



## Accurate Estimation of Effective Population Size in the Korean Dairy Cattle Based on Linkage Disequilibrium Corrected by Genomic Relationship Matrix

Dong-Hyun Shin<sup>a</sup>, Kwang-Hyun Cho<sup>1,a</sup>, Kyoung-Do Park<sup>2</sup>, Hyun-Jeong Lee<sup>3</sup>, and Heebal Kim\*

Department of Agricultural Biotechnology, Animal Biotechnology Major,  
Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul 151-921, Korea

**ABSTRACT:** Linkage disequilibrium between markers or genetic variants underlying interesting traits affects many genomic methodologies. In many genomic methodologies, the effective population size ( $N_e$ ) is important to assess the genetic diversity of animal populations. In this study, dairy cattle were genotyped using the Illumina BovineHD Genotyping BeadChips for over 777,000 SNPs located across all autosomes, mitochondria and sex chromosomes, and 70,000 autosomal SNPs were selected randomly for the final analysis. We characterized more accurate linkage disequilibrium in a sample of 96 dairy cattle producing milk in Korea. Estimated linkage disequilibrium was relatively high between closely linked markers ( $>0.6$  at 10 kb) and decreased with increasing distance. Using formulae that related the expected linkage disequilibrium to  $N_e$ , and assuming a constant actual population size,  $N_e$  was estimated to be approximately 122 in this population. Historical  $N_e$ , calculated assuming linear population growth, was suggestive of a rapid increase  $N_e$  over the past 10 generations, and increased slowly thereafter. Additionally, we corrected the genomic relationship structure per chromosome in calculating  $r^2$  and estimated  $N_e$ . The observed  $N_e$  based on  $r^2$  corrected by genomics relationship structure can be rationalized using current knowledge of the history of the dairy cattle breeds producing milk in Korea. (**Key Words:** Linkage Disequilibrium, Dairy Cattle, Effective Population Size, SNP)

### INTRODUCTION

The recent availability of SNP genotyping allows for the application of genomic techniques to domestic animals. Genomic techniques such as genome-wide association studies and genomic selection depend on the extent of linkage disequilibrium and its rate of decline with distance between loci within a population. These genomic tools depend on sample size and the quality of linkage disequilibrium estimation. Therefore, accurate

characterization of linkage disequilibrium will assist in planning future studies using genomic techniques.

Linkage disequilibrium can provide insights into the evolutionary history of a population through the effective population size ( $N_e$ ) (Falconer et al., 1996). Using  $N_e$ , we can monitor genetic diversity in livestock populations and explain the observed range and pattern of genetic variation with regard to population genetics. In addition, if pedigrees are incomplete or unavailable,  $N_e$  provides information into the inbreeding level of a population. The strength of linkage disequilibrium at different genetic distances between makers can be used to infer ancestral  $N_e$ . The pattern of historical  $N_e$  in animal populations can increase our understanding of the impact of selective breeding strategies on the genetic variation within the framework of population genetics.

Researchers have already applied SNP chips for genome-wide association study (Huang et al., 2010; Jiang et al. 2010; Sahana, 2010) and genomic selection (Schaeffer, 2006; Hayes et al., 2009) in dairy cattle. The pattern of linkage disequilibrium in cattle has been characterized, and

\* Corresponding Author: Heebal Kim. Tel: +82-2-880-4803, E-mail: heebal@snu.ac.kr

<sup>1</sup> Division of Animal Breeding and Genetics, National Institute of Animal Science, Rural Development Administration, Cheonan 331-801, Korea.

<sup>2</sup> Genomic Informatics Center, Hankyong National University, Anseong 456-749, Korea.

<sup>3</sup> Division of Animal Genomics and Bioinformatics, National Institute of Animal science, Rural Development Administration, #564 Omockchun-dong, Suwon, 441-706, Korea.

<sup>a</sup> These authors contributed equally to this work.

