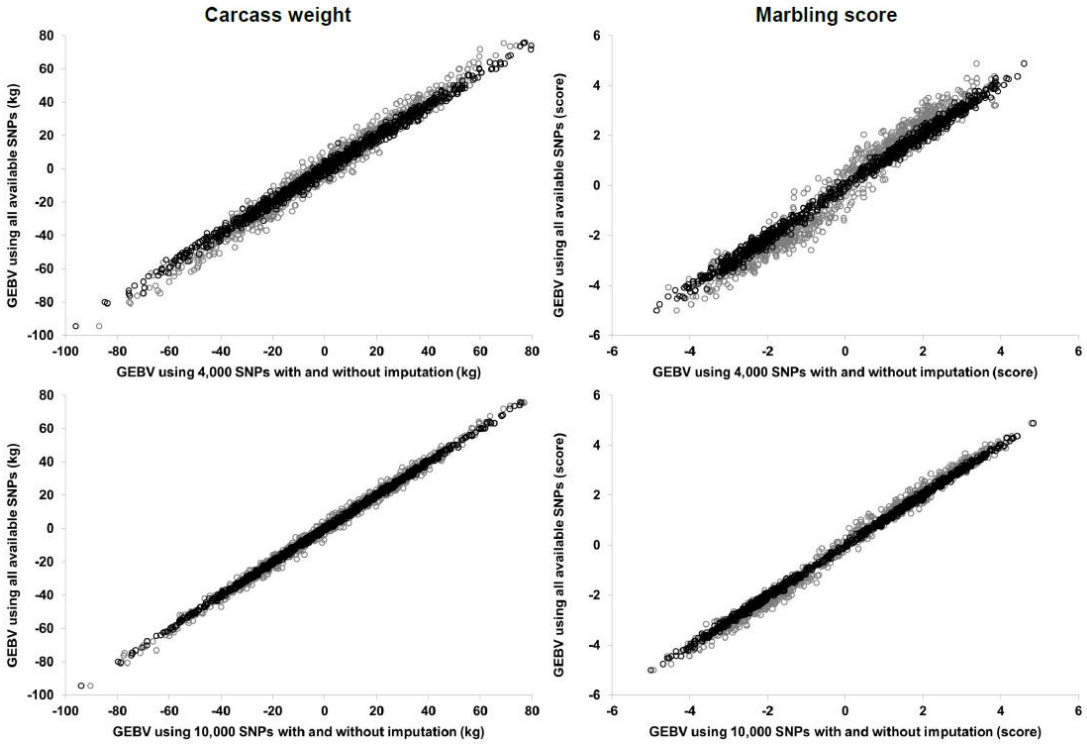

Figure 2-5. Scatter plots for GEBVs obtained with model 1 using 4,000 (top panels) or 10,000 (bottom panels) SNPs and those using all available SNPs, for carcass weight (left panels) and marbling score (right panels) with and without imputation. Black circles: with imputation; gray circles: without imputation.

2.3.6 Estimation using a threshold model

The proportions of genetic variances to phenotypic variances in the underlying scale estimated for MS using the threshold model (model 2) are presented in Table 2-5. For both the binary and more categorical treatments, the estimations were successful only when relatively small numbers of SNPs were used. The changes in the estimated proportions, relative to the proportions estimated using model 1, are depicted in Figure 2-6. The values of some correlations between GEBVs obtained with models 1 and 2 are listed in Table 2-6. It has been noted that generalized linear animal

models are plagued by extremely slow mixing in implementations of Markov chain Monte Carlo methods (Hoeschele and Tier, 1995).

Table 2-5. Proportion of genetic to phenotypic variance estimated with model 2 for marbling score

No. of SNPs selected	Binary*	Categorical*
100	0.15 ± 0.04	0.14 ± 0.03
200	0.15 ± 0.04	0.13 ± 0.03
500	0.33 ± 0.06	0.22 ± 0.04
1,000	0.51 ± 0.08	0.34 ± 0.06
2,000	0.66 ± 0.08	0.50 ± 0.06
4,000	-	0.65 ± 0.07
6,000	-	0.75 ± 0.08
8,000	-	0.70 ± 0.07
10,000	-	0.82 ± 0.07

*Binary: treated as a binary trait; Categorical: treated as a categorical trait with 11 categories.

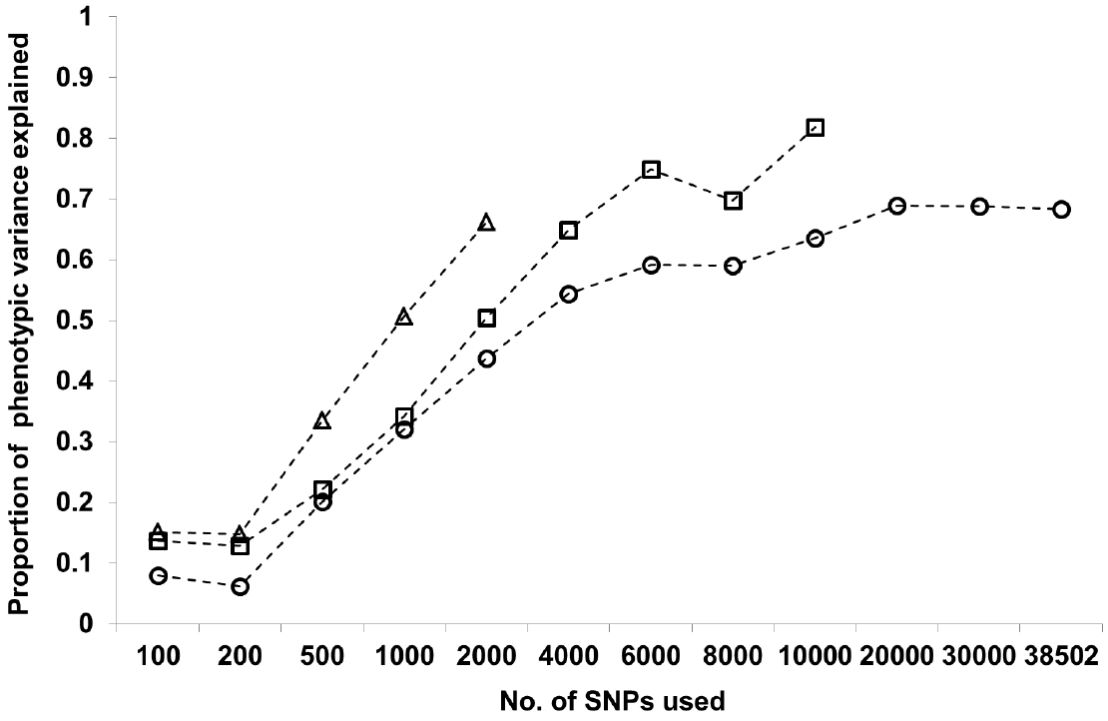

Figure 2-6. Changes in proportions of phenotypic variances explained with model 1 and 2 for marbling score, with increasing density of SNPs used to construct G matrix. Circles: model 1; triangle: model 2 (binary); square: model 2 (categorical with 11 categories).

Table 2-6. Correlations between GEBVs obtained with models 1 and 2 using a given SNP set and all available SNPs for marbling score

No. of SNPs selected	Model 1 with all 38,502 SNPs		Model 1 with each SNP set		Between binary and categorical
	Binary*	Categorical*	Binary	Categorical	
100	0.46	0.47	0.96	0.96	0.87
200	0.52	0.53	0.95	0.95	0.84
500	0.72	0.74	0.97	0.97	0.89
1,000	0.83	0.84	0.96	0.97	0.88
2,000	0.90	0.92	0.96	0.97	0.88
4,000	-	0.95	-	0.97	-
6,000	-	0.96	-	0.96	-
8,000	-	0.96	-	0.96	-
10,000	-	0.96	-	0.96	-

*Binary: treated as a binary trait; Categorical: treated as a categorical trait with 11 categories.

For both the successful and unsuccessful estimations, a single chain of 10,000,000 samples was run with the first 3,000,000 samples being discarded. The results showed that the estimates presented in Table 2-5 were not substantially different from those obtained, while there was still no convergence for any of the unsuccessful cases. The failures of the estimations using the larger numbers of SNPs may be attributed largely to the limited number of animals used in this chapter.

The proportions of the genetic to the phenotypic variances estimated with model 2 were observed to be consistently larger than the corresponding estimates with model 1. The estimates of heritability for MS of Japanese Black cattle reported in the literature (Oyama, 2011) indicate that the genetic variances in the underlying scale obtained with model 2 in this chapter may be somewhat inflated; for instance, 0.66 with 2,000 SNPs for the binary case and 0.82 with 10,000 SNPs for the case of 11 categories. However, for a given set of selected SNPs, the correlations between GEBVs obtained with model 2 and GEBVs obtained with model 1 using all the SNPs were found to be similar to the corresponding values in the analyses with model 1. The correlations between the GEBVs obtained with both the models using a given set of SNPs were very high overall. These correlations between the GEBVs would generally support the validity of the results for MS obtained with model 1.

2.3.7 Overall discussion

Equally spaced panels with various densities are already used in many situations. Such panels have the advantage of being applicable irrespective of trait and population, for which the density of SNPs plays an important role in GE and GS according to the extent of LD between SNP markers and real QTLs. At this stage, the 50K chip is most commonly used. Using higher density-panels, such as the 50K and 770K, may account for a high to very high proportion of genetic variation. In addition, as shown also in this chapter, use of genotypes imputed from low-density to high-density can take account of genetic variance largely. However, while increasing density of SNPs used could increase the extent of the LD, it could also increase the number of uninformative and collinear SNPs (Harris and Johnson, 2010). Thus, for

robust prediction it is important to exclude collinear nuisance SNPs, since their inclusion in the analyses may increase error and sampling variances in estimation of SNP effects on the training population or allow a single QTL to be attributed to a number of highly correlated SNPs, which would be likely to reduce the predictability of GEBVs and its persistence across generations. Schulz-Streeck et al. (2011) confirmed this by simulation, finding that excluding the markers with negligible or inconsistent effects by pre-selection increases the accuracy of GE.

From this perspective, even the 3K chip has been suggested to be a useful tool in dairy GE (Wiggans et al., 2011). Also, evaluating the predictive ability of subsets of SNPs, Moser et al. (2010) concluded that accurate GE of Holstein bulls and cows can be accomplished with 3,000–5,000 equally spaced SNPs. From the viewpoint of the relationship of r^2 to the accuracy of GEBV, a simulation study showed that while the accuracy of GEBVs for unphenotyped animals ranged from about 0.65, for the mean r^2 of 0.1 between adjacent markers, to more than 0.80, for the mean $r^2$of 0.2, the accuracy for phenotyped animals exceeded 0.8, with a mean r^2 of 0.1, with heritability of 0.5 (Calus et al., 2008). The mean of r^2 was almost 0.1 when 10,000 SNPs were used in this chapter (Table 2-1), and the level of heritability estimated using all available SNPs was more than 0.5 for both the traits (Tables 2-2 and 2-3). Therefore, using 10,000 equally spaced SNPs, which is relatively few compared with all available SNPs in the 50K chip, might be sufficient to cover both of the carcass traits, even in validation and application populations. Moreover, as far as genetic evaluation for ranking animals is concerned, the current results might suggest a possibility of using 4,000–6,000 equally spaced SNPs for these carcass traits in the Japanese Black population in Japan, since the downward bias in GEBV values

observed in this chapter with lower densities of SNPs would not substantially influence the ranking of animals. Such lower density panels could be used practically in pre-selection, especially of young breeding females whose number in the population is definitely high. This could be beneficial, even with the current degree of accuracy, in dramatically reducing the total cost of the genetic evaluation, since carcass traits are usually measured only on their relatives. If necessary, the imputation of SNP genotypes from the lower density panels to higher density panels, as indicated in (Boichard et al., 2012), could help to achieve an additional increase in the accuracy of GE. On the other hand, young breeding bulls to be selected as future elite AI sires should be genotyped with a high-density panel for more reliable GE and GS, since the contribution of elite AI sires to genetic improvement is significant.

There are several reports on ways of choosing unequally spaced SNPs, as well as equally spaced SNPs as a subset, particularly in a relatively low-density panel, and on the utility of low-density marker panels (e.g., Weigel et al., 2009; Vazquez et al., 2010; Zhang et al., 2011). Of these ways, choosing SNPs ranked highly in the magnitude of the absolute value of estimated effect is typical. In most cases, prediction of GEBVs with high-ranking SNPs is somewhat more accurate and reliable than with equally spaced SNPs, when the same number of SNPs is used in the prediction (e.g., Vazquez et al., 2010; Zhang et al., 2011). For Japanese Black cattle, only one previous study, conducted from the viewpoint of GE, performed the estimation of variance for carcass traits (Watanabe et al., 2014). This chapter used 50K SNP genotype data from 673 steers to simply perform linear regression analysis of each SNP for each trait, and subsets of SNPs with various significance levels for the association with each trait were used to account for variances. Including this chapter, however, use of SNPs

ranked highly based on certain criteria would generally be applicable only to a particular trait and population. One approach is to integrate the optimal subset of the SNPs for each of several important traits into one set, which is as cheap as possible to use in the target population, as ordinary selection is often implemented for certain multiple traits, although this strategy still requires the re-selection of SNPs with process of generation. While use of an equally spaced SNP panel deals with all the genome regions, according to density, a trait-specific panel would frequently deal with only parts of the genome. Thus, a compromise plan, as suggested by Weigel et al. (2009), might be practical, in which a large part of the whole SNP set is composed of equally spaced SNPs, and SNPs that are in high LD with the causative variants are also included. An example of the latter SNPs for CW in Japanese Black cattle is those linked tightly with *CW-1*, *-2* and *-3*, found by (Mizoshita et al., 2004; Takasuga et al., 2007; Setoguchi et al., 2009; Nishimura et al., 2012). In addition, since pedigree data are important information irrespective of traits (Vazquez et al., 2010), if deep and wide pedigree data can be combined with a SNP set, as mentioned above, more effective GE and GS might be possible. More studies of sophisticated approaches to construct an optimal SNP set for valid and cost-effective GE of carcass traits in beef cattle are required.

In this chapter, we employed a scheme of equally spaced selection of SNPs to investigate CW and MS, which are representative traits for carcass quantity and meat quality, respectively. We have provided important basic information on the relationships between SNP marker density and genetic variance explained and accuracy and bias of GEBVs obtained. However, as the size of the dataset available in this chapter was limited, all of the available animals were used in the estimation

analyses. In the analyses, the number of animals available (about 900) was well exceeded by the number of SNPs in most settings of SNP selection and use. Thus, we note that the genetic variance explained and the accuracy of GEBVs obtained in this chapter may be somewhat inflated, relative to those values obtained using many independent validation animals. Therefore, further research is needed to confirm the current findings, especially from the perspective of prediction, and accumulating a much larger volume of relevant data.

2.4 Summary

GE is expected to chase all the QTLs simultaneously using genome-wide SNPs, assuming a state of LD between the QTLs and the SNP markers. The accuracy of GE for carcass traits in beef cattle, including the Japanese Black, has been poorly studied. In this chapter, the extent of LD and effects of density of equally spaced genome-wide SNPs on genetic variance explained and GE for CW and MS for this breed were investigated, assuming two statistical models. It has been revealed that the extent of whole genome LD in Japanese Black cattle, whose Ne have been sharply declined and is relatively small (Nomura et al., 2011), is likely to be relatively high among major beef cattle breeds, that a large part of additive genetic variance for the carcass traits could be captured by using all available SNPs within the 50K chip, and that the degree of marbling is controlled by only QTLs with relatively small effects, compared with CW. The possibility of effective GE with at least 4,000 equally spaced SNPs was suggested for these traits in this breed.

CHAPTER THREE

Accuracy of imputation of single nucleotide polymorphism marker genotypes from low-density panels in Japanese Black cattle

3.1 Introduction

Genotype imputation has recently become an indispensable step in GP, as well as GWAS using SNPs. There are different situations that require imputation; these include the imputation of individual genotype data from low- to high-density, the prediction of missing genotype data, and the integration of genotype data from different individuals that were genotyped using different platforms with only some common SNPs (Hickey et al., 2012). For the imputation of low-density data to high-density genotype data in individuals, Habier et al. (2009) proposed the use of equally spaced low-density SNP panels. The combination of low-density SNP information and genotype imputation may be a powerful tool for cost-effective GP because it reduces genotyping costs while retaining prediction accuracy (e.g., Weigel et al., 2010b; Dassonneville et al., 2011; Mulder et al., 2012).

There are two types of information for genotype imputation, namely, the information about the relationships among individuals and that of LD structure among markers. Consequently, differences exist in imputation methods based on the use of the relationship information (family imputation) and/or LD information among markers without the knowledge of relationship (population imputation). Although some data on imputation performance were presented in Sasaki et al. (2013) and Ogawa et al. (2014), they used different statistical measures of accuracy and gave only a very limited

description of the results. In this chapter, we assess the performance of SNP genotype imputation in a Japanese Black cattle population, varying the settings of a few factors that affect imputation performance, such as SNP density, and using certain statistics as performance indicators.

3.2 Materials and Methods

3.2.1 Ethics statement

Animal care and use was according to the protocol approved by the Shirakawa Institute of Animal Genetics Animal Care and Use Committee, Nishigo, Japan (ACUCH21-1).

3.2.2 Target and reference populations

Genotype information from a total of 1,366 Japanese Black fattened steers was used in this chapter, which were identical to those used in the previous chapter. Carcass records were available for 872 and were not for 494 of these animals for the moment. Consequently, the former and the latter groups of animals were treated as the target and reference populations, respectively. The steers were sampled from 2000 through 2009 at two large meat markets in Japan, the Tokyo Metropolitan Central Wholesale Market and the Osaka Municipal South Port Wholesale Market. We note that pedigree information for the animals was not available in this chapter.

3.2.3 Genotype data

DNA samples were extracted from perirenal adipose tissue. Sample DNA was quantified and genotyped using the Illumina BovineSNP50 BeadChip (denoted as 50K

chip in this chapter). A total of 38,502 SNPs was used in this chapter, which was the same as those totally available in the previous chapter. Figure 3-1 shows the number of SNPs on each chromosome. With the exception of Bos taurus X chromosome (BTX), Bos taurus autosome (BTA) 1 and BTA28 had the largest and smallest numbers of SNPs, respectively. As 0.3% of genotype data on average was missing in the target population, missing genotype filling was conducted using a software package Beagle 3.3.2 (Browning and Browning, 2007), employing only the target population. After the missing genotype filling, the MAFs of some SNPs became lower than 0.01, but did not reach zero. In this chapter, genotype information after the imputation of missing values was treated as true genotype information.

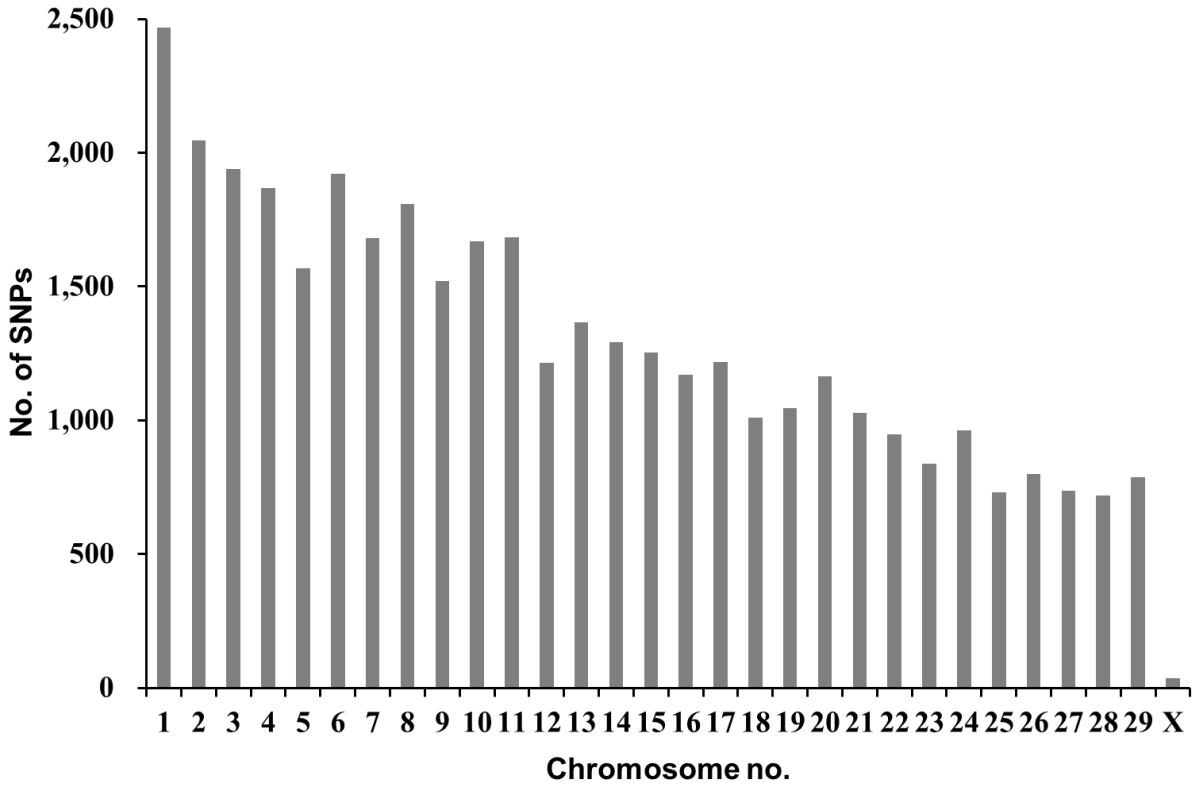

Figure 3-1. The number of SNPs on each chromosome.

3.2.4 Genotype imputation scheme

Six different low-density SNP panels were constructed by selecting 500, 1,000, 2,000, 3,000, 4,000, and 10,000 equally spaced SNPs (denoted as method 1 in this chapter), as with the case of the previous chapter. Then, with Beagle 3.3.2, we carried out the genotype imputation from these panels to all the 38,502 SNPs, exploiting phased haplotype data from the individuals in the reference population. Hozé et al. (2013) denoted that Beagle is commonly used in bovine datasets, and Nicolazzi et al. (2013) showed that, using Holstein genotypes, Beagle performed better for animals in the target population when their close relatives are not in the reference population, compared with other software using pedigree information for genotype imputation. Although Browning and Browning (2007) stated that using more than 10 iterations, which is the default in Beagle 3.3.2, yields only a very small additional improvement in imputation accuracy, we used 20 iterations in this chapter to insure maximum imputation accuracy.

To investigate the performance of imputation from a dense SNP marker panel, imputation was also implemented using SNP subsets made by omitting 500, 1,000, 2,000, 3,000, 4,000 and 10,000 equally spaced SNPs from all of the available 38,502 SNPs in the target population. Moreover, to examine the performance of imputation using a smaller reference population, 10, 25, 50 and 75% of individuals were randomly selected from the reference population. This random selection was performed five times, and only the SNPs on BTA28 were imputed.

An alternative method of SNP selection based on the level of LD between two SNPs (denoted as method 2 in this chapter) was also tested, with the expectation that this method would achieve higher imputation accuracy. Because Beagle software exploits LD structure among SNPs, using this method, the SNPs having the largest average r^2 value

(Hill and Robertson, 1968) relative to all other SNPs in the same segment, or SNP subset, were selected. Each segment was constructed by splitting all the SNPs on each chromosome into equally-sized SNP subsets, and the number of segments from each chromosome was set to be same as the number of SNPs on each chromosome selected with method 1.

3.2.5 Measures of imputation accuracy

As an indicator of genotype imputation performance, the genotype concordance rate (e.g., Badke et al., 2013; Corbin et al., 2014), which is the proportion of the genotypes from the sample population that were correctly imputed, was calculated for each SNP and each individual (denoted as C_{SNP} and C_{Ind}, respectively). The C_{SNP} was expected to be higher by chance for SNPs having a lower MAF (Hickey et al., 2012). There are multiple approaches for adjusting C_{SNP} (e.g., Vereijken et al., 2010; Hayes et al., 2011; Badke et al., 2013), some of these use genotype information from only the reference population and others employ genotype information from both the reference and target populations.

In this chapter, an adjustment to C_{SNP} using the expected concordance rate of genotypes (denoted as R_{SNP}), estimated using genotypic frequencies, was made using the formula presented by Hayes et al. (2011): $\left(C_{SNP}\text{-}R_{SNP}\right)/\left(1\text{-}R_{SNP}\right)$ with R_{SNP} given as $P_R(11)P_T(11)+P_R(12)P_T(12)+P_R(22)P_T(22)$ in this chapter, where $P_R(ij)$ are the frequencies of the genotypes for each SNP in the reference population and $P_T(ij)$ are the frequencies of the true genotypes for each SNP in the target population.

In addition, the correlation between the true and imputed genotypes (converted to 0, 1 and 2 when the genotype was 11, 12 and 22, respectively) was calculated as a measure of imputation accuracy robust to MAF (Hickey et al., 2012).

3.3 Results and Discussion

In this chapter, 12 of the available SNPs were monomorphic in the reference population, and after imputation from 500, 1,000, 2,000, 3,000, 4,000 and 10,000 SNPs, 466, 146, 67, 46, 34 and 18 SNPs, respectively, were predicted to be monomorphic in the target population. The C_{SNP} values against MAF in the target population are plotted in Figure 3-2a. In addition, C_{SNP} values of BTA6 against physical position are shown as an example in Figure 3-2b. C_{SNP} values tended to be lower for SNPs with higher MAF or those located nearer the terminals of the chromosome, especially when imputed from extremely low-density panels. This observation of the relationship between C_{SNP} and MAF is similar to the observations reported by Hickey et al. (2012) in maize, Badke et al. (2013) in pigs, and Corbin et al. (2014) in thoroughbred horse. Our finding regarding SNP location was also similar to that reported by Badke et al. (2013). Using dairy and beef cattle populations, Berry et al. (2014) also reported a decrease in genotype concordance rate with increased MAF when imputed from Illumina BovineLD to BovineHD platform. However, the decrease based on SNP location was not observed, which would be due to the increase of SNP density at the ends of chromosomes in Illunima BovineLD platform (Boichard et al., 2012).

The observations described above were less prominent when imputed from larger-sized SNP panels, however, the concordance rates of some SNPs were still not improved. Those SNPs could be classified into one of two groups; in one group the allele frequencies greatly differed between the target and reference populations, and in the other group the allele frequencies were not so different (Figure 3-2c). In a typical example of the former case, the MAF of the SNP with the lowest concordance rate of 2.4% in the target population was approximately 0.15 (Figure 3-2a), whereas the frequency of the

same allele in the reference population was 1, or monomorphic. In the latter group, the SNPs might have been mismapped, and if mismapping occurs then the estimation of the LD pattern from the genotype data would be affected. Therefore, remapping or removing of such SNPs may result in improved imputation accuracy (Erbe et al., 2012; Hozé et al., 2013).

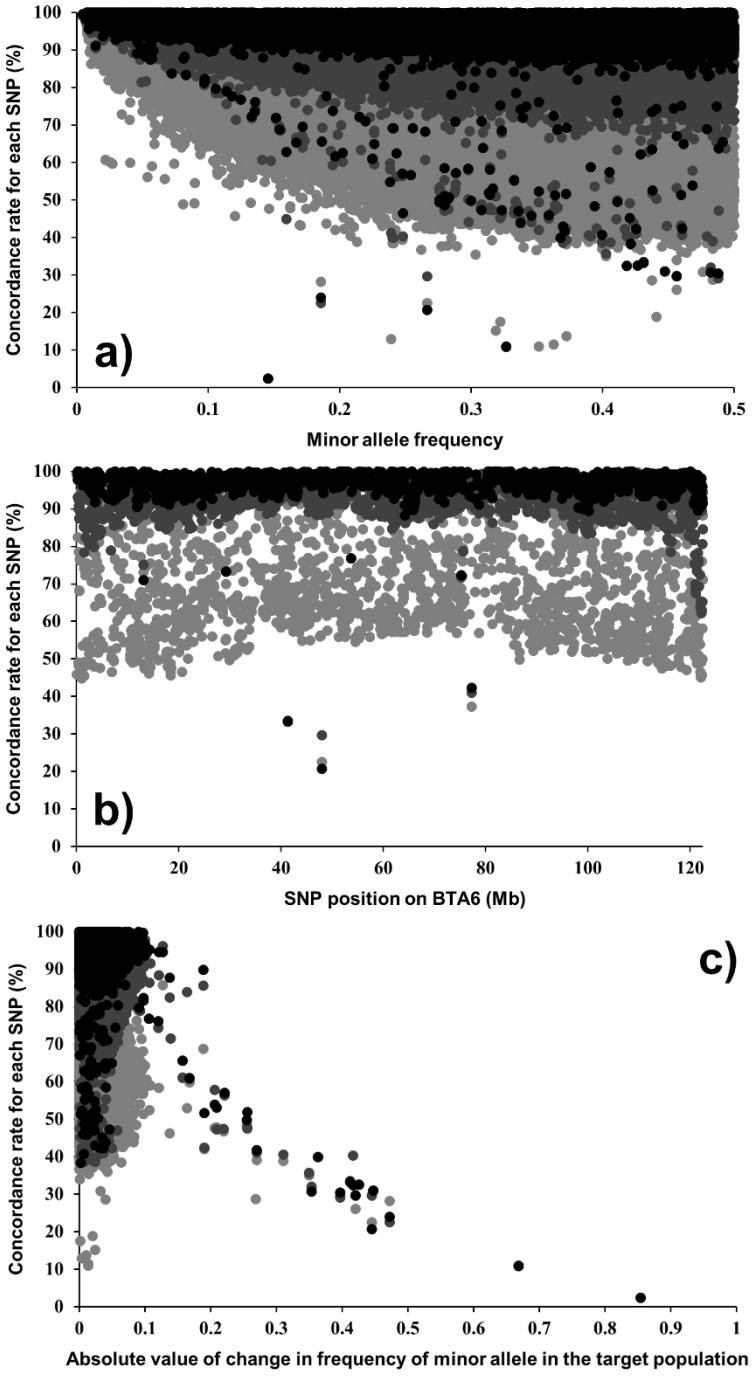

Figure 3-2. Concordance rate for each SNP (C_{SNP}) against minor allele frequency (a), physical position on BTA6 (b), and absolute value of change in the frequency of minor alleles in the target population (c). Light gray, dark gray and black dots represent the results when imputed from 500, 3,000 and 10,000 SNPs selected by method 1, respectively.

Figure 3-3 shows the change in C_{SNP} values for the 3 SNPs that had a relatively large effect on CW in Japanese Black cattle; these SNPs were first identified by Nishimura et al. (2012). In this chapter, the genotypes of the 3 SNPs were always imputed, and the C_{SNP} for SNP *Hapmap26308-BTC-057761* on BTA6 was always greater than 0.99, whereas those for SNPs *Hapmap46986-BTA-34282* on BTA14 and *BTA-52694-no-rs* on BTA8 had lower values when imputed from lower-density panels and were found to be 0.67 and 0.75, respectively, when 500 equally spaced SNPs were used. These results may be because of the distance between each of the 3 SNPs and its nearest SNP of all the 500 SNPs selected. Indeed, in the current analysis the actual distance between the SNP and its nearest SNP was about 0.4 Mb for *Hapmap26308-BTC-057761*, but more than 1.5 and 2.0 Mb for *Hapmap46986-BTA-34282* and *BTA-52694-no-rs*, respectively. For fat percentage, a dairy trait that is affected by QTLs including the *DGAT1* gene having a relatively larger effect, Chen et al. (2014) reported that the inclusion of two SNPs within the *DGAT1* gene region into lower-density SNP panels improved the accuracy of GP; these two SNPs were estimated to have the first and second largest effects when analysed using true genotypes obtained using 50K chips. Accordingly, SNPs that are highly associated with, or a true causative variant for, important traits (such as *Hapmap26308-BTC-057761*, *Hapmap46986-BTA-34282* and *BTA-52694-no-rs* for CW) should be included in low-density panels used for imputation to improve the accuracy of the genotype data used in the subsequent analyses including GP.

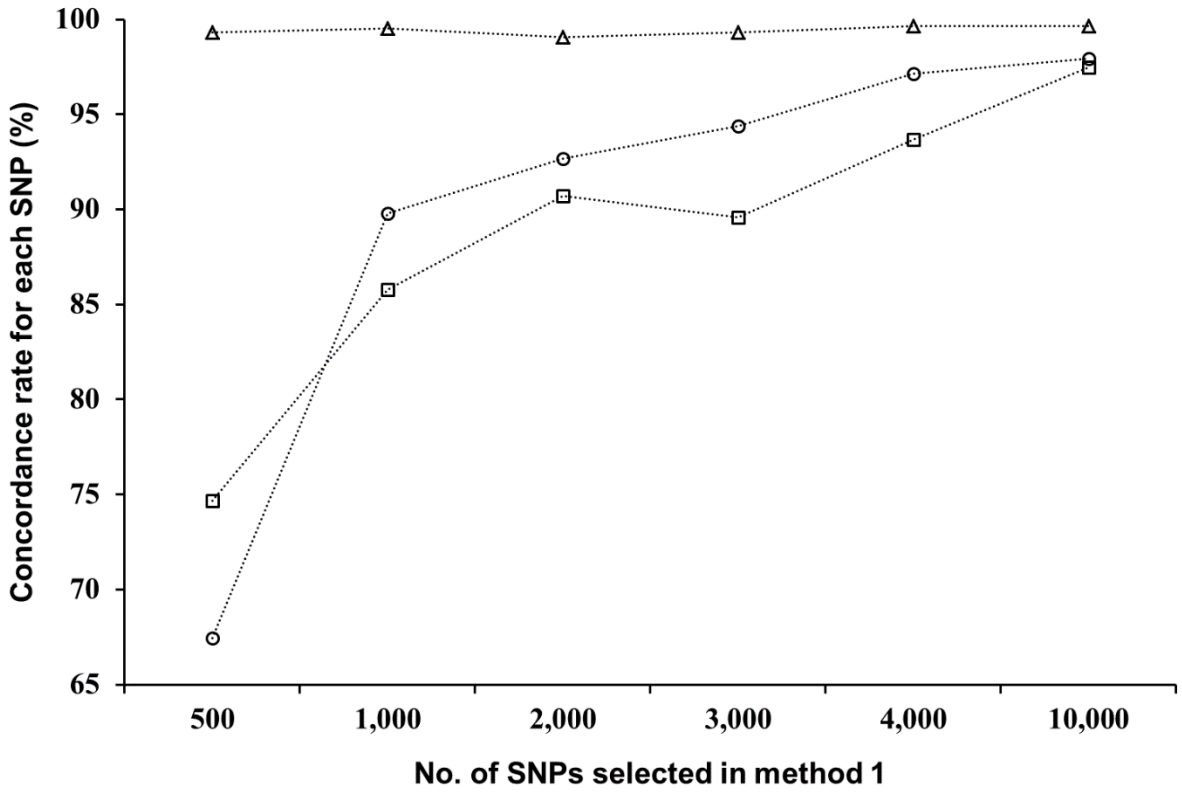

Figure 3-3. Changes in the concordance rates (C_{SNP}) for each of the 3 SNPs reported by Nishimura *et al*. (2012) against the number of equally spaced SNPs used. circle: SNP *Hapmap46986-BTA-34282*; triangle: SNP *Hapmap26308-BTC-057761*; square: SNP *BTA-52694-no-rs*.

The imputation accuracies for each SNP subset obtained using SNP selection methods 1 and 2 are shown in Table 3-1, together with the maximum and minimum values and SDs of C_{SNP} and C_{Ind}. The imputation accuracies obtained using method 1 were 69.3, 79.2, 88.0, 91.4, 93.1 and 96.5% when imputed from 500, 1,000, 2,000, 3,000, 4,000 and 10,000 equally spaced SNPs, respectively. While the SDs were always higher for C_{SNP} than C_{Ind}, the maximum values of C_{SNP} and C_{Ind} were consistently equal to and lower than 100%, respectively. Weigel et al. (2010a) reported that, with a suitable reference population, about 3,000 equally spaced SNPs achieved an imputation accuracy, defined

as the mean proportion of genotypes imputed correctly, of greater than 0.9 when imputation was implemented only on the testing set. Consequently, 3,000 SNPs provided approximately 95% of the predictive ability gained using a 50K chip. In this chapter, imputation accuracies of nearly 0.9 or greater were achieved when 2,000 and ≥3,000 equally spaced SNPs were used, respectively. This suggests that it may be possible to rank animals for carcass traits using less than 4,000 equally spaced SNPs, which is the number suggested by Ogawa et al. (2014). To confirm this, further analyses using imputed genotypes and carcass records are required, and it is also important to continue efforts to improve imputation accuracy.

Table 3-1. Imputation accuracy of low-density panels, with the maximum and minimum values (Max and Min) and standard deviation (SD) of concordance rate for each SNP (C_{SNP}) and individual (C_{Ind}).

Method	No. of SNPs selected	Imputation accuracy (%)	C_{SNP} (%)			C_{Ind} (%)		
			Max	Min	SD	Max	Min	SD
1	500	69.3	100	2.4	14.9	80.9	56.8	4.0
	1,000	79.2	100	2.4	11.5	89.0	61.1	4.4
	2,000	88.0	100	2.4	7.6	95.7	69.4	3.7
	3,000	91.4	100	2.4	5.9	97.1	74.7	3.2
	4,000	93.1	100	2.4	5.1	98.1	77.6	2.8
	10,000	96.5	100	2.4	3.6	99.2	86.7	1.6
2	500	71.0	100	2.4	15.0	85.1	59.6	4.0
	1,000	81.2	100	2.4	11.0	91.8	66.0	4.3
	2,000	89.7	100	2.4	7.1	96.1	73.7	3.4
	3,000	92.7	100	2.4	5.6	97.8	75.1	2.9
	4,000	94.0	100	2.4	4.8	98.3	79.2	2.5
	10,000	96.9	100	2.4	3.5	99.3	88.1	1.5

It would be relatively complicated to compare the current imputation accuracy results with those of previous studies, since there are many different factors affecting the imputation performance, such as breed or species, SNP density, the size of the reference population, the relatedness between reference and testing populations, the software used for imputation, the statistics used to summarize the data obtained, and so on. Corbin et al. (2014) tried to compare the results by calculating SNP density normalized by Ne. The

value of the Ne in the Japanese Black population would be about 30, or at least lower than 50 (Nomura et al., 2001). In this chapter, when using 500 equally spaced SNPs, assuming that a total genome in Japanese Black cattle is 2,820 centimorgan according to Odani et al. (2006), the normalized SNP density was calculated as 0.35 to 0.59 Ne/Morgan, which was larger than that for the 1K panel used by Corbin et al. (2014). However, the imputation accuracy in this case was found to be lower than that for the 1K panel in Corbin et al. (2014) and those used in other studies on other species when imputed from a similar SNP density normalized by Ne (Weigel et al., 2010a; Hayes et al., 2011). Assuming that this method of comparison is valid, one possible reason could be that the number of SNPs with extremely different allele frequencies between the target and reference populations may be relatively large in this chapter, as is seen in Figure 3-2c. Also, in this chapter, the individuals used were randomly-sampled fattened steers, they were not sampled based on relationships among individuals (pedigree information) and this may have resulted in poorer haplotype sharing between reference and target populations and/or a poorer construction of haplotype structure in the reference population, thus resulting in lower imputation accuracy. Hayes et al. (2011) reported the association between relatedness calculated using SNP genotype information and imputation performance, although such an association was not observed in this chapter. This may indicate the need to add close relatives of individuals in the target population into the reference population to improve imputation accuracy (e.g., Hickey et al., 2012; Ventura et al., 2014). If possible, the addition of ancestors, including the sires and dams of individuals in the target population, into the reference population (such as parent-offspring trios) would be more effective than the addition of randomly-sampled individuals because haplotypes are directly inherited from parents. Moreover, imputation

accuracy can be improved if pedigree information is directly used for genotype imputation (family imputation). Thus, the use of different imputation software may lead to further increases in imputation accuracy.

The differences between the maximum and minimum values of C_{Ind} were 24.1, 27.9, 26.2, 22.4, 20.4, and 12.5 points when using 500, 1,000, 2,000, 3,000, 4,000 and 10,000 SNPs for imputation by method 1, respectively (Table 3-1). Imputation accuracies with SNP selection method 2 were 1.7, 2.0, 1.7, 1.3, 0.9, and 0.4 points higher than the corresponding values obtained by method 1 when using 500, 1,000, 2,000, 3,000, 4,000 and 10,000 SNPs, respectively. Badke et al. (2013) also compared SNP selection methods and found that considering r^2, referred to as statistical search in the article, resulted in equal or better imputation accuracy than using equally spaced SNP subsets based on physical position. The maximum and minimum values of C_{Ind} with method 2 were always larger than the corresponding values obtained by method 1. However, not all values of C_{Ind} were increased, and furthermore, some individuals in the target population lower C_{Ind} values from method 2 than from method 1 (data not shown). The difference in C_{Ind} between the two SNP selection methods could be partly because of the amount and degree of haplotype sharing among individuals in the target and reference populations.

Figure 3-4a depicts the accuracy of imputation for each chromosome using method 1. As expected, the imputation accuracy for BTX was lower than for BTAs, mainly because of the difference in the coverage of the SNP markers on the original 50K chip. The correlations between imputation accuracy and the number of SNPs on each autosome were high, 0.75, 0.80, 0.76, 0.84, 0.75, and 0.73 when imputed from 500, 1,000, 2,000, 3,000, 4,000, and 10,000 SNPs, respectively. The length of the chromosome was proportional to the number of SNPs on that chromosome used in this chapter, and there

are some articles that reported the similar association between imputation performance and the length of chromosome (Khatkar et al., 2012; Segelke et al., 2012). Figure 3-4b represents the difference that was obtained by subtracting the value of imputation accuracy for each chromosome obtained using method 1 from the corresponding accuracy obtained using method 2. When imputed from 500 SNPs, the difference was negative and largest for BTA13 (−3.1 point) and positive and largest for BTA20 (+5.6 point). The differences that were obtained by subtracting the value of C_{SNP} for each SNP on BTA13 and BTA20 obtained using method 1 from the corresponding value of C_{SNP} obtained using method 2 are shown in Figures 3-4c and 3-4d, respectively. These results may reflect the similarity of LD structure among SNPs between the target and reference populations differing at the interchromosomal level.