Submitted Jun. 4, 2013; Accepted Oct. 2, 2013; Revised Oct. 17, 2013

predictions of  $N_e$  estimated using SNP chip data have already been made. Flury (2010), based on 128 Swiss Eringer breed genotyped using the Illumina BovineSNP50 BeadChip, estimated linkage disequilibrium-based actual  $N_e$  and ancestral pedigree-based  $N_e$  (Flury et al., 2010). Other studies have evaluated the historical  $N_e$  of a variety of cattle breeds, all of which suggested a continuous decrease in  $N_e$  since the time of domestication (Thevenon, et al., 2007; De Roos et al., 2008).

$N_e$  is an important measure in the global dairy cattle industry as well as the Korean dairy cattle industry in identifying the state of genetic resource. As Korea is a major-semen importing country in the dairy cattle industry (Table S4), dairy cattle in Korea have diverse genetic sources. In this study, we used  $r^2$  corrected by the genomic relationship structure based on SNPs, which is more accurate than the existing  $r^2$  to estimate linkage disequilibrium. We then estimated the current  $N_e$  and inferred an accurate  $N_e$ . These results are compared with other studies and considered in the context of current knowledge for the establishment of genomics methods in dairy cattle in Korea.

## MATERIALS AND METHODS

### Genotypic data

Almost all of the dairy cattle in Korea are Holstein, which have been produced by selected domestic seed bull or imported semen. Therefore, the Korean dairy cattle population in this manuscript is the Holsteins population in Korea. The country of origin for imported semen is shown in Table S4. The 96 Holstein samples were collected from four field dairy farms located in Gyeonggi and Gangwon provinces. The samples were collected by the Animal and Plant Quarantine Agency of Korea (a governmental agency) by random sampling for a survey on population disease resistance in Korea. The collected samples for this study represent the Korean Holstein cattle population despite the limitations of the sample size and is treated as a single population sample representing the dairy cattle of Korea. There are few regional differences in the dairy cattle in Korea because the supply and management of dairy cattle semen or seed bull takes place at the national level.

The Illumina BovineHD Genotyping BeadChip, which contains 777 kb SNP located across all autosomes, sex chromosomes, and mitochondria was used. These informative SNPs were selected from the Bovine Genome Database. Genotyping data were analyzed using Plink for quality control (GENO>0.05, MAF<0.1, HWE test  $p \leq 0.0001$ ), which removed 226,156 SNP from the analysis because of poor genotyping quality (Purcell et al., 2007). As the Illumina BovineHD Genotyping BeadChip contains numerous SNP at a high density, the distance between

adjacent SNPs is very short making the computing time long. To reduce the computational time, we used 70,000 randomly selected SNPs including only autosomal SNPs after quality control. Interbull (Uppsala, Sweden; www.interbull.org) uses BovineSNP50 Genotyping BeadChips, which contains 54,609 SNP to estimate genetic parameters and hence, we considered 70,000 SNPs sufficient for our purposes.

### Linkage disequilibrium

We estimated linkage disequilibrium using the "LDcorSV" package implemented in R. This package provides a set of functions to measure the existing  $r^2$  and the  $r^2$  corrected by the sample structure (Mangin et al., 2012). Estimating the  $r^2$  of all pairwise per chromosome using high-density SNP chips is time-consuming. To reduce computing time, the input for each chromosome was split into files containing 100 loci, and  $r^2$  was calculated for all syntenic marker pairs in each file as has been done in previous studies (Flury et al., 2010).

The standard measures of existing  $r^2$  are respectively equivalent to the covariance and the correlation between alleles at two different loci (Hill and Robertson, 1968), computed as:

$$r^2 = \frac{D^2}{P_A P_a P_B P_b} \quad (1)$$

where  $P_A$ ,  $P_a$ ,  $P_B$ , and  $P_b$  are the respective frequencies of alleles A, a, B and b, and D is  $P_{AB} - P_A P_B$ . Sample structure information is required for the corrected  $r^2$ . We used the genomic relationship matrix based on SNP data instead of pedigree data. In this study, the genomic relationship value of individual pairs was the number of common SNP between two individuals/number of total allele sites. In this way, we proposed a simple genomic relationship matrix (96×96). A process of calculating  $r^2$  corrected by the genomic relationship structure based on SNP was identical to the existing  $r^2$  calculation.