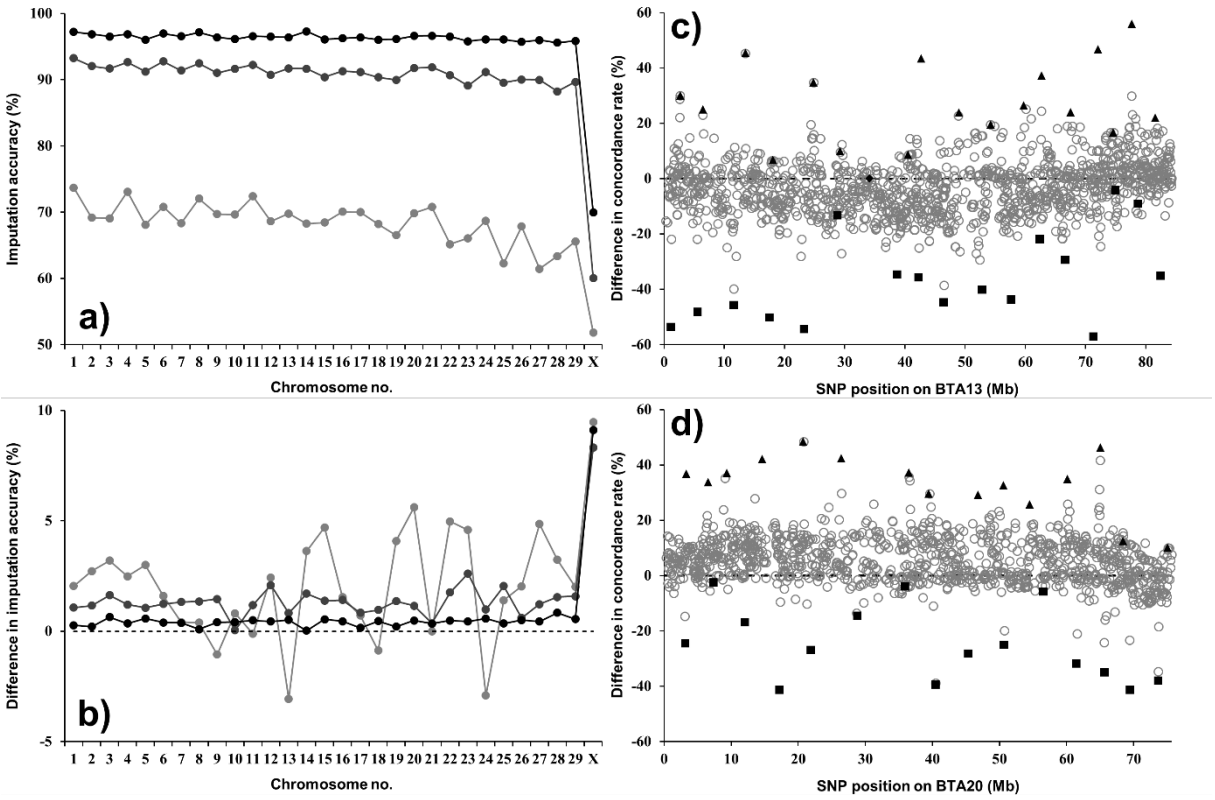

Figure 3-4. Imputation accuracy at the chromosome level using method 1 (a), the difference between methods 1 and 2 in imputation accuracy (b), the difference between methods 1 and 2 in concordance rate for each SNP (C_{SNP}) of the SNPs on BTA13 (c) and BTA20 (d). The axis of ordinate in (b), (c) and (d) represents the difference between the values obtained by method 2 and the corresponding values obtained by method 1. For (a) and (b), light gray, dark gray and black dots represent the results when imputed from 500, 3,000 and 10,000 SNPs selected by method 1, respectively. For (c) and (d), gray circle: SNP not selected by either methods 1 or 2; black triangle: SNP selected by method 1 but not method 2; black square: SNP selected by method 2 but not method 1; black rhombus: SNP selected by both methods 1 and 2.

Figure 3-5 depicts imputation accuracy relative to SNP marker density. Imputation accuracy values were 95.4, 96.4, 97.2, 98.0, 98.1 and 98.1% when 500, 1,000,

2,000, 3,000, 4,000 and 10,000 equally spaced SNPs were masked, respectively, showing the trend that imputation accuracy was improved when the number of SNP genotypes to be imputed became larger. This was an opposite trend found when 500 to 10,000 equally spaced SNPs were selected and used. This finding may be attributed to poorer sharing of shorter haplotypes than longer ones between the reference and target populations.

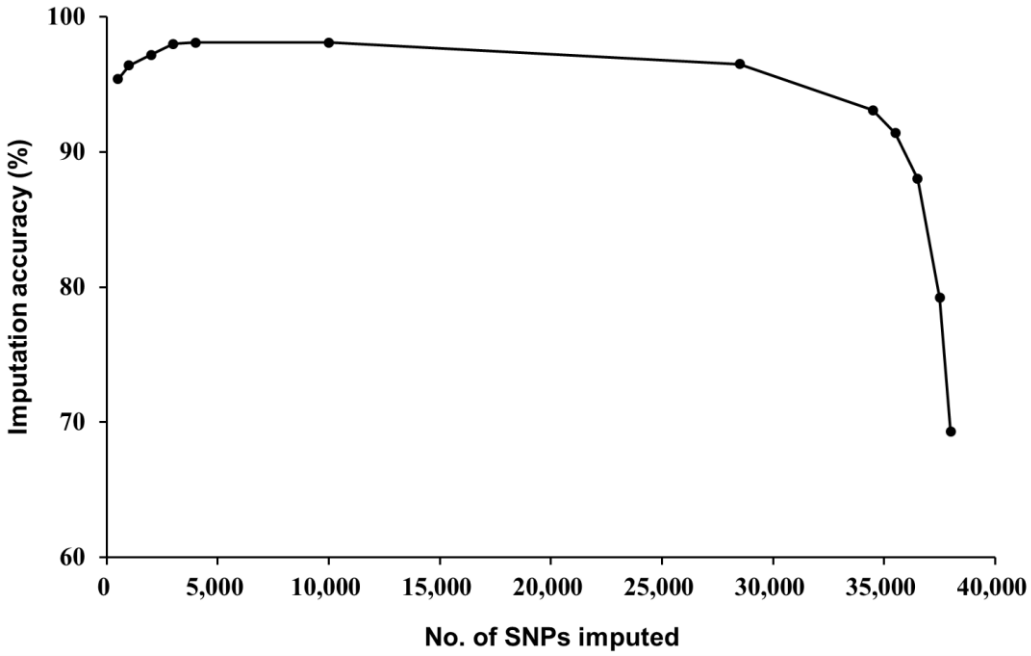

Figure 3-5. Imputation accuracy against SNP panel density.

The changes in the mean and SD values for imputation accuracy for BTA28 that occurred when the size of reference population was varied are depicted in Figure 3-6, together with the case where all of 494 individuals were in the reference population. For every density, imputation accuracy increased with enlarging the size of reference population. This pattern has also been reported in several livestock species (e.g., Zhang and Druet 2010; Hayes et al. 2011; Badke et al. 2013). However, the degree of increase

tended to be lower for lower-density SNP panels. This result indicates the possibility that the addition of individuals into the reference population may lead to improvements in imputation accuracy.

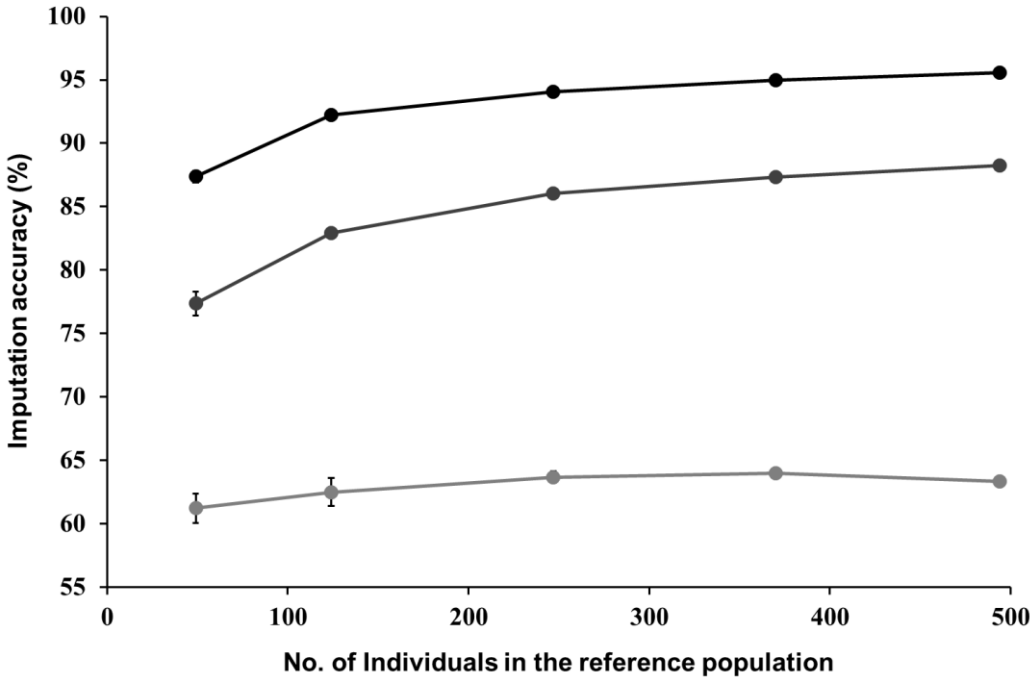

Figure 3-6. Imputation accuracy against reference population size. Values are expressed as mean ± SD for 5 replicates. Light gray, dark gray and black dots represent the results when imputed from 500, 3,000 and 10,000 SNPs selected by method 1, respectively.

Adjusted C_{SNP} values and correlations between true and imputed genotypes against MAF in the target population and location information on BTA6 are plotted in Figures 3-7a, 3-7b, 3-7c and 3-7d, respectively. In this chapter, correlation could not be obtained for SNPs predicted to be monomorphic in the target population. After adjustment, 662, 62, 19, 19, 12 and 7 imputed SNPs had concordance rate values that did not exceed

zero when genotypes were imputed from 500, 1,000, 2,000, 3,000, 4,000 and 10,000 SNPs, respectively. In contrast with Figure 3-2a, the concordance rate tended to be lower for SNPs with lower MAF, especially for those with MAF < 0.1, after adjustment (Figure 3-7a). The difference in C_{SNP} between pre- and post-adjustment has also been reported in other articles (e.g., Hickey et al., 2012; Badke et al., 2013). Hayes et al. (2011) described a similar result after adjustment, and suggested that SNPs with very low MAF should be treated with caution. On the other hand, SNPs located closer to the terminals of the chromosome had lower concordance rates than others (Figure 3-7b). A similar pattern was observed (Figures 3-7c and 3-7d).

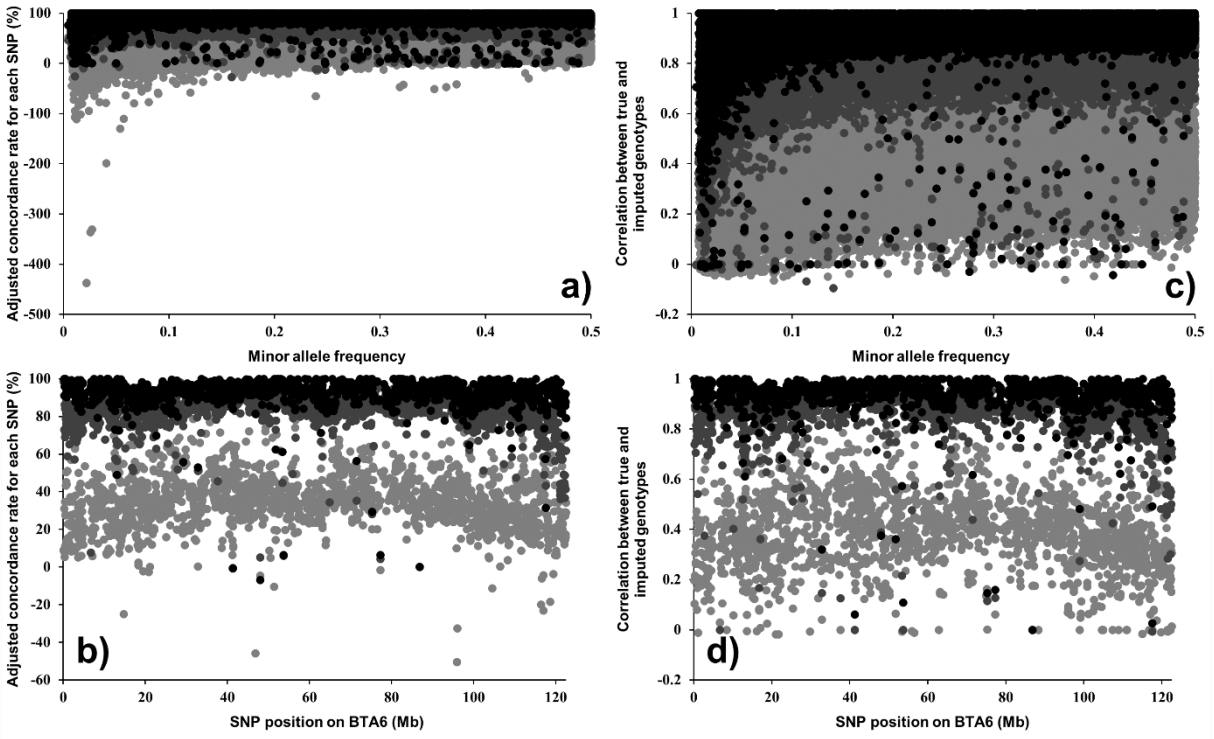

Figure 3-7. Adjusted concordance rates for each SNP and the correlation between true and imputed genotypes against minor allele frequency—(a) and (c), respectively—and against physical position on BTA6—(b) and (d), respectively. Light gray, dark gray and black dots represent the results when imputed from 500, 3,000 and 10,000 SNPs selected by method 1, respectively.

Ethnicity was also mentioned by Browning and Browning (2011) as one factor that affects the performance of imputation. On this point, there are several historical subpopulations in the Japanese Black population in Japan, and therefore, there could be differences in haplotype structure that characterize each sub-population at the genome level. Because the DNA samples used in this chapter were obtained at the 2 major carcass markets in Japan, it is likely that the target and reference populations in this chapter consisted of animals from several different subpopulations. Using Holstein cattle, Pryce

et al. (2014) reported that there were regions of the genome where differences between the imputed genotypes were more likely to arise when using reference populations from different countries. Similarly, since the Japanese Black cattle population is composed of several subpopulations, it would be possible to construct different reference populations for the same target population, and this might directly influence the results of imputation. Some studies have used a multi-breed reference population, resulting in lower imputation accuracy compared with the use of a single-breed reference population. This was despite the multi-breed reference population being larger than the single-breed reference population, and was mainly attributed to poor sharing of haplotype structure across breeds (Hayes et al., 2011; Ventura et al., 2014). From their findings, it is suggested that in Japanese Black cattle, imputation performance might improve when imputation is carried out using subpopulation-specific target and reference populations rather than when the reference population is constructed across subpopulations.

3.4 Summary

Genotype imputation with low-density SNP genotype information may be effective approach for GP in terms of reducing genotyping costs while retaining accuracy of GP, provided that the accuracy of the genotype imputation is high. However, only few studies have reported the limited results about the genotype imputation in Japanese Black cattle population. In this chapter, we investigated in more detail the genotype imputation from low-density marker panels in this breed, implementing population imputation with Beagle software and varying the settings of a few factors affecting imputation accuracy. As has been found in previous studies, the MAF and location of the SNPs affected the concordance rate for each SNP, longer autosomes had greater imputation accuracy than

shorter ones, and there was the possibility of further improving imputation accuracy by adding additional individuals to the reference population. Use of SNPs selected based on LD information slightly improved imputation accuracy compared with using equally spaced SNPs, although not all imputation accuracy was improved at the chromosome and SNP level. Imputation accuracy became greater than 90% when imputed from ≥3,000 equally spaced SNPs, implying that it could be practical and cost-effective to use a lower-density SNP chip together with genotype imputation. We discussed a possible approach that might further improve imputation accuracy in Japanese Black cattle.

CHAPTER FOUR

Estimation of variance and genomic prediction using genotypes imputed from low-density marker subsets for carcass traits in Japanese black cattle

4.1 Introduction

To boost the efficacy of implementing GP, reducing of genotyping cost is the subject to be tackled, since more individuals both measured and genotyped are needed to achieve more accurate GP. Studies have reported the performance of low-density SNP genotype data for GP using simulated and real data (Weigel et al., 2009; Moser et al., 2010; Rolf et al., 2010; Vazquez et al., 2010; Zhang et al., 2011). Ogawa et al. (2014) also investigated the possibility of using low-density SNP genotyping information for GP of CW and MS in Japanese Black cattle, giving an estimated number of equally spaced SNPs at least necessary for the valid ranking of animals genetically for the carcass traits in this breed.

Genotype imputation could be an important means to reduce genotyping costs for GP. Habier et al. (2009) proposed the use of equally spaced, low-density SNP panels to impute higher-density genotype data for individuals. This strategy could be promising for Japanese Black cattle, which are likely to have a relatively high extent of whole genome LD among beef cattle breeds (Ogawa et al., 2014). Several studies have reported the results of GP using genotype data imputed using low-density SNP panels, indicating the validity of this approach (Weigel et al., 2010; Dassonneville et al., 2011; Mulder et al., 2012). For population imputation in the Japanese Black population, Ogawa et al. (2016) examined the accuracies of genotype imputation using low-density SNP subsets,

namely equally spaced 500 to 10,000 SNP subsets, with Beagle (Browning and Browning, 2007). They assessed some measures of imputation performance including the averages of the genotype concordance rates for each SNP and each individual and the correlation between the true and imputed genotypes. However, it may be of further concern what results are brought to estimation of genetic variance and GP accuracy by the use of imputed SNP genotypes with the accuracies obtained.

Therefore, following and using the results of Ogawa et al. (2016), this chapter reports the findings from investigating the influences of the imputed SNP genotypes on G matrix, amount of genetic variance explained, and accuracies of GEBVs for CW and MS in Japanese Black cattle.

4.2 Materials and Methods

4.2.1 Ethics statement

Animal care and use was according to the protocol approved by the Shirakawa Institute of Animal Genetics Animal Care and Use Committee, Nishigo, Japan (ACUCH21-1).

4.2.2 Phenotype and genotype data

Records of cold CW and MS in 872 Japanese Black fattened steers, ranging from 15.3 to 43.0 months of age, were used for this chapter. These are the same animals used in the study of imputation accuracy by Ogawa et al. (2016) and were sampled from 2000 to 2009 at the Tokyo Metropolitan Central Wholesale and Osaka Municipal South Port Wholesale Markets. MS were expressed in terms of BMS (beef marbling standard) scores, ranging from null (1) to very abundant (12), assessed at the ribeye dissected between the

sixth and seventh ribs (Japan Meat Grading Association, 1988). The means and SDs were 496.6 and 48.0 kg for CW and 6.8 and 3.5 for MS, respectively.

DNA samples were extracted from perirenal adipose tissues and quantified and genotyped using the Illumina BovineSNP50 BeadChip (denoted as 50K chip in this chapter). While the 50K chip contains 54,001 SNPs, a total of 38,502 SNPs was used in this chapter. Details about quality control procedures and filling in missing genotype data are given in the previous chapter.

After these steps, population genotype imputation using low-density SNP subsets was conducted employing the reference data (Ogawa et al., 2016). Briefly, six different low-density SNP subsets were constructed by selecting 500, 1,000, 2,000, 3,000 4,000 and 10,000 equally spaced SNPs. Beagle 3.3.2 (Browning and Browning, 2007) was then used for genotype imputation from these subsets to all 38,502 SNPs, exploiting phased haplotype data from the individuals in the reference population of Japanese Black fattened steers. The imputation accuracy, which is expressed as the average genotype concordance rate, was 69.3%, 79.2%, 88.0%, 91.4%, 93.1%, and 96.5% for the subsets of 500, 1,000, 2,000, 3,000, 4,000, and 10,000 SNPs, respectively.

4.2.3 Statistical analyses

The following statistical model was assumed for this analysis:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Mu} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of records, $\mathbf{b}$ is the vector of the discrete effects of overall mean, carcass market, and year at slaughter as well as the continuous effects of the linear and quadratic covariates of month of age at slaughter, $\mathbf{u}$ is the vector of SNP allele substitution effects with zero mean, $\mathbf{M}$ is the matrix with element m_{ij} equal to $-2p_j$,

$1-2p_j$, or $2-2p_j$, if the genotype of individual *i* at SNP *j* is 11, 12, or 22, with the frequency of second allele of SNP *j* represented by p_j, $\mathbf{e}$ is the vector of residuals with mean zero, and $\mathbf{X}$ is an incidence matrix. Using the assumption of ridge regression-best linear unbiased prediction that each SNP has an equal variance, the variance of $\mathbf{y}$ is given as:

$$V(\mathbf{y}) = \mathbf{MM}'\sigma_u^2 + \mathbf{I}\sigma_e^2,$$

where σ_u^2 is the variance of SNP allele substitution effects, σ_e^2 is the residual variance, and $\mathbf{I}$ is the identity matrix. From this model, additive genetic values or genomic breeding values, denoted as $\mathbf{g}$, are given by $\mathbf{g} = \mathbf{Mu}$, with mean zero, and additive genetic variance captured by SNPs (σ_g^2) equals $\sigma_u^2 \sum_{j=1}^{no.of\ SNPs} 2p_j(1-p_j)$.

The above model can be rewritten as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{g} + \mathbf{e},$$

which was used in this chapter (denoted as model 1 in this chapter). The variance of $\mathbf{y}$ is expressed as:

$$V(\mathbf{y}) = \mathbf{G}\sigma_g^2 + \mathbf{I}\sigma_e^2,$$

where $\mathbf{G}$ is the G matrix constructed according to VanRaden (2008), given as:

$$\mathbf{G} = \frac{\mathbf{MM}'}{\sum_{j=1}^{no.of\ SNPs} 2p_j(1-p_j)}.$$

In this chapter, three different G matrices were constructed: one using information from the actual genotypes of all the available 38,502 SNPs, one using only SNPs in the low-density subsets, and one using the imputed genotypes for all the SNPs based on a low-density subset ($\mathbf{G_{ALL}}$, $\mathbf{G_{SUB}}$, and $\mathbf{G_{IMP}}$, respectively). To assess the similarity between the matrices, correlations between the diagonal and upper triangular elements of $\mathbf{G_{ALL}}$ and the corresponding elements of $\mathbf{G_{SUB}}$ or $\mathbf{G_{IMP}}$ were calculated

(denoted as r_D and r_N, respectively, in this chapter). Linear regressions were fit, where the dependent variables were the diagonal and the upper triangular elements of $\mathbf{G}_{\mathbf{SUB}}$ or $\mathbf{G}_{\mathbf{IMP}}$ and the independent variable was the corresponding elements of $\mathbf{G}_{\mathbf{ALL}}$ (denoted as b_D and b_N, respectively, in this chapter). To make the G matrices always positive definite, $\mathbf{I}\times 10^{-4}$ was added to all the matrices. Pedigree information, and consequently the additive relationship matrix, was not available for the animals in this chapter.