Details on the physical position of the markers can be found in the Illumina product literature ([http://support.illumina.com/array/array\\_kits/bovinehd\\_dna\\_analysis\\_kit/downloads.ilmn](http://support.illumina.com/array/array_kits/bovinehd_dna_analysis_kit/downloads.ilmn)). To determine linkage disequilibrium in relation to the physical distance between markers, marker pairs were divided into distance bins. We established two kinds of classes (0 to 0.5Mb, 0 to 5Mb) and subsequently, applicable marker pairs to each class were put into 50 distance bins with bin ranges dependent on the class (see Table S2). The two types of mean  $r^2$  for each of the distance bins were then plotted against the median of the distance bin range (Mb). This estimation of  $r^2$  was performed on a chromosome by chromosome basis; the

pooled results are presented in Figure 2.

Also  $r^2$  was calculated for a random selection of non-syntenic SNP. Across the autosome, 700 SNP were randomly selected and  $r^2$  values were calculated for only non-syntenic marker pairs. This resulted in a total of 250,096 pairwise comparisons that were not corrected by the genomic relationship structure.

### Construction model of linkage disequilibrium with distance

Under the assumption of an isolated population with random mating, Sved (1971) derived an approximate expression for the expectation of  $r^2$  (Sved, 1971):

$$E(r^2) = (1+4Nc)^{-1} \quad (2)$$

where  $N$  is the effective population size, and  $c$  is the recombination frequency. In this study, as in previous studies,  $c$  was replaced by linkage distance in Morgans (Hayes et al., 2003; Tenesa et al., 2007; Thevenon et al., 2007; De Roos et al., 2008; Villa-Angulo et al., 2009; Corbin et al., 2010; Flury et al., 2010; Qanbari et al., 2010). This was justified by the approximation of the more precise equation for  $E(r^2)$  given by Sved (1971). Based on this formula, a non-linear least-squares approach to statistically model the observed  $r^2$  was implemented within R (nl function) using this model:

$$y_i = \frac{1}{(a + 4bd_i)} + e_i \quad (3)$$

where  $y_i$  is the value of  $r^2$  for marker pair  $i$ , at linkage distance  $d_i$  in Morgans. Parameters  $a$  and  $b$  were estimated iteratively using the least-squares method. Chromosome-specific megabase-to-centimorgan conversion rates were calculated based on total physical chromosome length as stated on the UCSC Web site, and total chromosome genetic length derived from the bovine linkage map (Arias et al., 2009) (see Table S2). We used marker pairs for which  $r^2$  values were higher than the mean of  $r^2$  for non-syntenic marker pairs. This model was applied to each chromosome in turn and the parameters were estimated. Similar to Corbin (2010), estimated parameters were combined by meta-analysis in R using an inverse variance method for pooling and random effects method based on the DerSimonian-Laird method (DerSimonian and Laird, 1986; Corbin et al., 2010).

### Ancestral effective population size estimation

Equation (2) can predict the effective population size at a given point in time, expressed as generations in the past (Hayes et al., 2003; De Roos et al., 2008).

$$N_T(t) = (4c)^{-1}[(r_c^2)^{-1} - 1] \quad (4)$$

where  $N_T$  is the effective population size  $t$  generations ago,  $c$  is the distance between markers in Morgans,  $r_c^2$  is the mean value of  $r^2$  for marker pairs  $c$  Morgans apart, and  $c = (2t)^{-1}$  when assuming linear growth (Hayes et al., 2003). As previously mentioned, marker pairs with linkage distances less than the mean of  $r^2$  for non-syntenic marker pairs ( $<0.014248$ ) were excluded from this analysis. To estimate  $N_T$ , the number of prior generations was selected and a suitable range for  $c$  was calculated. The binning process was designed to ensure sufficient marker pairs within each bin and obtain a representative  $r^2$  mean (see Table S3). This process was performed for markers pooled across autosomes.

### Estimation of effective population size based on the genomic relationship structure per chromosome

For the 96 genotyped individuals, we were able to establish genomic relationship structure per chromosome using SNP information of each chromosome. We proposed  $r^2$  corrected by the genomic relationship structure per chromosome using "LDcorSV" package implemented in R (Mangin et al., 2012). A process of estimating  $N_e$  based on genomic relationship per chromosome and ancestral  $N_e$  was identical to that described above.