Analyses with model 1 were conducted using each of the three G matrices ($\mathbf{G}_{\mathbf{ALL}}$, $\mathbf{G}_{\mathbf{SUB}}$, and $\mathbf{G}_{\mathbf{IMP}}$). All parameters in the model were estimated via the Bayesian framework using Gibbs sampling in the BLR package (de los Campos et al., 2009) in R environment (R Core Team, 2013). A flat prior distribution was used for the nuisance parameters ($\mathbf{b}$), and multivariate normal distributions were employed as priors for the additive genetic values and residuals. As prior distributions for σ_g^2 and σ_e^2, independent scaled inverted chi-square distributions were used with degree of belief and scale parameters of −2 and 0, respectively, assuming no prior information. A single chain of 110,000 samples was run, and the first 10,000 samples were discarded as burn-in. Posterior summaries, or means and SDs, were computed with a thinning rate of 10. The proportion of additive genetic to phenotypic variances explained by all the SNPs considered in model 1 was estimated by averaging the values of $\hat{\sigma}_g^2/\left(\hat{\sigma}_g^2+\hat{\sigma}_e^2\right)$ calculated using samples obtained. Correlation was calculated to assess the relationships between GEBVs ($\hat{\mathbf{g}}$) obtained using $\mathbf{G}_{\mathbf{ALL}}$ and those obtained using $\mathbf{G}_{\mathbf{IMP}}$. Linear regressions were also computed, where the dependent variables were GEBVs using $\mathbf{G}_{\mathbf{SUB}}$ and $\mathbf{G}_{\mathbf{IMP}}$ and the independent variable was those using $\mathbf{G}_{\mathbf{ALL}}$.

The data were also analysed using the linear model (model 2 in this chapter):

$$\mathbf{y}=\mathbf{Xb}+\mathbf{g}_1+\mathbf{M}_2\mathbf{u}_2+\mathbf{e},$$

where $\mathbf{g_1}$ is the vector of the additive genetic values with mean zero due to any one of the six different low-density SNP subsets including n_t (t=1～6) SNPs, $\mathbf{M_2}$ is the matrix with element m_{ik} equal to $-2p_k$, $1-2p_k$, or $2-2p_k$, if the genotype of individual i at the kth of the remaining $38{,}502-n_t$ SNPs is 11, 12, or 22, with the frequency of the second allele represented by p_k, and $\mathbf{u_2}$ is the vector of allele substitution effects of the remaining SNPs with mean zero. The variance of $\mathbf{y}$ is:

$$V(\mathbf{y}) = \mathbf{G_1}\sigma^2_{g_1} + \mathbf{M_2}\mathbf{M}'_2\sigma^2_{u_2} + \mathbf{I}\sigma^2_e,$$

where $\sigma^2_{g_1}$ is the additive genetic variance captured by the selected n_t SNPs, $\sigma^2_{u_2}$ is the variance of allele substitution effects of the remaining SNPs, and, in this chapter, $\mathbf{G_1}$ was equal to $\mathbf{G_{SUB}}$. Model 2 can therefore be rewritten as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{g_1} + \mathbf{g_2} + \mathbf{e},$$

and the variance of $\mathbf{y}$ as:

$$V(\mathbf{y}) = \mathbf{G_1}\sigma^2_{g_1} + \mathbf{G_2}\sigma^2_{g_2} + \mathbf{I}\sigma^2_e,$$

where $\mathbf{g_2}$ equals $\mathbf{M_2u_2}$, $\mathbf{G_2}$ is the G matrix for the genotypes of the $38{,}502-n_t$ SNPs, and $\sigma^2_{g_2}$ is the additive genetic variance equal to $\sigma^2_{u_2}\sum_{k=1}^{38{,}502-n_t} 2p_k(1-p_k)$.

Imputed SNP genotypes were not directly obtained on the given animal, but were rather constructed using other animals' genotype data. Model 2 was just used to hopefully partition the total additive genetic variance explained by all available SNPs into two parts, distinguishing the part due to imputed genotypes from that due to true ones, and to hopefully investigate the influences of genotype imputation in some details. With this model, two types of analyses were conducted; for all analyses with model 2, $\mathbf{G_1}$ was always $\mathbf{G_{SUB}}$, and for analyses "without genotype imputation" and "with genotype imputation", $\mathbf{G_2}$ was based on the actual or imputed genotypes of $38{,}502-n_t$ SNPs, respectively.

Similar to the case of model 1, assuming similar prior distributions, the parameters in model 2 were estimated using the BLR package. However, the number of iterations, burn-in period, and thinning rate were increased to 600,000, 100,000, and 50, respectively, since, in some cases, the results were judged to have poor mixing and slow convergence by visual inspection of sample plots, when using the same settings as those for model 1. $\sigma^2_{g_1}+\sigma^2_{g_2}$ and $\sigma^2_{g_1}\Big/\left(\sigma^2_{g_1}+\sigma^2_{g_2}\right)$ were estimated by averaging the values of $\hat{\sigma}^2_{g_1}+\hat{\sigma}^2_{u_2}\sum_{k=1}^{38,502-n_t}2p_k\left(1-p_k\right)$ and $\hat{\sigma}^2_{g_1}\Big/\left(\hat{\sigma}^2_{g_1}+\hat{\sigma}^2_{u_2}\sum_{k=1}^{38,502-n_t}2p_k\left(1-p_k\right)\right)$, respectively, calculated using the samples obtained. Correlations were computed between GEBVs ($\hat{\mathbf{g}}$) obtained using $\mathbf{G_{ALL}}$ with model 1 and those ($\hat{\mathbf{g}}_\mathbf{1}$, $\hat{\mathbf{g}}_\mathbf{2}$ or $\hat{\mathbf{g}}_\mathbf{1}+\hat{\mathbf{g}}_\mathbf{2}$) obtained using model 2. Linear regressions were fit, where the dependent variables were $\hat{\mathbf{g}}_\mathbf{1}$, $\hat{\mathbf{g}}_\mathbf{2}$ and $\hat{\mathbf{g}}_\mathbf{1}+\hat{\mathbf{g}}_\mathbf{2}$ obtained using model 2, and the independent variable was $\hat{\mathbf{g}}$ obtained using $\mathbf{G_{ALL}}$ with model 1.

4.3 Results and Discussion

4.3.1 Comparison of G matrices

Table 4-1 shows correlation and linear regression coefficients for the comparison of matrices $\mathbf{G_{SUB}}$ and $\mathbf{G_{IMP}}$ and matrix $\mathbf{G_{ALL}}$ for the diagonal elements (r_D and b_D, respectively) and those for the upper triangular elements (r_N and b_N, respectively).

Table 4-1. Correlation (r_D and r_N) and single regression coefficients (b_D and b_N) for elements of matrices $\mathbf{G_{SUB}}$, $\mathbf{G_{IMP}}$ and $\mathbf{G_{ALL}}$ [1)].

No. of SNPs	$\mathbf{G_{SUB}}$				$\mathbf{G_{IMP}}$			
selected	r_D	r_N	b_D	b_N	r_D	r_N	b_D	b_N
500	0.82	0.79	0.99	1.00	0.83	0.77	0.59	0.63
1,000	0.89	0.88	0.99	1.00	0.95	0.93	0.80	0.80
2,000	0.94	0.94	0.96	1.00	0.98	0.98	0.92	0.93
3,000	0.96	0.96	1.01	0.99	0.99	0.99	0.96	0.96
4,000	0.97	0.97	0.99	1.00	0.99	0.99	0.97	0.98
10,000	0.99	0.99	0.99	1.00	0.99	0.99	0.98	1.00

[1)] r_D and r_N: correlation coefficient between the diagonal elements and between the upper triangular elements, respectively, of $\mathbf{G_{SUB}}$ and $\mathbf{G_{ALL}}$ and of $\mathbf{G_{IMP}}$ and $\mathbf{G_{ALL}}$; b_D and b_N: linear regression coefficient, where the dependent variable was the diagonal elements and the upper triangular elements, respectively, of $\mathbf{G_{SUB}}$ and $\mathbf{G_{IMP}}$ and the independent variable was the corresponding elements of $\mathbf{G_{ALL}}$.

The values of r_D and r_N were approximately 0.8 for the 500 SNP subset and above 0.9 for subsets with more than 1,000 selected SNPs. Both r_D and r_N for $\mathbf{G_{IMP}}$ were higher than the corresponding values for $\mathbf{G_{SUB}}$ when the number of selected SNPs was 1,000 to 4,000. Values of r_D and r_N for $\mathbf{G_{IMP}}$ imputed from 2,000 and 3,000 SNPs exceeded those for $\mathbf{G_{SUB}}$ using 4,000 SNPs. The values of b_D and b_N were nearly one for $\mathbf{G_{SUB}}$, whereas those for $\mathbf{G_{IMP}}$ were approximately 0.6, 0.8, and 0.9 for 500, 1,000, and 2,000 SNP subsets, respectively. This is unexpected and needs to be confirmed in further studies. The differences in regression coefficients between $\mathbf{G_{SUB}}$

and $\mathbf{G}_{\mathbf{IMP}}$ are likely to be dependent on the genotype imputation results as well as the imputation accuracy (Ogawa et al., 2016), and both b_D and b_N obtained using $\mathbf{G}_{\mathbf{IMP}}$ were highly correlated, with correlation coefficients of more than 0.9.

4.3.2 Estimation of variance components

Figure 4-1 depicts the marginal posterior means and SDs of variance components in model 1, which were determined using $\mathbf{G}_{\mathbf{ALL}}$ and $\mathbf{G}_{\mathbf{SUB}}$ without genotype imputation and using $\mathbf{G}_{\mathbf{IMP}}$ based on genotype imputation, for CW and MS. Variance components in model 2 obtained using $\mathbf{G}_{\mathbf{1}}$ and $\mathbf{G}_{\mathbf{2}}$ are also shown in Figure 4-1. Those of variance components estimated were σ_e^2 and σ_g^2 in model 1 and σ_e^2, $\sigma_{g_1}^2$, and $\sigma_{g_2}^2$ in model 2. For imputation in model 2, estimation of variance was performed using $\mathbf{G}_{\mathbf{SUB}}$ as the $\mathbf{G}_{\mathbf{1}}$ matrix; the $\mathbf{G}_{\mathbf{2}}$ matrix was constructed using the imputed, rather than the actual, genotypes of the corresponding SNPs.

The estimated phenotypic variance (marginal posterior mean ± SD) in model 1 using $\mathbf{G}_{\mathbf{ALL}}$ based on the actual genotypes of all available 38,502 SNPs was 2,024.4 ± 113.7 kg^2 for CW and 12.1 ± 0.7 $point^2$ for MS. However, when using model 1 without genotype imputation, it was clearly observed for both traits that as the number of the SNPs used to construct $\mathbf{G}_{\mathbf{SUB}}$ decreased, the residual and the genetic variances were over- and underestimated, respectively, compared with using $\mathbf{G}_{\mathbf{ALL}}$. Use of genotype imputation employing the low-density SNP subsets moderated this to some degree.

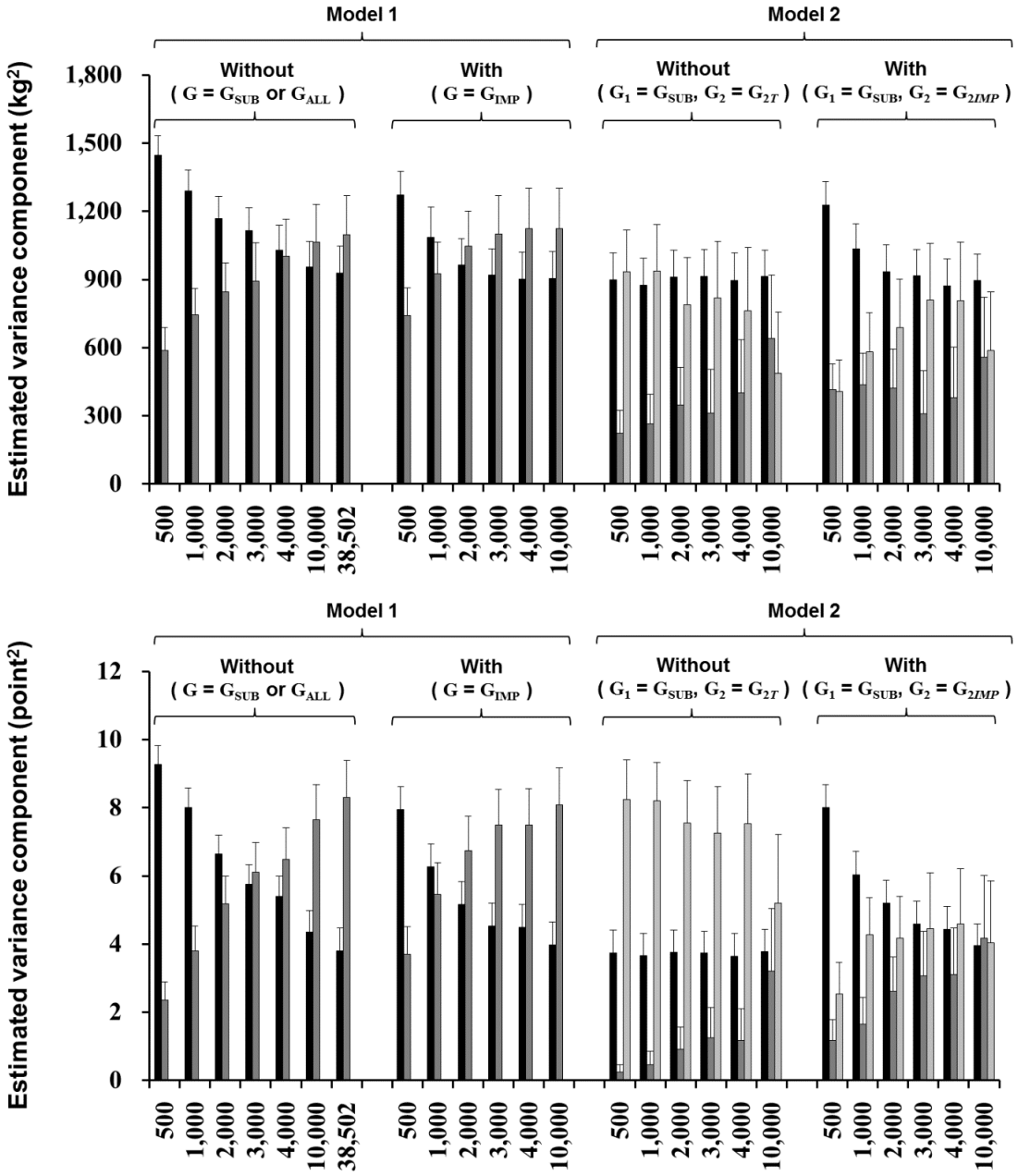

Figure 4-1. The marginal posterior means and standard deviations of variance components for carcass weight (above) and marbling score (below). For model 1, black and dark-gray bars represent the estimates of residual variance (σ_e^2) and additive genetic variances (σ_g^2), respectively. For model 2, black, dark-gray and light-gray bars represent the estimates of residual variance (σ_e^2) and two additive genetic variances ($\sigma_{g_1}^2$ and $\sigma_{g_2}^2$),

respectively. With and without refer to the use of genome imputation. In the cases with and without imputation using model 2, estimation was conducted using the matrix $\mathbf{G}_{\mathbf{SUB}}$ as $\mathbf{G}_1$ and the matrix $\mathbf{G}_2$ was constructed using the imputed genotypes and the true genotypes of the corresponding SNPs ($\mathbf{G}_2 = \mathbf{G}_{2IMP}$ and $\mathbf{G}_2 = \mathbf{G}_{2T}$, respectively).

Proportions of the additive genetic to phenotypic variance estimated using model 1 without and with genotype imputation are shown in Table 4-2, together with those obtained via $\mathbf{G}_{\mathbf{ALL}}$ based on the actual genotypes of all available SNPs. When $\mathbf{G}_{\mathbf{ALL}}$ was used, the estimate of the proportion of the additive genetic variance, or heritability, was 0.54 for CW and 0.68 for MS. Conversely, when $\mathbf{G}_{\mathbf{SUB}}$, based on the actual genotypes of only 500 SNPs, was used without imputation, the estimated value of the proportion of the additive genetic variance was only about 50% and 30% of those values for CW and MS, respectively. When using $\mathbf{G}_{\mathbf{IMP}}$ based on genotypes imputed from 500 SNPs, however, the proportions estimated were approximately 15% better for both traits. When using the 500 to 3,000 SNP subsets, while keeping the estimated value of phenotypic variance almost the same (data not shown), the proportion of the genetic variance estimated via genotype imputation was approximately 0.1 and more than 0.1 higher for CW and MS, respectively, relative to the corresponding estimates without genotype imputation. For both traits, when using $\mathbf{G}_{\mathbf{IMP}}$ based on the genotypes imputed from the 2,000 to 3,000 SNP subsets, the estimated proportions of the genetic variance were comparable to those using only the actual genotypes of the 4,000 to 10,000 SNP subsets.

Table 4-2. The proportion of additive genetic to phenotypic variance estimated with model 1 using $\mathbf{G}_{SUB}$, $\mathbf{G}_{ALL}$ and $\mathbf{G}_{IMP}$ for carcass weight (CW) and marbling score (MS).

Trait	No. of SNPs selected	$\mathbf{G}_{SUB}$ or $\mathbf{G}_{ALL}$		$\mathbf{G}_{IMP}$	
CW	500	0.29 (53.3)[1)]	± 0.04	0.37 (68.0)	± 0.05
	1,000	0.36 (67.6)	± 0.05	0.46 (85.0)	± 0.06
	2,000	0.42 (77.5)	± 0.05	0.52 (96.3)	± 0.07
	3,000	0.44 (82.1)	± 0.05	0.54 (100.6)	± 0.06
	4,000	0.49 (91.1)	± 0.06	0.55 (102.5)	± 0.07
	10,000	0.53 (97.3)	± 0.06	0.55 (102.3)	± 0.07
	38,502	0.54 (100)	± 0.07	-	
MS	500	0.20 (29.5)	± 0.04	0.32 (46.3)	± 0.06
	1,000	0.32 (46.8)	± 0.05	0.46 (68.0)	± 0.06
	2,000	0.44 (64.0)	± 0.05	0.56 (82.6)	± 0.06
	3,000	0.51 (75.1)	± 0.06	0.62 (90.9)	± 0.06
	4,000	0.54 (79.6)	± 0.06	0.62 (91.3)	± 0.06
	10,000	0.63 (92.9)	± 0.06	0.67 (97.8)	± 0.06
	38,502	0.68 (100)	± 0.06	-	

[1)] Values in parentheses represent the percentage relative to the estimate obtained with model 1 using $\mathbf{G}_{ALL}$ constructed using the actual genotypes of all of the available 38,502 SNPs.

For model 2, when the actual genotypes of SNPs were used without genotype imputation, there was no substantial overestimation of the residual variance σ_e^2 for each of the two traits in any of the limited SNP subsets, compared with using $\mathbf{G_{ALL}}$ in model 1. For both traits, the estimate of the sum of the genetic variances $\sigma_{g_1}^2 + \sigma_{g_2}^2$ was almost the same as that obtained with model 1 using $\mathbf{G_{ALL}}$ (data not shown).

For model 2 without genotype imputation, as expected, the estimate of $\sigma_{g_1}^2$ became higher when a larger number of SNPs were considered to construct $\mathbf{G_1}$, or $\mathbf{G_{SUB}}$, for both traits. The estimate of $\sigma_{g_1}^2 \big/ \left(\sigma_{g_1}^2 + \sigma_{g_2}^2\right)$ was always smaller for MS than for CW (data not shown). This was probably due to the difference in the genetic architecture between the traits. When the same $\mathbf{G_{SUB}}$ matrix was used as matrix $\mathbf{G}$ in model 1 and matrix $\mathbf{G_1}$ in model 2, the estimate of the additive genetic variance σ_g^2 obtained from model 1 was higher than that for $\sigma_{g_1}^2$ obtained via model 2. This was clearly observed, especially when fewer SNPs were used in $\mathbf{G_{SUB}}$. The higher estimate from model 1, given the same matrix $\mathbf{G_{SUB}}$, could be caused by the additional variance due to the correlated effects of SNPs around those used to construct the $\mathbf{G_{SUB}}$ matrix that was accounted for by this model according to the extent of LD.

When model 2 with genotype imputation was used, the values of $\sigma_{g_1}^2 + \sigma_{g_2}^2$ and σ_e^2 were almost the same as σ_g^2 and σ_e^2 obtained with model 1 using $\mathbf{G_{IMP}}$, respectively. However, taking into consideration the SDs of the estimates, using the same $\mathbf{G_{SUB}}$ matrix based on subsets with lower numbers of SNPs (500 to 1,000 SNPs for CW and 500 to 4,000 SNPs for MS), there was a tendency for $\sigma_{g_1}^2$ to be larger and $\sigma_{g_2}^2$ to be smaller when estimated using model 2 with genotype imputation than without, although the estimated $\sigma_{g_1}^2$ was always smaller than the σ_g^2 estimated with model 1. Under the use of $\mathbf{G_{SUB}}$ based on the low numbers of SNPs, it may be likely that model

2 with $\mathbf{G}_2$ based on the imputed remaining SNPs cannot capture $\sigma^2_{g_2}$ so well, then considering the term $\mathbf{g}_1$ and using $\mathbf{G}_{\mathbf{SUB}}$ based on the true genotypes of the SNPs results in the captured genetic variance partially including the true variance explained by SNPs around those used to construct $\mathbf{G}_{\mathbf{SUB}}$.

4.3.3 Accuracy of genomic estimated breeding values

For model 1, the changes in correlation coefficients between GEBVs obtained using $\mathbf{G}_{\mathbf{ALL}}$ and GEBVs using $\mathbf{G}_{\mathbf{SUB}}$ and $\mathbf{G}_{\mathbf{IMP}}$ are shown in Figure 4-2. In addition, changes in linear regression coefficients, where the dependent variables are GEBVs obtained using $\mathbf{G}_{\mathbf{SUB}}$ and $\mathbf{G}_{\mathbf{IMP}}$ and the independent variable is GEBVs using $\mathbf{G}_{\mathbf{ALL}}$, are also depicted in Figure 4-2. Correlation coefficients were obviously higher when using $\mathbf{G}_{\mathbf{IMP}}$ than when using $\mathbf{G}_{\mathbf{SUB}}$ for both traits. The coefficients were observed to be almost one when using genotype data imputed from at least 2,000 to 3,000 equally spaced SNPs. Also, for both traits, regression coefficients were nearer to one when using $\mathbf{G}_{\mathbf{IMP}}$ than when using $\mathbf{G}_{\mathbf{SUB}}$. When genotypes were imputed from the 3,000 SNP subset, the coefficient of regression of GEBVs obtained using $\mathbf{G}_{\mathbf{IMP}}$ on those using $\mathbf{G}_{\mathbf{ALL}}$ was approximately one for CW and higher than 0.9 for MS. These results indicate the validity of using suitable genotype imputation based on information obtained from equally spaced, low-density SNPs.

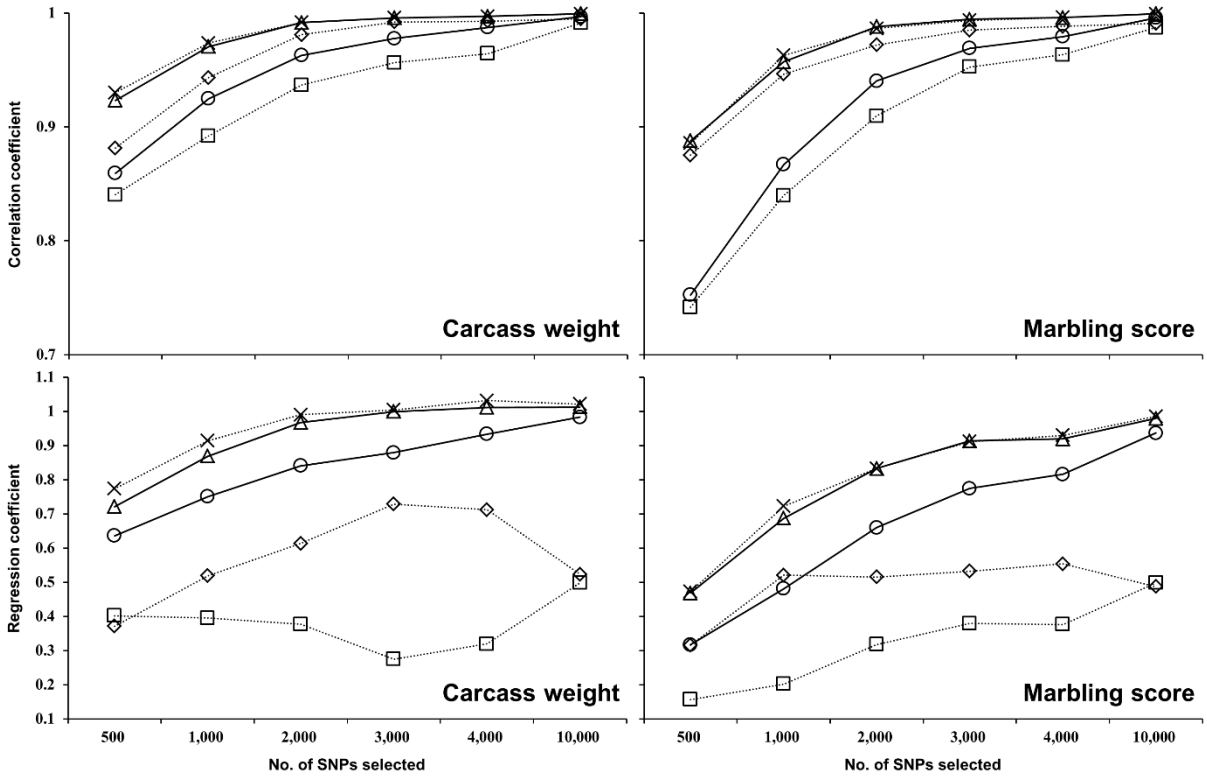

Figure 4-2. Changes in correlation coefficients (above) between, and linear regression coefficients (below) for, GEBVs for carcass weight and marbling score. Circles and triangles with solid lines represent the correlation coefficients between GEBVs obtained using $\mathbf{G}_{\mathbf{ALL}}$ based on the actual genotypes of all available SNPs and GEBVs using $\mathbf{G}_{\mathbf{SUB}}$ based on the actual genotypes of only selected SNPs as well as GEBVs using $\mathbf{G}_{\mathbf{IMP}}$ based on the imputed genotypes in model 1, respectively. Crosses, squares and rhombuses with dotted lines represent the coefficients between GEBVs obtained with model 1 using the actual genotypes of all the available SNPs and GEBVs, $\hat{\mathbf{g}}_1 + \hat{\mathbf{g}}_2$, $\hat{\mathbf{g}}_1$ and $\hat{\mathbf{g}}_2$ from model 2, which were obtained with $\mathbf{G}_2$ constructed using the imputed genotypes of the corresponding SNPs. For linear regression, circles and triangles with solid lines represent the coefficients where the independent variable was GEBVs obtained using the actual genotypes of all the available SNPs and the dependent variables were GEBVs obtained using the actual genotypes of only selected SNPs and using the imputed

genotypes in model 1, respectively, while crosses, squares and rhombuses with dotted lines represent the coefficients where the independent variable was GEBVs obtained with model 1 using the actual genotypes of all available SNPs and the dependent variables were GEBVs $\hat{\mathbf{g}}_1 + \hat{\mathbf{g}}_2$, $\hat{\mathbf{g}}_1$ and $\hat{\mathbf{g}}_2$ in model 2, respectively.

When model 2 was used without imputed genotypes, correlations were in the range of 0.997 to 0.999 for CW and above 0.999 for MS between $\hat{\mathbf{g}}_1 + \hat{\mathbf{g}}_2$ and $\hat{\mathbf{g}}$ obtained with model 1 using $\mathbf{G}_{\mathbf{ALL}}$. The linear regression coefficient, where the dependent variable was $\hat{\mathbf{g}}_1 + \hat{\mathbf{g}}_2$, and the independent variable was $\hat{\mathbf{g}}$, ranged from 1.011 to 1.034 for CW and from 1.005 to 1.022 for MS It was observed that model 2 without use of imputed genotypes worked well.

When using model 2 with genotype imputation, $\mathbf{G}_1$ was $\mathbf{G}_{\mathbf{SUB}}$ and $\mathbf{G}_2$ was constructed using the imputed genotypes of the corresponding SNPs. For both traits, correlation coefficients between $\hat{\mathbf{g}}_1 + \hat{\mathbf{g}}_2$ and $\hat{\mathbf{g}}$ obtained with model 1 using $\mathbf{G}_{\mathbf{ALL}}$ and the linear regression coefficients were similar to the corresponding values for GEBVs obtained using $\mathbf{G}_{\mathbf{IMP}}$ and $\mathbf{G}_{\mathbf{ALL}}$ (Figure 4-2).

4.3.4 Overall discussion

It may be a feasible approach for a pre-selection of young bulls and females to implement GP with genotypes imputed using a low-density SNP panel (Dassonneville et al., 2011). For genotype imputation in Japanese Black cattle, Ogawa et al. (2016) assessed certain measures of imputation performance in some details. Following their work, this chapter investigated what results were brought by the use of their results of imputation, particularly on the estimation of variance components and the accuracy of breeding value

prediction. Our underlying concern was the degree of the utility to combine the use of low-density SNP panel and genotype imputation in carcass traits in Japanese Black cattle.

This chapter has revealed certain profiles of influences of imputed SNP genotypes using low to lower-density SNP subsets on genetic variance estimation and GP for the carcass traits. The results obtained using model 1 were improved with genotype imputation to higher-density SNP information, relative to the results using only equally spaced low-density SNP subsets (Table 4-2 and Figure 4-2). This finding would indicate certain availability of the imputed genotype information in genetic evaluation for the carcass traits in Japanese Black cattle. However, it should be mentioned that results obtained using imputed genotype information are heavily dependent on genotype imputation accuracy. Obviously, comparing the genetic variances estimated using imputed SNP genotypes showed that the whole variance was not fully caught accordingly, when there were a relatively large number of imputed SNPs used and their accuracies were relatively low.

In the current study, genotypes were imputed using the Beagle software, which exploits LD information among markers (population imputation) and does not use the knowledge of genetic relationships among individuals. In usual cases of Wagyu cattle, however, the knowledge of pedigree information and the relationships among individuals are available for genotype imputation. Then, other software implementing family imputation, such as FImpute (Sargolzaei et al., 2014), is likely to give more accurate imputed genotypes.

A model like model 2 was used to treat with separating all available SNPs into two groups such as SNPs on a particular chromosome and those on the remaining chromosomes independent of the original chromosome (e.g., Jensen et al., 2012). In the

setting of the current study, equally spaced SNPs used for constructing $\mathbf{G}_{SUB}$ and the remaining SNPs were distributed across all chromosomes. Nevertheless, we also used model 2 as one operational model. For the genotype imputation from low-density SNP sets to high density data, the genotype data on individuals consist of the genotypes obtained from themselves and those imputed completely using a different population as the reference population. There might be some room to investigate whether actual and imputed genotypes should be explicitly regarded as the different types of SNP data from each other for GP. At least conceptually, the model like model 2 would take SNPs with actual and imputed genotypes into consideration simultaneously, but separately. Our findings indicate that model 2 may be likely to work to partition the total additive genetic variance. However, there was no substantial improvement of GEBV accuracy observed using this model (Figures 4-1 and 4-2). For model 2, the actual performance in various settings of SNP distribution needs to be further investigated carefully, for instance, using a simulation study.

In this chapter, equally spaced SNP subsets were used as low-density panels. This type of SNP selection is independent of the trait examined. From the viewpoint of GP, several studies (e.g., Weigel et al., 2009; Vazquez et al., 2010; Zhang et al., 2011) reported accuracies using trait-specific SNP subsets which were constructed based on the magnitude of the estimated allele substitution effect. Such SNP subsets generally resulted in equal to higher accuracies than those obtained using equally spaced subsets, especially when the number of selected SNPs was low. However, these trait-specific SNP subsets often failed to cover the entire genome even when the number of selected SNPs was increased. This fact may lead poor results in genotype imputation using trait-specific SNP subsets. On this point, Moser et al. (2010) merged the subsets of equally spaced, selected

SNPs and trait-specific SNP subsets. This treatment could improve the limited imputation accuracy in the case of using only trait-specific SNP subsets by enlarging the genome coverage of only trait-specific SNP subsets. Merged trait-specific SNP subsets often include SNPs that are highly associated with traits of concern but may not be included in equally spaced, low-density SNP subsets. Therefore, genotype imputation using panels merged in such a way would give improved GP accuracy relative to those obtained with genotype imputation using only equally spaced low-density SNP panels.

4.4 Summary

Accurate genotype imputation with low-densiy SNP data may enable to perform accurate GP as same as one using real SNP genotype information of higher density while reducing genotyping cost. It has been not obvious the effects of using genotype information imputed from low-density SNP subsets in the previous chapter on, especially, variance component estimation and GP for the carcass traits. In this chapter, we examined the influence of genotype imputation with low-density marker subsets on the G matrix, amount of genetic variance explained, and accuracies of GP for the two carcass traits in Japanese Black cattle, using the two linear models. The results obtained indicate the validity of using genotype information accurately imputed from lower-density SNP information in GP for carcass traits, showing the possibility of cost-effective GP for in Japanese Black cattle.

CHAPTER FIVE

Genomic prediction for carcass traits in Japanese Black cattle using single nucleotide polymorphism markers of different densities

5.1 Introduction

GP, or breeding value prediction for SNP markers, can be conducted even when pedigree information is not available, once prediction equation is appropriately provided from the data of individuals with phenotype and marker genotype information. Even in countries where registration systems have been well developed, small-scale rural natural breeding operations of native animals may therefore easily and effectively use information on genomic relationships and inbreeding in genetic evaluation and pre-selection while also helping maintain genetic diversity. The same thing is of course true for countries where pedigree information is not well and widely available.

Several factors have been reported to affect the results of GP (e.g. Goddard and Hayes, 2009), including Ne and the degree of LD. While the use of fully high-density SNP markers is desirable to capture the total additive genetic variance of a quantitative trait and to predict genomic breeding values accurately, GP using SNPs of a lower-density would be an attractive alternative if found to work relatively well compared to GP using higher-density SNP markers. It is likely that the Ne of Japanese Black cattle is small (Nomura et al., 2001), resulting in the degree of whole-genome LD being higher than in other major beef breeds (Ogawa et al., 2014). This may make it possible to implement a valid GP utilising lower-density SNP information and effectively capture the total additive genetic variance by using only lower-density SNP genotypes rather than high-

density ones. In addition, GP via GBLUP uses a G matrix based on a sufficient density of SNP genotype information in place of an additive relationship matrix based on pedigree information. There are different procedures used to construct G matrices, which may also affect the results of estimating variance components and GP.

Ogawa et al. (2014) recently outlined the possibility of using GP to rank Japanese Black cattle genetically for carcass traits; however, the analyses of GP using different animals as the validation population were not performed in the study. The purpose of this chapter was therefore to use validation as well as training populations to conduct GP analyses for two economically important carcass traits in Japanese Black cattle, using several SNP marker subsets of varying densities and two representative G matrices.

5.2 Materials and Methods

5.2.1 Trait data and training and validation populations

The phenotypic records of 1,791 Japanese Black fattened steers ranging from 15.3 to 32.5 months of age were used. These were collected from 2000 to 2010 at two domestic carcass markets, the Tokyo Metropolitan Central Wholesale market and the Osaka Municipal South Port Wholesale market. These animals, whose pedigrees were unavailable, were treated as the training population.

The carcass traits studied were CW and MS, which were assessed by official graders according to carcass grading standards (Japan Meat Grading Association, 1988). MS is an ordered categorical trait (1–12) that is evaluated between the 6th and 7th ribs.

An additional 189 animals were used as the validation population. These were from one prefecture and had predicted breeding values determined by the official genetic evaluation in this prefecture, employing pedigree data, using an average information

REML and the empirical BLUP procedure (Ashida and Iwaisaki, 1998). The linear model used in the official genetic evaluation included the fixed effects of overall mean, year, sex, carcass market, slaughter age (linear and quadratic covariates) and inbreeding of animal (linear covariate) and the random effects of farm, animal and residual. The base population was assumed to be that in 1975. Estimated heritability (standard error) was 0.43 (0.02) for CW and 0.55 (0.02) for MS. The accuracies of the predicted animal effects, or predicted breeding values, for these 189 animals were higher than 0.77, with average (SD) of 0.80 (0.01) for CW and of 0.83 (0.01) for MS. In the present chapter, the number of sires of 1,050 of 1,791 steers in training population were 65, who also had 164 steer progenies in validation population.

5.2.2 Genotype data

Genomic DNA was extracted from the peri-renal adipose tissues of the steers according to a standard protocol; proteinase K digestion, phenol-chloroform extraction, and ethanol precipitation. Sample DNA was quantified by the DU640 spectrophotometer (Bekman Coulter Inc.) or the Nanodrop ND-1000 Spectrophotometer (Thermo Fisher Scientific Inc.). Of the training population, 458 animals were genotyped using the Illumina BovineHD BeadChip (Illumina Inc.), while the remaining animals were genotyped using either of the Illumina BovineSNP50 BeadChip (Illumina Inc.) and the Affymetrix AXIOM BOS1 Array (Affymetrix Inc.). The genotype data were analysed using GenomeStudio software (Illumina Inc.). Phasing and genotype imputation were performed using the Beagle 3.3.2 package (Browning and Browning, 2007). There were 565,837 autosomal SNPs available with a MAF higher than 0.01 and a p-value from

Hardy–Weinberg equilibrium higher than 0.001 in the training population. These SNPs were denoted as the high-density set (HD set) in the chapter.

Two additional SNP sets containing 31,231 and 6,073 SNPs (denoted as the 50K and the low-density (LD) set, respectively) were also constructed from the HD set; these included SNPs genotyped using the Illumina BovineSNP50 and the BovineLD BeadChip respectively. Furthermore, in total, 27 subsets of SNPs of low- to lower-density were constructed using the HD set and used, with every ($i \times j$)-th SNP being taken and kept in order, where i = 2–10 and j = 1, 10 and 100.

5.2.3 Statistical analyses

GEBVs were obtained by GBLUP, with the following linear model being fitted:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of records; $\mathbf{b}$ is the vector of overall mean and macro-environmental effects, including year of slaughter, carcass market and age at slaughter (linear and quadratic covariates); $\mathbf{g}$ is the vector of genomic breeding values and is assumed to be $V(\mathbf{g}) = \mathbf{G}\sigma_g^2$, with G matrix and additive genetic variance denoted as $\mathbf{G}$ and σ_g^2, respectively; $\mathbf{e}$ is the vector of residuals and is assumed to be $V(\mathbf{e}) = \mathbf{I}\sigma_e^2$, with the identity matrix and residual variance represented by $\mathbf{I}$ and σ_e^2, respectively; $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices for $\mathbf{b}$ and $\mathbf{g}$, respectively.

In this chapter, two different G matrices constructed using the same SNPs were used. One was the matrix described by VanRaden (2008) and the other was the modification of the matrix proposed by Yang et al. (2010) (denoted as $\mathbf{G_V}$ and $\mathbf{G_Y}$, respectively). The modification of the matrix of Yang et al. (2010) was suggested by Meuwissen et al. (2011) who stated that the estimation of the diagonal elements of the G

matrix in Yang et al. (2010) may lead to the resulting matrix no longer being positive semi-definite. Two G matrices were in the following form:

$$\mathbf{G} = \frac{\mathbf{MDM}'}{n},$$

where $\mathbf{M}$ is the matrix whose element m_{ij} corresponds to the standardised genotype at SNP *j* of animal *i* and is calculated as $\left(w_{ij} - 2p_j\right) \Big/ \sqrt{2p_j\left(1-p_j\right)}$, using the number of counted alleles at SNP *j* of animal *i* represented by w_{ij} and the frequency of the counted alleles at SNP *j* represented by p_j; $\mathbf{D}$ is the diagonal matrix; n is the number of SNPs used for constructing the G matrix. For $\mathbf{G_V}$ and $\mathbf{G_Y}$, the *j*-th diagonal element of $\mathbf{D}$ was set to $2np_j\left(1-p_j\right) \Big/ \sum_{k=1}^{n} 2p_k\left(1-p_k\right)$ and 1, respectively. Consequently, each SNP was weighted by the value $p_j\left(1-p_j\right) \Big/ \bar{p}\left(1-\bar{p}\right)$ in $\mathbf{G_V}$, compared with $\mathbf{G_Y}$, where $\bar{p} = 0.5 - \sqrt{0.25 - \sum_{k=1}^{n} p_k\left(1-p_k\right)/n}$, giving larger and smaller weighting to SNPs with higher and lower MAFs than $\bar{p}$. The elements of $\mathbf{G_V}$ and $\mathbf{G_Y}$ were a weighted average and an unweighted average, respectively, of the relationships at the SNPs (Meuwissen et al., 2011).

The R package 'rrBLUP' (Endelman, 2011) was used to conduct GBLUP analyses, in which variance components were estimated by the REML procedure by using the spectral decomposition algorithm described by Kang et al. (2008). The heritability of each trait was estimated as $\hat{\sigma}_g^2 \Big/ \left(\hat{\sigma}_g^2 + \hat{\sigma}_e^2\right)$. According to Legarra et al. (2008), the prediction accuracy of GP was assessed as the correlation between the GEBVs and the records of animals in the validation population corrected by all the fixed effects estimated in the official genetic evaluation, divided by the square root of estimated heritability. To investigate how $\mathbf{G_Y}$ and $\mathbf{G_V}$ constructed using the same SNP sets were different from each other, the correlation and single regression coefficients were calculated for the

elements of $\mathbf{G_Y}$ and those of $\mathbf{G_V}$. In the regression, dependent and independent variables were the elements of $\mathbf{G_V}$ and $\mathbf{G_Y}$, respectively. The diagonal and upper-triangular elements were analysed separately.

5.3 Results

Table 5-1 shows estimates of the components of variance and heritability obtained using HD, 50K and LD SNP sets and incorporating the two types of G matrix. For both traits, the estimates of the additive genetic and residual variances obtained using the 50K set, which included about 30,000 SNPs, were substantially higher and lower, respectively, than those using the LD set, which contained about 6,000 SNPs. The estimates of the genetic and residual variances obtained using the HD set were, in turn, higher and lower, respectively, than those estimated using the 50K set. However, the estimates differed less between the HD and 50K sets than between the 50K set and the LD set. Fortunately, the estimates of the phenotypic variance of both traits ($\hat{\sigma}_g^2 + \hat{\sigma}_e^2$) were almost the same when using the 50K and LD sets as when using the HD set, regardless of the type of G matrix used. These results indicated that the use of SNP sets of higher-density contributed to capture an additional additive genetic variance that was not explained by lower-density markers, irrespective of which of two methods was adopted to construct the G matrix. The heritability estimated using the LD set was about 0.4 for CW and slightly lower than 0.5 for MS. Contrary to these findings, the estimates using the 50K and HD sets were higher for both CW (0.06–0.07 and 0.08–0.1, respectively) and MS (0.06–0.07 and 0.08–0.09, respectively). The heritability estimate obtained using the HD set was thus found to be around 0.5 for CW and slightly lower than 0.6 for MS. These estimates compared approximately to the reported average values from about 40 estimates

obtained from the official traditional genetic evaluation using pedigree data conducted in each prefecture (0.46 for CW and 0.58 for MS; Wagyu Registry Association, 2014).

The genetic and residual variances estimated using $\mathbf{G}_{\mathbf{Y}}$ were found to be consistently higher and lower, respectively, than those estimated using $\mathbf{G}_{\mathbf{V}}$, irrespective of which SNP set was used. However, the actual differences were very small. When using the HD set, the estimates of heritability for CW and MS obtained using $\mathbf{G}_{\mathbf{Y}}$ were only 0.03 and 0.02, respectively, higher than those obtained using $\mathbf{G}_{\mathbf{V}}$.

The correlations between the GEBVs and the corrected records in the validation population divided by the square root of estimated heritability, or the accuracy of GP, are also given in Table 5-1. The correlations obtained using the three SNP sets were about 0.85 for CW and 0.60 for MS. The prediction accuracy for either trait was not substantially affected by the SNP set or G matrix used.

Table 5-1. The estimates of components of variance and heritability and prediction accuracy obtained using HD, 50K and LD SNP sets and incorporating two types of G matrix [A]

Trait	SNP set	No. of SNPs	$\mathbf{G_V}$				$\mathbf{G_Y}$			
			$\hat{\sigma}_e^2$	$\hat{\sigma}_g^2$	$\hat{h}^2$	r	$\hat{\sigma}_e^2$	$\hat{\sigma}_g^2$	$\hat{h}^2$	r
Carcass	HD	565,837	1748.6	1630.5	0.48	0.85	1651.1	1741.2	0.51	0.84
weight	50K	31,231	1822.4	1534.7	0.46	0.86	1754.7	1612.9	0.48	0.86
	LD	6,073	2020.0	1327.0	0.40	0.85	1982.6	1366.4	0.41	0.84
Marbling	HD	565,837	4.3	5.5	0.56	0.58	4.1	5.7	0.58	0.58
score	50K	31,231	4.5	5.3	0.54	0.59	4.3	5.5	0.56	0.59
	LD	6,073	5.1	4.7	0.48	0.62	5.0	4.8	0.49	0.63

[A] $\mathbf{G_V}$ and $\mathbf{G_Y}$: G matrices of VanRaden (2008) and the modification of the matrix of Yang *et al*. (2010), respectively; $\hat{\sigma}_e^2$ and $\hat{\sigma}_g^2$: estimated residual and additive genetic variances (kg^2 for carcass weight and $point^2$ for marbling score), respectively; $\hat{h}^2$: estimated heritability; r: prediction accuracy assessed as the correlation between the genomic estimated breeding values calculated in this chapter and the records corrected by all the fixed effects estimated in the official genetic evaluation using pedigree data for animals in the validation population, divided by the square root of estimated heritability.

Table 5-2 shows the results of the comparison of $\mathbf{G_V}$ and $\mathbf{G_Y}$. For all three SNP sets, correlation coefficients between the diagonal elements and those between the upper-triangular elements of $\mathbf{G_V}$ and $\mathbf{G_Y}$ were higher than 0.90 and 0.99, respectively. The regression coefficients for the diagonal elements and the upper-triangular elements were smaller and larger than 1, respectively. The differences among the coefficients using

different SNP sets were mainly due to the distribution of MAFs of SNPs in each SNP set (data not shown).

Table 5-2. The comparison of the two different G matrices constructed using the same SNP set [A]

SNP set	No. of SNPs	$\overline{p}$	Diagonal element			Upper-triangular element		
			Corr	Reg	Int	Corr	Reg	Int
HD	565,837	0.20	0.93	0.63	0.37	0.99	1.09	0.00
50K	31,231	0.20	0.93	0.60	0.39	0.99	1.10	0.00
LD	6,073	0.24	0.95	0.71	0.29	0.99	1.06	0.00

[A] $\overline{p}$: the value calculated as $0.5-\sqrt{0.25-\sum_{k=1}^{n} p_k\left(1-p_k\right)/n}$, where p_k is the frequency of counted allele at SNP *k* and n is the number of SNPs; Corr: correlation coefficient between the corresponding elements of $\mathbf{G_V}$ and $\mathbf{G_Y}$; Reg and Int: single regression coefficient and the intercept where independent and dependent variables were the corresponding elements of $\mathbf{G_Y}$ and $\mathbf{G_V}$, respectively; $\mathbf{G_V}$ and $\mathbf{G_Y}$: G matrices of VanRaden (2008) and the modification of the matrix of Yang *et al*. (2010), respectively.

The heritabilities estimated using all of the low- to lower-density SNP subsets and incorporating the two types of G matrices are depicted in Figure 5-1, together with those using the HD, 50K and LD SNP sets. The scale of the abscissa axis in the figure is $\log_{10}(x)$ where x is the number of SNPs included in the given SNP set or subset. The phenotypic variances estimated using all the SNP subsets were almost the same as those obtained using the HD SNP set, regardless of the type of G matrix used (data not shown). This is similar to what was found for the 50K and LD sets. The pattern of change in the

estimated heritability when increasing the number of SNPs used was very similar for both traits. When the lowest SNP density was considered, namely when the number of SNPs in the subset was 1/1000 of all available ones in the HD set, the values of the estimated genetic variance and heritability were both around 40% of those obtained using all the available SNPs in the HD set. As the density of SNPs in the subset was increased, the proportions of the genetic variance and heritability explained gradually increased. When using the SNP subset with a density of 1/100 of the HD set, the proportions reached around 80% of the values obtained using the HD set. When using lower-density SNP subsets, G matrix was calculated with a larger sampling variance, as indicated by VanRaden (2008), and subsequently, G matrix did not explain all the variance, which caused the resulting lower value of estimated heritability (Figure 5-1).

While there was a tendency for the use of $\mathbf{G}_{\mathbf{Y}}$ to result in slightly higher estimates of heritability, the effect of which of the two G matrices was used was very small for both traits, especially when the SNP density of the subset was low.

The correlations between the GEBVs and the corrected records in the validation population divided by the square root of estimated heritability (the prediction accuracy) obtained using low- to lower-density SNP subsets are shown in Figure 5-2, along with those obtained using the HD, 50K and LD sets. The correlations were above 0.7 for CW and around 0.4 for MS even when SNP subsets of a low-density were used. The correlations were around 80% and 90% of those obtained using the HD set when the SNP density of the subset was 1/1000 and 1/100 or greater, respectively, of that of the HD set. The differences between the correlations obtained using the two different G matrices were very small for both traits, in particular when using SNP subsets of a higher-density.

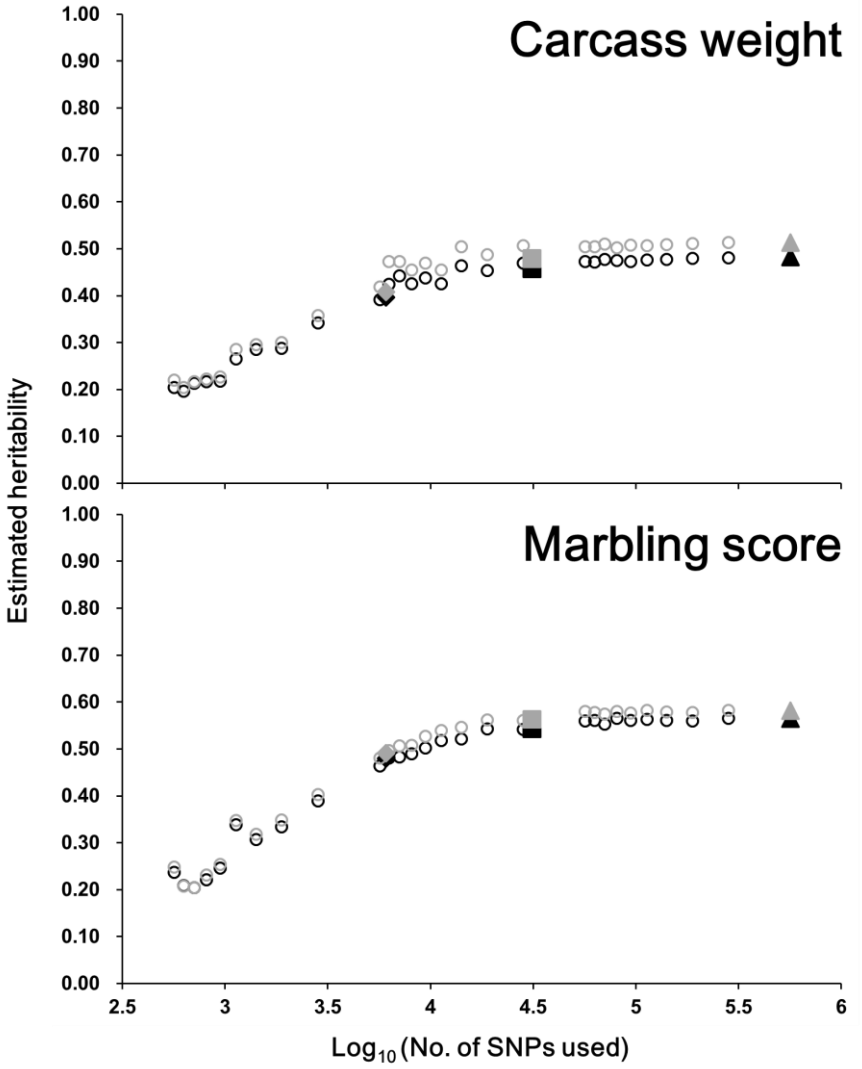

Figure 5-1. Changes in the value of estimated heritability. Triangles, squares and rhombuses show the results obtained using HD, 50K and LD SNP sets, respectively. Triangles, squares and rhombuses show the results obtained using HD, 50K and LD SNP sets, respectively. Circles represent the results using lower-density SNP subsets. Black and gray show the results using G matrices of VanRaden (2008) and the modification of the matrix of Yang et al. (2010), respectively.

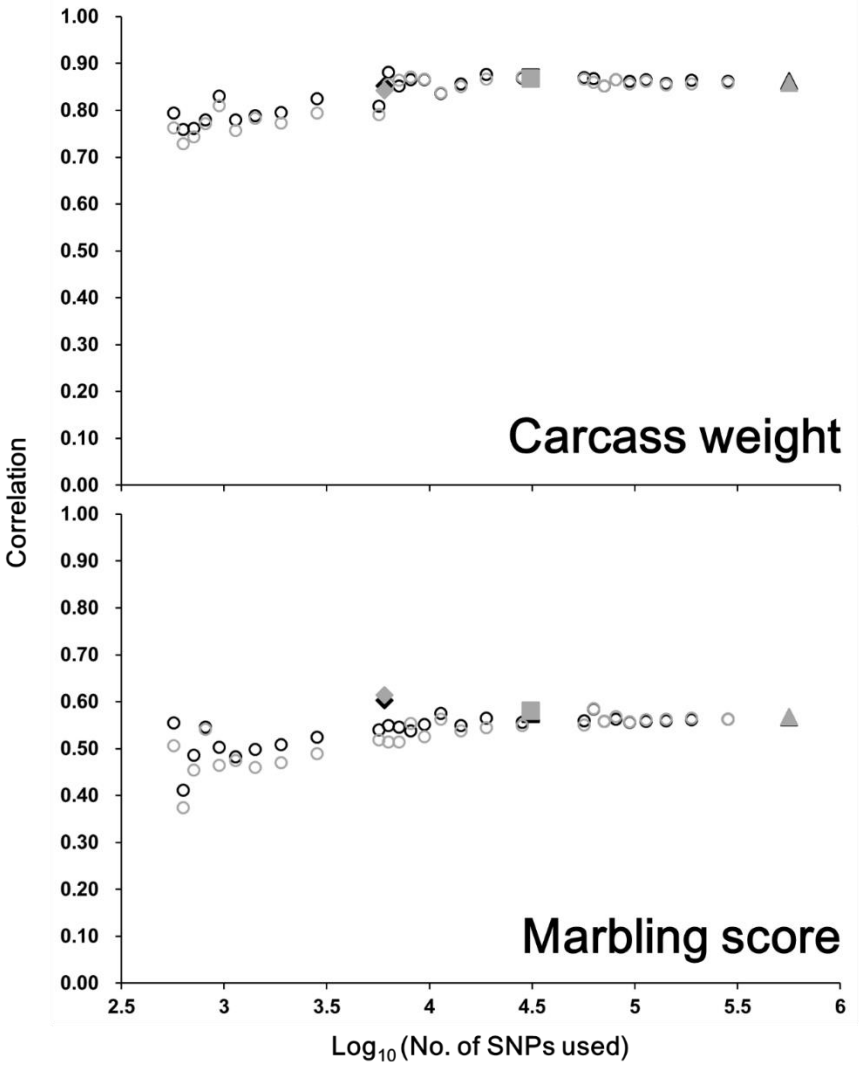

Figure 5-2. Changes in the prediction accuracy assessed as the correlation between the genomic estimated breeding values and the records corrected by all the fixed effects estimated in the official genetic evaluation using pedigree data for animals in the validation population, divided by the squared root of estimated heritability. Triangles, squares and rhombuses show the results obtained using HD, 50K and LD SNP sets, respectively. Circles represent the results using lower-density SNP subsets. Black and gray show the results using G matrices of VanRaden (2008) and the modification of the matrix of Yang et al. (2010), respectively.

5.4 Discussion

The differences were very small between the estimates of variances and heritability and the accuracy of GEBVs obtained using the two different G matrices (Table 5-1, Figures 5-1 and 5-2), although the estimates using $\mathbf{G}_{\mathbf{Y}}$ were slightly higher than those obtained using $\mathbf{G}_{\mathbf{V}}$. As stated by Meuwissen et al. (2011), elements of $\mathbf{G}_{\mathbf{V}}$ and $\mathbf{G}_{\mathbf{Y}}$ are a weighted and an unweighted mean of the relationships at the SNPs, respectively. While the approach of Yang et al. (2010) does not down-weight the information from low MAF SNPs, that of VanRaden (2008) gives extra weights to SNPs with high heterozygosity. It is likely that the quantitative trait loci of a selected trait may have rare alleles, and therefore, use of $\mathbf{G}_{\mathbf{Y}}$ rather than $\mathbf{G}_{\mathbf{V}}$ for such a trait may generally be preferable from the perspective of minimising the missing heritability. However, there did not appear to be a substantial advantage of using $\mathbf{G}_{\mathbf{Y}}$ in the present chapter. This might partly be due to the genetic architecture and the current genetic status of the two carcass traits in Japanese Black cattle and/or the quality-control criterion used for MAF. Both G matrices resulted in similar estimates of variance components, which agreed with the finding in previous studies (e.g. Clark et al., 2013). For single-step GBLUP, Forni et al. (2011) looked at effects of using different G matrices, including one of those considered in the present chapter.

As would be expected, using the HD set resulted in a larger proportion of the additive genetic variance being explained than was found for the other two sets, especially the LD set. However, the additional account of the genetic variance over that found using the 50K set was only very slight, as has been reported in the literature (e.g. Snelling et al., 2013). Even when using the HD SNP set, there might still be unaccounted genetic

variance for both traits, due to ascertainment bias in a SNP chip. Developing new SNP markers specific to the Japanese Black population may be a further necessary approach.

The levels of the estimated heritability, or genomic heritability, were found in this chapter to be approximately 0.5 for CW and 0.6 for MS, which were practically comparable to the levels of the unweighted means of previous estimates in Japanese Black cattle given by Oyama (2011).

There have been several studies about single-breed genomic prediction for CW and/or MS in beef cattle breed. Bolormaa et al. (2013), Boerner et al. (2014) and Chen et al. (2015) reported values lower than 0.4 as a prediction accuracy for CW in Angus breed, while Saatchi et al. (2011) and Boddhireddy et al. (2014) found a higher prediction accuracy. Use of different types of response variable, such as corrected record and estimated breeding value, resulted in different prediction accuracies (Boddhireddy et al., 2014; Fernandes Júnior et al., 2016). In the previous studies, there were also differences in training population size, SNP marker density, statistical model used, validation scheme, and so on. Thus, a direct comparison of prediction accuracies among those studies seems to be difficult; however, the current values of prediction accuracy shown in Table 5-1 were higher than many of the values reported in the previous studies. In this chapter, sires of a large part of the steers in the validation population were also sires of more than half of steers in the training population. This might have somewhat inflated the prediction accuracy.

Even in the cases where SNP subsets of a low-density were used, the GP accuracies were found to be above 0.7 for CW and around 0.4 for MS. The accuracies were around 80% and 90% of those obtained using the HD set, when the SNP density of a given subset was 1/1000 and 1/100, or greater, respectively. Assuming that the genomic

heritability found using the HD data in this chapter is the heritability of the trait, the GP accuracy obtained in this chapter is likely to correspond to progeny testing using several to 10 progenies for CW, even if the lower-density SNP subsets are used. This may indicate a certain validity of the use of GS for CW in this breed.

Habier et al. (2009) proposed the use of evenly-spaced low-density marker panels for cost-effective GS. This strategy may perform relatively well in Japanese Black cattle compared to the use of higher-density panels, as the Ne of this breed in Japan is relatively small compared with other typical cattle breeds (Nomura et al., 2001). Some studies reported the results about the use of low-density SNP subsets in GP with real dairy cattle data (e.g. Weigel et al., 2009; Moser et al., 2010; Vazquez et al., 2010), indicating that at least several thousands of evenly spaced SNPs may be needed to achieve the same degree of prediction accuracy as obtained using HD SNPs genotyped by the Illumina BovineSNP50 assay. Ogawa et al. (2014) indicated the possibility to rank animals genetically for carcass traits in Japanese Black cattle, with the information on at least 4000 of evenly spaced SNPs. In this chapter, with the independent validation population, when the SNP density was 1/100 of that of the HD set, the estimated heritability and prediction accuracy were around 80% and 90%, respectively, of values obtained using the HD set, irrespective of trait and type of G matrix, which was in agreement with the findings in Ogawa et al. (2014). However, for MS with a higher genomic heritability estimated, an obviously lower prediction accuracy was obtained in the validation step in this chapter, and such an accuracy corresponds to a progeny testing using only a few progenies. The use of a larger training population would be necessary to further improve the GP accuracy.

This chapter may indicate a possibility that using lower-density SNP information is at least useful in pre-selecting young Japanese Black breeding animals for the two carcass traits, especially for CW, while maintaining the highest possible genetic diversity. Also, a GP and GS strategy could be attractive when pedigree information is unavailable, such as is found in local native breeds, when the prediction equation is appropriately provided.

5.5 Summary

GP of breeding values using SNP markers can be conducted even when pedigree information is unavailable. In this chapter, using genotype information on a maximum of 565,837 SNP markers and the independent validation population from the training population, GP accuracy of CW and MS was assessed with varying marker density. The GP accuracy assessed as the correlation between the GEBVs and the corrected records divided by the square root of estimated heritability was around 0.85 for CW and 0.60 for MS, when using all available 565,873 SNPs. The difference in the way of calculating G matrix used, VanRaden (2008) and Yang et al. (2010) with minor modification, did not substantially affect the results for the two traits in this chapter. When SNP density was 1/1000 of that of all available 565,837 SNP, around 80% of prediction accuracy obtained using all available SNPs was retained. These results may indicate a possibility that using lower-density SNP information is at least useful in pre-selecting young Japanese Black breeding animals for the two carcass traits, especially for CW.

CHAPTER SIX
General discussion

GS proposed by Meuwissen et al. (2001) may enhance the rate of genetic progress for traits, including those that are measured relatively late in animal's production cycle and/or can only be measured by sacrificing potential breeding candidates such as carcass traits in beef cattle, because accurate GP can ideally perform once the genotype information of genome-wide high-density markers of candidates, provided that accurate marker effects are already estimated. GP is an important step in GS, because selection is carried out based on the result of GP. As with other livestock, it is critically important to investigate GP in Japanese Black cattle. Thus, this thesis provided fundamental scientific information about the possibility of GP using a commercial SNP chip for the two economically important representatives of carcass traits, CW and MS, in Japanese Black cattle.

Many animals need to be genotyped for GP, because a larger training population is preferable to perform more accurate GP (e.g., Goddard and Hayes, 2009). However, the cost of genotyping with a high-density SNP chip is still high. Use of low-density marker panels may be one possible approach to address this problem, and the possibility of using a larger training population would be increased by combining use of a low-density marker panel with genotype imputation. Consequently, use of low(er)-density SNP subsets in GP and the performance of genotype imputation with low(er)-density SNP subsets were investigated.

First, the effects of equally spaced SNP density on the G matrix, the genetic variance explained and the possibility of GP, as well as the extent of whole-genome LD,

were investigated using the genotype data of approximately 40,000 SNPs and two statistical models. Using all pairs of two adjacent SNPs throughout the whole marker sets, the means of r^2 spanned 0.22 at the range 0–0.1 Mb to 0.08 at the range 0.5–1 Mb, and approximately 26 and 6% of the r^2 values of the former and latter ranges exceeded 0.3, respectively. Whereas approximately 90% of the genetic variance for CW estimated using all available SNPs was explained using 4,000–6,000 SNPs, the corresponding percentage for MS was consistently lower. With the conventional linear model that incorporates the G matrix based on VanRaden (2008), correlation between the GEBVs obtained using 4,000 SNPs and all available SNPs was almost one for both traits, with an underestimation of the former GEBVs, especially for MS. Although a limited number of animals were available in this study, the Japanese Black is likely to be in a breed group with a relatively high extent of whole-genome LD. The results obtained were very suggestive and important, indicating that the degree of marbling is controlled by only QTLs with relatively small effects, unlike CW which is controlled by a few major genes (Mizoshita et al., 2004; Takasuga et al., 2007; Setoguchi et al., 2009; Nishimura et al., 2012) and many polygenes, and that there is a possibility of genetically ranking animals well using a kind of GP that employs at least several thousands of equally spaced SNPs for carcass traits in Japanese Black cattle.

As far as we know, this was the first report on the estimating the degree of whole-genome LD using genome-wide high-density SNP markers in Japanese Black cattle population, together with comparing the results obtained with those of other reports for different cattle breeds. Uemoto et al. (2015) also calculated r^2 values using more than 1,000 Japanese Black steers and cows with genotype information with the Illumina BovineHD BeadChip and reported that the extent of LD for the population

used was relatively higher and lower than that in foreign beef and dairy breeds, respectively, which agrees with our findings.

Next, the performance of genotype imputation using low(er)-density marker panels in Japanese Black cattle was evaluated. A target population of genotype imputation was the same as used in chapter two, and the independent reference population was provided. Population imputation was performed using Beagle 3.3.2 software. Imputation accuracy was assessed based on the average concordance rates of the genotypes, varying equally spaced SNP densities and the number of individuals in the reference population. Two additional statistics were calculated as indicators of imputation performance. It was found that the concordance rates tended to be lower for SNPs with greater minor allele frequencies, or those located near the ends of chromosomes. Additionally, longer autosomes were found to yield greater imputation accuracies than shorter ones. When SNPs were selected based on LD information, relative imputation accuracy slightly improved. When 3,000 and 10,000 equally spaced SNPs were used, the imputation accuracies were greater than 90% and 95%, respectively. This study clearly indicated that combining the genotyping using a low(er)-density SNP chip with genotype imputation based on a population of individuals genotyped using a higher-density SNP chip is a cost-effective and a valid approach for GP. It should be noted that use of imputed genotypes with such a low-density SNP panel for GWAS is strongly discouraged, because in classical GWAS, in contrast to GP, an analysis is performed using a model in which only one marker is fitted.

Imputation accuracy was assessed as the average of genotype concordance rates. Although the imputation accuracy was observed to be high when using

higher-density markers, the range of the values of the concordance rate was wide for both SNPs and individuals. Improving genotype imputation performance for particularly problematic SNPs and individuals is very important for further utilization of imputed genotype information in analyses.

Furthermore, the influence of genotype imputation using low(er)-density SNP marker subsets on the G matrix, genetic variance explained, and GP accuracy were investigated for CW and MS in Japanese Black fattened steers using genotype data on a relatively low number of SNPs. Genotypes were imputed using equally spaced SNP subsets of different densities. Two different linear models were used, incorporating only one G matrix constructed using a SNP marker subset and two different G matrices constructed using the selected and remaining SNPs. When using the model that incorporates only one G matrix, the estimated additive genetic variance was always larger when using all SNPs obtained via genotype imputation than when using only equally spaced SNP subsets. Correlations between the GEBVs obtained using genotype imputation with at least 3,000 SNPs and those using all available SNPs without imputation were found to be higher than 0.99 for both traits. Although additive genetic variance was likely to be partitioned with the model that incorporates two different G matrices, this model did not enhance GP accuracy compared with the model that incorporates only one G matrix. The findings in this study indicated that genotype imputation using an equally spaced low-density panel of an appropriate size can be used to conduct cost-effective and valid GP.

To date, there are several available SNP chips of different densities, including the Illumina BovineHD BeadChip. Use of higher-density chips could lead to an increase in the amount of additive genetic variance explained and improved GP accuracy. Then,

taking the results obtained in chapter two into consideration, GP using a lower-density marker subset would be expected to relatively perform well. However, the study conducted in chapter two did not use an independent validation population, and GP accuracy needed to be assessed using an independent validation population. Thus, in chapter five, GP was studied using SNP markers of varying densities and considering an independent validation population. Fattened steers with phenotypic data and animals with PBVs provided by the official genetic evaluation using pedigree data were treated as the training and validation populations, respectively. Genotype data on approximately 570,000 autosomal SNPs were available, and SNPs were selected to provide different equally spaced lower-density SNP subsets. GEBVs were obtained using GBLUP by incorporating one of two types of G matrices. GP accuracy, which was assessed as the correlation between the GEBVs and the corrected records divided by the square root of estimated heritability, was around 0.85 for CW and 0.60 for MS when using approximately 570,000 SNPs. The type of G matrix used gave no substantial difference in the results at a given SNP density for the traits examined. Around 80% of the GP accuracy was retained when SNP density was decreased to 1/1000 of that of all available SNPs. These results indicated that, even when a lower-density SNP panel is used, GP may be beneficial to the preselection for carcass traits in Japanese Black young breeding animals.

The GP accuracy assessed was relatively high, especially for CW, in Japanese Black cattle compared with other foreign beef breeds. One possible explanation for this finding is that the degree of genetic connectedness among animals might be higher in the Japanese Black population, because Ne of this breed is approximately 30 (Nomura et al., 2001), which could be lower than those of most of other major beef breeds, and

almost all Japanese Black breeding females are bred with proven sires by AI using frozen semen in Japan, whereas a smaller proportion of calves is usually generated from AI in foreign beef breeds. In chapter five, the training population was constructed of fattened steers marketed in two large domestic carcass markets in Osaka and Tokyo prefectures, and the validation population included fattened steers marketed in carcass markets in Tottori prefecture. This case of constructing the training population was quite different from the most common case, where the training population consists of the ancestors of selection candidates or animals sampled from the population from which the validation population was sampled. However, the genetic relationships between Japanese Black steers used in the training and validation populations are potentially higher for the reason mentioned above, resulting in high GP accuracy. The observed difference in GP accuracies between CW and MS is partly due to the different genetic architectures of the traits.

GP accuracy is derived from two sources: markers that capture additive genetic relationships but are in linkage equilibrium with QTL (linkage), and markers that are in LD with QTL (e.g., Habier et al., 2007). From this perspective, use of low-density marker panels in GP heavily depends on the former information, compared with use of higher-density panels. Jannink et al. (2010) broke down GP accuracy into that contributed by LD and that resulting from linkage using the method described by Habier et al. (2007) for varying training population size, marker number, and QTL number; the proportion of accuracy attributable to LD increased as the marker density and training population size increased. Habier et al. (2009) and Solberg et al. (2009) showed that the reduction in prediction accuracy in later generations was stronger when using lower-density compared with higher-density markers. These findings may indicate a

relative disadvantage of using lower-density rather than higher-density markers for GP with a larger training population; however, this disadvantage could be compensated when accurate genotype imputation is performed.

Although the corresponding information for the Japanese Black was not considered when developing the Illumina BovineSNP50 BeadChip (Matukumalli et al., 2008), genotype data on most of the SNPs contained (almost 40,000 out of 54,001 SNPs) were available after the standard quality control process in the study described in chapter two. There are two possible reasons that some SNPs occurred before the differentiation, and that some SNPs were introduced through crossing with British and Continental breeds in the early 1900s, ignoring the occurrence of the identical mutant after differentiation at the breed-level. Use of these available SNPs explained about half of the phenotypic variance for CW and MS, which were in similar levels of previously reported heritabilities estimated using pedigree information; this indicated the possibility of using SNPs genotyped by existing commercial chips in GP for Japanese Black cattle. Therefore, in chapter five, GP accuracy was assessed using an independent validation population in which individuals had PBVs based on official genetic evaluation in Tottori prefecture, resulting in considerably high prediction accuracy for CW. However, official genetic evaluation of carcass traits including CW and MS was performed separately in each prefecture, and a population in some prefecture may have more different genetic character than others. Genetic relationships between the training and validation populations also affect GP accuracy (e.g. Hayes et al., 2009). Therefore, additional investigations are necessary to assess GP accuracy with different validation populations with larger training population sizes, because the training population size was not fully large.

In this thesis, it was found that the assessed GP accuracy was higher than and almost the same as the accuracy of parent average for CW and MS, respectively, but lower than that of PBV obtained through the genetic evaluation for each trait. Based on these fact, it seems to be reasonable to judge that GP results could be used instead of parent average for pre-selection of young breeding animals. However, there is still what to toward a full-scale implementation of GS for carcass traits in this breed. For example, updating the content of the training population, as well as enlarging the training population, would be necessary to improve prediction accuracy, because LD between markers and QTLs changes across populations and generations (e.g., Meuwissen et al., 2001). Additionally, the fact that updating the training population is needed indicates the necessity to continue trait record collection. Moreover, no study has investigated aspects of long-term GS; for example, how often the training population needs to be updated and how often the SNP effects need to be re-estimated. Furthermore, the impact of implementing GS for carcass traits on the results of future genetic evaluations of carcass and other traits is entirely unknown. Therefore, continued and fundamental researches are still definitely necessary to provide further valuable insights into GP and GS, and to accomplish successful introduction of GS scheme in Japanese Black cattle.

In the near future, whole-genome sequence data may also become available for Japanese Black cattle. The advantage of whole-genome sequence data is that there is a potential increase of causative variants in marker data. Because the causative mutations are present in the sequence data, prediction accuracy may be further improved by addressing the ascertainment bias in a SNP chip, and the issue of decay in associations between causative mutations and SNPs, which results in the decline in accuracy over time, may be overcome (e.g., Meuwissen and Goddard, 2010). This situation will favor

statistical methods that assume a (large) fraction of the SNPs have no effect, such as BayesB, more than GBLUP using the G matrix based on the first method of VanRaden (2008), which was confirmed by Meuwissen and Goddard (2010). However, Hayes et al. (2013) denoted that this requires a training population in which LD between causative mutations and other variants is as limited as possible: if the extent of the LD is too great, the GP algorithms will distribute the effect of the causative mutation over a large number of SNPs in LD with the mutation, which has already been observed when increasing SNP marker density, for example, from 50,000 SNPs to approximately 800,000 SNPs (Erbe et al., 2012). It will be an especially challenging issue for Japanese Black cattle population, because Japanese Black cattle population is likely to be in a breed group with a relatively high extent of whole-genome LD, which was found in chapter two. Furthermore, the number of markers dramatically increases when using sequence data, which means that the problem of large p, small n is much severe. One strategy to deal with this problem will be to exploit prior knowledge, such as biological information (Hayes et al., 2013; MacLeod et al., 2016).

Single-step GBLUP (Aguilar et al., 2010; Christensen and Lund, 2010) may be one reasonable approach to incorporate genome-wide SNP information into conventional genetic evaluation. In this approach, a relationship matrix among animals is constructed by combining the A and G matrices to take advantage of all available information and to maximize the prediction accuracy, because it is very rarely the case that all animals in the evaluation have been genotyped, which is also the case in Japanese Black cattle. However, there are only a few reports about the results of single-step GBLUP in beef cattle (Onogi et al., 2014; Lourenco et al., 2015; Onogi et al., 2016). For all studies of this thesis, the pedigree information of the steers to construct

the A matrix was not available. Recently, Fernando et al. (2014) proposed a single-step Bayesian regression model approach that is not limited to normally distributed marker effects. The study may be required to investigate the performance of this kind of approach also in the Japanese Black cattle when the data are available.

In conclusion, genotype imputation for Japanese Black cattle population and GP for two economically important carcass traits, CW and MS, were studied. Although the number of samples was not so high, the results obtained provided positive scientific basis for successful application of GP in Japanese Black cattle. This thesis showed the possibility of relatively good performance of equally spaced low-density SNP panels in GP for some situations such as preselection of young breeding animals. It is indicated that genotype imputation with low-density SNP marker subsets could contribute to producing a larger training population while reducing genotyping cost for GP.

General summary

Japanese Black beef cattle are representative of the four breeds in Wagyu, or Japanese native cattle, and well known to excel in meat quality, especially in marbling. The current official genetic evaluation system for carcass traits in this breed using trait and pedigree information provides PBVs with high and higher accuracies for sires and breeding females, respectively, and contributes to successful genetic improvement of the traits. However, this system is cost- and time-consuming. GS based on GP using genome-wide, high-density SNP markers is expected to further accelerate the genetic improvement, but a very large number of animals must be genotyped to perform accurate GP. In this thesis, the possibility of GP using a commercial SNP chip was investigated for the two economically important carcass traits, CW and MS, in Japanese Black cattle.

First, as well as the extent of whole-genome LD in this breed, the effects of equally spaced SNPs densities on the G matrix, genetic variance explained, and GP were investigated using 38,502 available SNPs and employing linear and threshold models. Although the number of animals available was limited to approximately 900 fattened steers, the results indicated that the Japanese Black is likely to be in a breed group with a relatively high extent of whole-genome LD, use of genome-wide SNPs genotyped by a commercial chip with sufficient density has the potential to explain a large amount of additive genetic variance, and there is a possibility of genetically ranking animals for the traits using at least 4,000 equally spaced SNPs.

Next, using the same animals as before, the performance of genotype imputation with lower-density SNP panels was evaluated. Population imputation with

an independent reference population was performed using Beagle 3.3.2 software. When 3,000 and 10,000 equally spaced SNPs were used, the imputation accuracies were greater than approximately 90% and 97%, respectively. When SNPs were selected based on LD information, relative accuracy slightly improved. The influences of genotype imputation using equally spaced low-density SNP subsets on genetic variance explained and GP were also investigated. The estimated additive genetic variance was always larger when using all SNPs provided via genotype imputation than when using only equally spaced SNP subsets. The correlations between the GEBVs obtained using genotype imputation with at least 3,000 SNPs and those using all available SNPs without imputation were higher than 0.99 for both traits. These results indicated that genotype imputation with an equally spaced low-density panel of an appropriate size could be a promising approach to producing a cost-effective, valid GP for the two carcass traits in Japanese Black cattle.

Furthermore, GP was performed for the two carcass traits with a training population of 1,791 fattened steers and an independent validation population that consisted of 189 steers with PBVs obtained by the official genetic evaluation using trait and pedigree information. The total number of SNPs available was 565,837, and different equally spaced SNP subsets of lower-density were provided, selecting a part of SNPs from among all available ones. GP was performed via GBLUP, and two different G matrices were calculated from the same SNP data. The GP accuracy assessed was around 0.85 for CW and 0.60 for MS when using all available SNPs. The two types of G matrix used gave no substantial difference in the results at a given SNP density. For both traits, around 80% of GP accuracy was retained when the SNP density was decreased to 1/1000 of that of all available SNPs. These results indicated that, even

when a lower-density SNP panel is used, GP may be beneficial to the preselection for the carcass traits in Japanese Black cattle.

This thesis provided positive scientific bases for the successful application of GP for the carcass traits in Japanese Black cattle and showed the possibility of relatively good performance of using equally spaced low-density SNP panels for GP in some situations, such as preselection of young breeding animals. It is indicated that genotype imputation with low-density SNP marker subsets would contributes to producing a larger training population while reducing genotyping cost for GP.

ACKNOWLEDGMENTS

My deepest appreciation goes to Dr. Hiroaki Iwaisaki, Emeritus Professor at Graduate School of Agriculture, Kyoto University, for his excellent and inspiring guidance, valuable discussions and the kindest support. I have received a lot of motivation and encouragement from him during all my studies. I am also grateful to Dr. Yukio Taniguchi and Dr. Hirokazu Matsuda, Associate and Assistant Professors at Graduate School of Agriculture, Kyoto University, respectively, for their kind advice.

I am deeply indebted to the sub-supervisors Dr. Hiroshi Imai and Dr. Hiroyuki Hirooka, Professors at Graduate School of Agriculture, Kyoto University, for their critical review of this thesis and encouragement. I also wish to thank Dr. Shinichi Kume and Dr. Tohru Matsui, Professors at Graduate School of Agriculture, Kyoto University, for their counselling and encouragement.

Furthermore, special thanks are expressed to Dr. Yoshikazu Sugimoto and Dr. Toshio Watanabe at the Shirakawa Institute of Animal Genetics, Mr. Tabuchi Ichiro and Ms. Yuki Kitamura at the Tottori Prefectural Agriculture and Forest Research Institute, Livestock Research Center for their kind provision of the data as study material in this thesis. I also thank the staff of the Shirakawa Institute of Animal Genetics for technical assistance. I would also like to express my gratitude to the financial support by the Research Fellowship of the Japanese Society for the Promotion of Science for Young Scientists (No. 15J02417).

I would like to express my gratitude to all those related directly or indirectly to completion of this thesis.

Lastly, I wish to thank my family for their constant and generous support.

REFERENCES

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

Almasy, L., and J. Blangero 1998. Multipoint quantitative-trait linkage analysis in general pedigrees. Am. J. Hum. Genet. 62:1198–1211.

Ashida, I., and H. Iwaisaki. 1998. A numerical technique for REML estimation of variance components using average information algorithm and its computing property. Anim. Sci. Technol. (Jpn.) 69:631–636.

Ashida, I., and H. Iwaisaki. 1999. An expression for average information matrix for a mixed linear multi-component of variance model and REML iteration equations. Anim. Sci. J. 70:282–289.

Badke, Y. M., R. O. Bates, C. W. Ernst, C. Schwab, J. Fix, C. P. Van Tassell, and J. P. Steibel. 2013. Methods of tagSNP selection and other variables affecting imputation accuracy in swine. BMC Genet. 14:8.

Barendse, W., S. M. Armitage, L. M. Kossarek, A. Shalom, B. W. Kirkpatrick, A. M. Ryan, D. Clayton, L. Li, H. L. Neibergs, N. Zhang, W. M. Grosse, J. Weiss, P. Creighton, F. McCarthy, M. Ron, A. J. Teale, R. Fries, R. A. McGraw, S. S. Moore, M. Georges, M. Soller, J. E. Womack, and D. J. S. Hetzel1. 1994. A genetic linkage map of the bovine genome. Nat. Genet. 6:227–235.

Bernardo, R. 2008. Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci. 48:1649–1664.

Berry D. P., J. F. Garcia, and D. J. Garrick. 2016. Development and implementation of

genomic predictions in beef cattle. Animal Frontiers 6:1.

Berry, D. P., M. C. McClure, and M. P. Mullen. 2014. Within- and across-breed imputation of high-density genotypes in dairy and beef cattle from medium- and low-density genotypes. J. Anim. Breed. Genet. 131:165–172.

Bishop M. D., S. M. Kappes, J. W. Keele, R. T. Stone, S. L. Sunden, G. A. Hawkins, S. S. Toldo, R. Fries, M. D. Grosz, and J. Yoo. 1994. A genetic linkage map for cattle. Genetics. 67:35–43.

Boddhireddy P., M. J. Kelly, S. Northcutt, K. C. Prayaga, J. Rumph, and S. DeNise. 2014. Genomic prediction in Angus cattle: Comparisons of sample size, response variables, and clustering methods for cross-validation. J. Anim. Sci. 92:485–497.

Boerner, V., D. J. Johnston, and B. Tier. 2014. Accuracies of genomically estimated breeding values from pure-breed and across-breed predictions in Australian beef cattle. Genet. Sel. Evol. 46:61.

Boichard, D., H. Chung, R. Dassonneville, X. David, A. Eggen, S. Fritz, K. J. Gietzen, B. J. Hayes, C. T. Lawley, T. S. Sonstegard, C. P. Van Tassell, P. M. VanRaden, K. A. Viaud-Martinez, and G. R. Wiggans. 2012. Design of a bovine low-density SNP array optimized for imputation. PLoS ONE 7:e34130.

Bolormaa, S., J. E. Pryce, K. Kemper, K. Savin, B. J. Hayes, W. Barendse, Y. Zhang, C. M. Reich, B. A. Mason, R. J. Bunch, B. E. Harrison, A. Reverter, R. M. Herd, B. Tier, H.-U. Graser, and M. E. Goddard. 2013. Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality in *Bos Taurus*, *Bos indicus*, and composite beef cattle. J. Anim. Sci. 91:3088–3104.

Browning, B. L., and S. R. Browning. 2011. A fast, powerful method for detecting identity by descent. Am. J. Hum. Genet. 88:173–182.

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. Am. J. Hum. Genet. 81:1084–1097.

Bulmer, M. G. 1980. The mathematical theory of quantitative genetics. Oxford University Press, Oxford, UK.

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

Cantet, R. J. C., and C. Smith. 1991. Reduced animal model for marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. 23:221–233.

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.

Chen, L., M. Vinsky, and C. Li. 2015. Accuracy of predicting genomic breeding values for carcass merit traits in Angus and Charolais beef cattle. Anim. Genet. 46:55–59.

Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:1–8.

Clark, S. A., B. P. Kinghorn, and J. H. J. van der Werf. 2013. Comparisons of identical by state and identical by descent relationship matrices derived from SNP markers in genomic evaluation. Proc. Assoc. Advmt. Anim. Breed. Genet. 20:261–265.

Corbin, L. J., A. Kranis, S. C. Blott, J. E. Swinburne, M. Vaudin, S. C. Bishop, and J. A. Wooliams. 2014. The utility of low-density genotyping for imputation in the Thoroughbred horse. Genet. Sel. Evol. 46:9.

Dassonneville, R., R. F. Brøndum, T. Druet, S. Fritz, F. Guillaume, B. Guldbrandtsen, M.

S. Lund, V. Ducrocq, and G. Su. 2011. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. J. Dairy Sci. 94:3679–3686.

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

de los Campos, G., A. Pataki, and P. Pérez. 2013a. The BGLR (Bayesian Generalized Linear Regression) R-Package. http://bglr.r-forge.r-project.org/BGLR-tutorial.pdf.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J. M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385.

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

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

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.

Endelman, J. B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250–255.

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

Espigolan, R., F. Baldi, A. A. Boligon, F. R. P. Souza, D. G. M. Gordo, R. L. Tonussi, D. F. Cardoso, H. N. Oliveira, H. Tonhati, M. Sargolzaei, F. S. Schenkel, R. Carvalheiro, J. A. Ferro, and L. G. Albuquerque. 2013. Study of whole genome linkage disequilibrium in Nellore cattle. BMC Genomics 14:305.

Fernandes Júnior, G. A., G. J. M. Rosa, B. D. Valente, R. Carvalheiro, F. Baldi, D. A. Garcia, D. G. M. Gordo, R. Espigolan, L. Takada, R. L. Tonussi, W. B. F. de Andrade, A. F. B. Magalhaes, L. A. L. Chardulo, H. Tonhati, and L. G. de Albuquerque. 2016. Genomic prediction of breeding values for carcass traits in Nellore cattle. Genet. Sel. Evol. 48:7.

Fernando, R. L., and M. Grossman. 1989. Marker-assisted selection using best linear unbiased prediction. Genet. Sel. Evol. 21:467–477.

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

Fisher, R. A. 1918. The correlation between relatives on the supposition of mendelian inheritance. Trans. R. Soc. Edinb. 52:399–433.

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

García-Ruiz, A., J. B. Cole, P. M. VanRaden, G. R. Wiggans, F. J. Ruiz-López a, and C. P. Van Tassel. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc. Natl. Acad. Sci. USA 113:E3995–E4004.

Garrick, D. J. 2011. The nature, scope and impact of genomic prediction in beef cattle in

the United States. Genet. Sel. Evol. 43:17.

Gianola, D., G, de los Campos, W. G. Hill, E. Manfredi, and R. L. Fernando. 2009. Additive genetic variability and the Bayesian alphabet. Genetics 183:347–363.

Goddard, M. E. 1992. A mixed model for analyses of data on multiple genetic markers. Theor. Appl. Genet. 83:878–886.

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

Goddard, M. E., B. J. Hayes, and T. H. E. Meuwissen. 2010. Genomic selection in livestock populations. Genet. Res., Camb. 92:413–421.

Gunia, M., R. Saintilan, E. Venot, C. Hozé, M. N. Fouilloux, and F. Phocas. 2014. Genomic prediction in French Charolais beef cattle using high-density single nucleotide polymorphism markers. J. Dairy Sci. 92:3258–3269.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2009. Genomic selection using low-density marker panels. Genetics 182:343–353.

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

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

Harris, B. L., F. E. Creagh, A. M. Winkelman, and D. L. Johnson. 2011. Experiences with the Illumina high density Bovine BeadChip. Interbull Bull. 44:3–7.

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

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009a. Invited review:

genomic selection in dairy cattle: progress and challenge. J. Dairy Sci. 92:433–443.

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

Hayes, B. J., P. J. Bowman, H. D. Daetwyler, J. W. Kijas, and J. H. J. van der Werf. 2011. Accuracy of genotype imputation in sheep breeds. Anim. Genet. 43:72–80.

Henderson C. R. 1973. Sire evaluation and genetic trends. In: Proceedings of the Animal Breeding and Genetics Symposium in Honour of J. L. Lush. American Society for Animal Science, Blackburgh, Champaign, IL, pp. 10–41.

Henderson C. R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447.

Henderson, C. R. 1984. Application of Linear Models in Animal Breeding. University of Guelph, Guelph, Ontario.

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

Hill, W. G., and A. Robertson. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38:226–231.

Hoeschele, I., and B. Tier. 1995. Estimation of variance components of threshold characters by marginal posterior modes and means via Gibbs sampling. Genet. Sel. Evol. 27:519–540.

Hozé, C., M. N. Fouilloux, E. Venot, F. Guillaume, R. Dassonneville, S. Fritz, V. Ducrocq, F. Phocas, D. Boichard, and P. Croiseau. 2013. High-density marker imputation accuracy in sixteen French cattle breeds. Genet. Sel. Evol. 45:33.

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

Japan Meat Grading Association (JMGA). 1988. New Beef Carcass Grading Standards. Japan Meat Grading Association, Tokyo.

Jensen, J., G. Su, and P. Madsen. 2012. Partitioning additive genetic variance into genomic and remaining polygenic components for complex traits in dairy cattle. BMC Genet. 13:44.

Johnson D. L., and R. Thompson. 1995. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. J. Dairy Sci.78:449–456.

Kachman, S. D., M. L. Spangler, G. L. Bennett, K. J. Hanford, L. A. Kuehn, W. M. Snelling, R. M. Thallman, M. Saatchi, D. J. Garrick, R. D. Schnabel, J. F. Taylor, and E. J. Pollak. 2013. Comparison of molecular breeding values based on within- and across-breed training in beef cattle. Genet. Sel. Evol. 45:30.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. 2008. Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723.

Kashi, Y., E. Hallerman, and M. Soller. 1990. Marker-assisted selection of candidate bulls for progeny testing programmes. Anim. Prod. 51:63–74.

Khansefid, M., J. E. Pryce, S. Bolormaa, S. P. Miller, Z. Wang, C. Li, and M. E. Goddard. 2014. Estimation of genomic breeding values for residual feed intake in a multibreed cattle population. J. Anim. Sci. 92:3270–3283.

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

13:538.

Kizilkaya, K., R. L. Fernando, and D. J. Garrick. 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J. Anim. Sci. 88:544–551.

Lee, S. H., B. H. Park, A. Sharma, C. G. Dang, S. S. Lee, T. J. Choi, Y. H. Choy, H. C. Kim, K. J. Jeon, S. D. Kim, S. H. Yeon, S. B. Park, and H. S. Kang. 2014. Hanwoo cattle: origin, domestication, breeding strategies and genomic selection. J. Anim. Sci. Technol. 56:2.

Legarra, A., C. R. Granie, E. Manfredi, and J. M. Elsen. 2008. Performance of genomic selection in mice. Genetics 180:611–618.

Liu, Y., X. Qin, X.-Z. H. Song, H. Jiang, Y. Shen, K. J. Durbin, S. Lien, M. P. Kent, M. Sodeland, Y. Ren, L. Zhang, E. Sodergren, P. Havlak, K. C. Worley, G. M. Weinstock, and R. A. Gibbs. 2009. Bos taurus genome assembly. BMC Genomics 10:180.

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

Lu, D., M. Sargolzaei, M. Kelly, C. Li, G. V. Voort, Z. Wang, G. Plastow, S. Moore, and S. P. Miller. 2012. Linkage disequilibrium in Angus, Charolais, and Crossbred beef cattle. Front. Genet. 3:152.

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

Matsuda, H., and H. Iwaisaki. 2001. A mixed model method to predict QTL-cluster effects using trait and marker information in a multi-group population. Genes Genet. Syst. 76:81–88.

Matsuda, H., and H. Iwaisaki. 2002a. Prediction of additive genetic effects for the QTL-cluster on the basis of data on surrounding markers in outbred populations. J. Appl. Genet. 43:193–207.

Matsuda, H., and H. Iwaisaki. 2002b. The genetic variance for multiple linked quantitative trait loci conditional on marker information in a crossed population. Heredity 88:2–7.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. C'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4:e5350.

McKay, S. D., R. D. Schnabel, B. M. Murdoch, L. K. Matukumalli, J. Aerts, W. Coppieters, D. Crews, E. D. Neto, C. A. Gill, C. Gao, H. Mannen, P. Stothard, Z. Wang, C. P. Van Tassel, J. L. Williams, J. F. Taylor, and S. S. Moore. 2007. Whole genome linkage disequilibrium maps in cattle. BMC Genet. 8:74.

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

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

Meuwissen, T. H. E., and J. A. M. van Arendonk. 1992. Potential improvements in rate of genetic gain from marker-assisted selection in dairy cattle breeding schemes. J. Dairy Sci. 75:1651–1659.

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

Meuwissen, T. H. E., and M.E. Goddard. 1996. The use of marker haplotypes in animal breeding schemes Genet. Sel. Evol. 28:161–176.

Meuwissen, T. H. E., and M. E. Goddard. 1997. Estimation of effects of quantitative trait loci in large complex pedigrees. Genetics 146:409–416.

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.

Mizoshita, K., T. Watanabe, H. Hayashi, C. Kubota, H. Yamakuchi, J. Todoroki, and Y. Sugimoto. 2004. Quantitative trait loci analysis for growth and carcass traits in a half-sib family of purebred Japanese Black (Wagyu) cattle. J. Anim. Sci. 82:3415–3420.

Moser, G., M. S. Khatkar, B. J. Hayes, and H. W. Raadsma. 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet. Sel. Evol. 42:37.

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. J. Dairy Sci. 95:876–889.

Namikawa, K. 1992. Wagyu: Japanese Beef Cattle - Historical Breeding Processes of Japanese Beef Cattle and Preservation of Genetic Resources as Economic Farm Animal. Wagyu Registry Association, Kyoto, Japan.

Nejati-Javaremi, A., C. Smith, and J. P. Gibson. 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. J. Anim. Sci. 75:1738–1745.

Neves, H. H. R., R. Carvalheiro, A. M. Pérez O'Brien, Y. T. Utsunomiya, A. S. do Carmo,

F. S. Schenkel, J. Sölkner, J. C. McEwan, C. P. Van Tassell, J. B. Cole, M. V. G. B. da Silva, S. A. Queiroz, T. S. Sonstegard., and J. F. Garcia. 2014. Accuracy of genomic predictions in Bos indicus (Nellore) cattle. Genet. Sel. Evol. 46:17.

Nicolazzi, E. L., S. Biffani, and G. Jansen. 2013. Short communication: imputing genotypes using PedImpute fats algorithm combining pedigree and population information. J. Dairy Sci. 96:2649–2653.

Nishimura, S., T. Watanabe, K. Mizoshita, K. Tatsuda, T. Fujita, N. Watanabe, Y. Sugimoto, and A. Takasuga. 2012. Genome-wide association study identified three major QTL for carcass weight including the *PLAG1-CHCHD7* QTN for stature in Japanese Black cattle. BMC Genet. 13:40.

Nomura, T., T. Honda, and F. Mukai. 2001. Inbreeding and effective population size of Japanese Black cattle. J. Anim. Sci. 79:366–370.

Odani, M., A. Narita, T. Watanabe, K. Yokouchi, Y. Sugimoto, T. Fujita, T. oguni, M. Matsumoto, and Y. Sasaki. 2006. Genome-wide linkage disequilibrium in two Japanese beef cattle breeds. Anim. Genet. 37:139–144.

Ogawa, S., H. Matsuda, Y. Taniguchi, T. Watanabe, A. Takasuga, Y. Sugimoto, and H. Iwaisaki. 2016. Accuracy of imputation of single nucleotide polymorphism marker genotypes from low-density panels in Japanese Black cattle. Anim. Sci. J. 87:3–12.

Ogawa, S., H., Matsuda, Y., Taniguchi, T., Watanabe, S., Nishimura, Y., Sugimoto, and H. Iwaisaki. 2014. Effects of single nucleotide polymorphism marker density on degree of genetic variance explained and genomic evaluation for carcass traits in Japanese Black beef cattle. BMC Genet. 15:15.

Onogi, A., A. Ogino, T. Komatsu, N. Shoji, K. Simizu, K. Kurogi, T. Yasumori, K. Togashi, and H. Iwata. 2014. Genomic prediction in Japanese Black cattle:

Application of a single-step approach to beef cattle. J. Anim. Sci. 92:1931–1938.

Onogi, A., A. Ogino, T. Komatsu, N. Shoji, K. Simizu, K. Kurogi, T. Yasumori, K. Togashi, and H. Iwata. 2015. Whole-genome prediction of fatty acid composition in meat of Japanese Black cattle. Anim. Genet. 46:557–559.

Oyama, K. 2011. Genetic variability of Wagyu cattle estimated by statistical approaches. Anim. Sci. J. 82:367–373.

Pagnacco, G., and G.B. Jansen. 2001. Use of marker haplotypes to refine covariances among relatives for breeding value estimation. J. Anim. Breed. Genet. 114:185–189.

Park, T., and G. Casella. 2008. The Bayesian Lasso. J. Am. Stat. Assoc. 103:681–686.

Patterson, H. D., and R. Thompson. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58:545–554.

Pimentel, E. C. G., and S. König. 2012. Genomic selection for the improvement of meat quality. J. Anim. Sci. 90:3418–3426.

Porto-Neto, L. R., W. Barendse, J. M. Henshall, S. M. McWilliam, S. A. Lehnert, and A. Reverter. 2015. Genomic correlation: harnessing the benefit of combining two unrelated populations for genomic selection. Genet. Sel. Evol. 47:84.

Pryce, J. E., B. Gredler, S. Bolormaa, P. J. Bowman, C. Egger-Danner, C. Fuerst, R. Emmerling, J. Sölkner, M. E. Goddard, and B. J. Hayes. 2011. Short communication: Genomic selection using a multi-breed, across-country reference population. J Dairy Sci. 94:2625–2630.

Pryce, J. E., J. Johnston, B. J. Hayes, G. Sahana, K. A. Weigel, S. McParland, D. Spurlock, N. Krattenmacher, R. J. Spelman, E. Wall, and M. P. L. Calus. 2014. Imputation of genotypes from low density (50,000 markers) to high density (700,000 markers) of cows from research herds in Europe, North America, and Australasia using 2 reference

populations. J. Dairy Sci. 97:1799–1811.

Quaas, R. L. 1976. Computing the diagonal elements of a large numerator relationship matrix. Biometrics 32:949–953.

R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. Vienna, Austria. http://www.r-project.org/.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rolf, M. M., J. F. Taylor, R. D. Schnabel, S. D. McKay, M. C. McClure, S. L. Northcutt, M. S. Kerley, and R. L. Weaber. 2010. Impact of reduced marker set estimation of genomic relationship matrices on genomic selection for feed efficiency in Angus cattle. BMC Genet. 11: 24.

Saatchi, M., M. C. McClure, S. D. McKay, M. M. Rolf, J. W. Kim, J. E. Decker, T. M. Taxis, R. H. Chapple, H. R. Ramey, S. L. Northcutt, S. Bauck, B. Woodward, J. C. M. Dekkers, R. L. Fernando, R. D. Schnabel, D. J. Garrick, and J. F. Taylor. 2011. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genet. Sel. Evol. 43:40.

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

Saito, S., and H. Iwaisaki. 1996. A reduced animal model with elimination of quantitative trait loci equations for marker-assisted selection. Genet. Sel. Evol. 28:465–477.

Saito, S., and H. Iwaisaki. 1997a. A reduced animal model approach to predicting total additive genetic merit for marker-assisted selection. Genet. Sel. Evol. 29:25–34.

Saito, S., and H. Iwaisaki. 1997b. Back-solving in combined-merit models for marker-assisted best linear unbiased prediction of total additive genetic merit. Genet. Sel.

Evol. 29:611–616.

Saito, S., H. Matsuda, and H. Iwaisaki. 1998. Best linear unbiased prediction of additive genetic merit using a combined-merit sire and dam model for marker-assisted selection. Genes Genet. Syst. 73:65–69.

Sasaki, S., T. Ibi, T. Watanabe, T. Matsuhashi, S. Ikeda, and Y. Sugimoto. 2013. Variants in the 3′ UTR of General Transcription Factor IIF, polypeptide 2 affect female calving efficiency in Japanese Black cattle. BMC Genet. 14:41.

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

Schulz-Streeck, T., J. O. Ogutu, and H.-P. Piepho. 2011. Pre-selection of markers for genomic selection. BMC Proc. 5(Suppl 3):S12.

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

Setoguchi, K., M. Furuta, T., Hirano, T. Nagao, T. Watanabe, Y. Sugimoto, and A. Takasuga. 2009. Cross-breed comparisons identified a critical 591-kb region for bovine carcass weight QTL (*CW-2*) on chromosome 6 and the Ile-442-Met substitution in *NCAPG* as a positional candidate. BMC Genet. 10:43.

Silva, C. R., H. H. R. Neves, S. A. Queiroz, J. A. D. Sena, and E. C. G. Pimentel. 2010. Extent of linkage disequilibrium in Brazilian Gyr dairy cattle based on genotypes of AI sires for dense SNP markers. In: Proc. 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany. pp. 1–29.

Smith, C., and S. P. Simpsom, 1986. The use of genetic polymorphisms in livestock improvement. J. Anim. Breed. Genet. 103:1–5.

Snelling, W. M., R. A. Cushman, J. W. Keele, C. Maltecca, M. G. Thomas, M. R. S. Fortes, and A. Reverter. 2013. BREEDING AND GENETICS SYMPOSIUM: Networks and pathways to guide genomic selection. J. Anim. Sci. 91:537–552.

Solberg, T. R., A. K. Sonesson, J. A. Wooliams, J. Ødegard, and T. H. E. Meuwissen. 2009. Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. Genet. Sel. Evol. 41:53.

Soller, M. 1978. The use of loci associated with quantitative traits in dairy cattle improvement. Anim. Prod. 27:133–139.

Soller, M., and J. S. Beckmann. 1983. Genetic polymorphism in varietal identification and genetic improvement. Theor. Appl. Genet. 67:25–33.

Takahata, N. 1993. Allelic genealogy and human evolution. Mol. Biol. Evol. 10:2–22.

Takasuga, A., T. Watanabe, Y. Mizoguchi, T. Hirano, N. Ihara, A. Takano, K. Yokouchi, A. Fujikawa, K. Chiba, N. Kobayashi, K. Tatsuda, T. Oe, M. Furukawa-Kuroiwa, A. Nishimura-Abe, T. Fujita, K. Inoue, K. Mizoshita, A. Ogino, and Y. Sugimoto. 2007. Identification of bovine QTL for growth and carcass traits in Japanese Black cattle by replication and identical-by-descent mapping. Mamm. Genome 18:125–136.

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.

Tibshirani, R. 1994. Regression Shrinkage and Selection Via the Lasso. J. R. Statist. Soc. B 58:267–288.

Uemoto, Y., S. Sasaki, Y. Sugimoto, and T. Watanabe. 2015. Accuracy of high-density

genotype imputation in Japanese Black cattle. Anim. Genet. 46:388–394.

van Arendonk, J. A. M., B. Tier, and B. P. Kinghorn, 1994. Use of multiple genetic markers in prediction of breeding values. Genetics 137:319–329.

Van Eenennaam, A. L., J. H. J. van der Werf, and M. E. Goddard. 2011. The value of using DNA markers for beef bull selection in the seed stock sector. J. Anim. Sci. 89:307–320.

Van Eenennaam, A. L., K. A. Weigel, A. E. Young, M. A. Cleveland, and J. C. M. Dekkers. 2014. Applied animal genomics: results from the field. Annu. Rev. Anim. Biosci. 2:105–139.

VanRaden, P. M. 2008. Efficient methods to compute genomic prediction. J. Dairy Sci. 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. J. Dairy Sci. 92:16–24

Vazquez, A. I., G. J. M. Rosa, K. A. Weigel, G. de los Campos, D. Gianola, and D. B. Allison. 2010. Predictive ability of subset of single nucleotide polymorphisms with and without parent average in US Holsteins. J. Dairy Sci. 93:5942–5949.

Ventura, R. V., D. Lu, F. S. Schenkel, Z. Wang, C. Li, and S. P. Miller. 2014. Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle. J. Anim. Sci. 92:1433–1444.

Vereijken, A. L. J., G. A. A. Albers, and J. Visscher. 2010. Imputation of SNP genotypes in chicken using a reference panel with phased haplotypes. In: Proc. 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany. pp. 1–6.

Villa-Angulo, R., L. K. Matukumalli, C. A. Gill, J. Choi, C. P. Van Tassell, and J. J. Grefenstette. 2009. High-resolution haplotype block structure in the cattle genome. BMC Genet. 10:19.

Visscher, P. M., W. G. Hill, and N. R. Wray. 2008. Heritability in the genomics era-concepts and misconceptions. Nat. Rev. Genet. 9:255–266.

Wagyu Registry Association. 2007. Breeding and Improvement of Wagyu. 2nd edition. Wagyu Registry Association, Kyoto, Japan. (in Japanese)

Wagyu Registry Association. 2014. 'Wagyu.' No. 269. Kyoto, Japan. p. 23. (in Japanese)

Watanabe, T., H. Matsuda, A. Arakawa, T. Yamada, H. Iwaisaki, S. Nishimura, and Y. Sugimoto. 2014. Estimation of variance components for carcass traits in Japanese Black cattle using 50K SNP genotype data. Anim. Sci. J. 85:1–7.

Weigel, K. A., G. de los Campos, O. González-Recio, H. Naya, X. L. Wu, N. Long, G. J. M. Rosa, and D. Gianola. 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J. Dairy Sci. 92:5248–5257.

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. J. Dairy Sci. 93:5423–5435.

Weigel, K. A., C. P. Van Tassell, J. R. O'Connell, P. M. VanRaden, and G. R. Wiggans. 2010a. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population based imputation algorithms. J. Dairy Sci. 93:2229–2238.

Whittaker, J. C., R. Thompson, and M. C. Denham. 2000. Marker-assisted selection using ridge regression. Genet. Res. Camb. 75:249–252

Wiggans, G. R., T. A. Cooper, P. M. VanRaden, K. M. Olson, and M. E. Tooker. 2011. Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation. J. Daily Sci. 95:1552–1558.

Yan, J, T. Shah, M. L. Warburton, E. S. Buckler, M. D. McMullen, and J. Crouch. 2009. Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PLoS ONE. 4:e8451.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42:565–569.

Yang, J., T. A. Manolio, L. R. Pasquale, E. Boerwinkle, N. Caporaso, J. M. Cunningham, M. de Andrade, B. Feenstra, E. Feingold, M. G. Hayes, W. G. Hill, M. T. Landi, A. Alonso, G. Lettre, P. Lin, H. Ling, W. Lowe, R. A. Mathias, M. Melbye, E. Pugh, M. C. Cornelis, B. S. Weir, M. E. Goddard, and P. M. Visscher. 2011. Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. 45:519–525.

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

Zhang, Z., X. Ding, J. Liu, Q. Zhang, and D. J. de Koning. 2011. Accuracy of genomic prediction using low-density marker panels. J. Daily Sci. 94:3642–3650.

Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard, G. Marçais, M. Roberts, P. Subramanian, J. A. Yorke, and S. L. Salzberg. 2009. A whole-genome assembly of the domestic cow, Bos Taurus. Genome Biol. 10:R42.