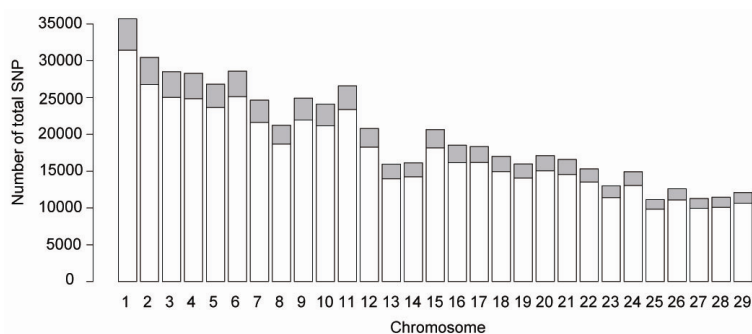
## RESULTS

### Genotypic data

Of the 734,862 genotyped autosomal SNPs, 508,707 (69.2%) remained after quality control processes and 70,000 were selected for analysis. The number of SNPs per autosome remaining after filtering and selection ranged from 1,300 to 4,280 and was closely related to chromosome length and total number of SNP as shown in Figure 1. The minor allele frequency of remaining SNP followed a uniform distribution and averaged ( $\pm$ SD) to be  $0.31 \pm 0.11$ . The average distance between marker pairs ( $\pm$ standard deviation) for this analysis was  $1,202.77 \pm 932$  kb, with the distance between markers ranging from 0.134 kb to 8,398 kb.

### Linkage disequilibrium

Linkage disequilibrium declined with increasing distances between marker pairs, as shown in Figure 2a and b. The most rapid decrease was seen over the first four bins, with the mean  $r^2$  decreasing by about half. The mean existing  $r^2$  decreased more slowly with increasing distance and was constant after 0.2 Mb of distance. The mean  $r^2$  corrected by the genomic relationship structure based on SNP with distance between marker pairs was slightly less



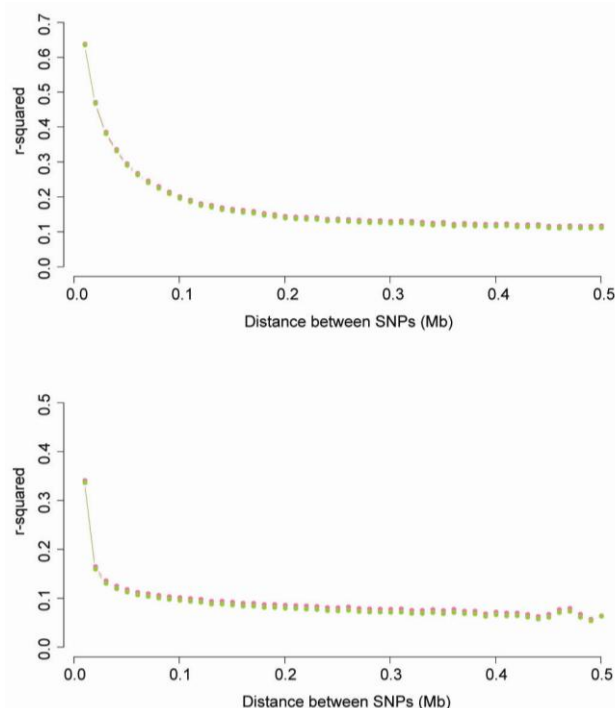
**Figure 1.** Number of SNPs per chromosome. Gray indicates the portion of used SNPs and white denotes the unused portion.

than that of the existing  $r^2$ . However, change patterns in the  $r^2$  means with distance for both methods were similar. The decline in linkage disequilibrium showed slight differences with log-transformed distance (Figures S1 and S2). According to the existing  $r^2$ , 23,797 of the 3,385,800 marker pairs were in complete linkage disequilibrium; 12,225 of these were adjacent pairs. For  $r^2$  corrected by the genomic relationship structure based on SNP, 23,794 of the 3,385,800 marker pairs were in complete linkage disequilibrium; 12,223 of these were adjacent pairs. The mean ( $\pm$ standard deviation)  $r^2$  between random non-syntenic marker pairs was  $0.014248 \pm 0.0197$ . 835,324 using

the existing  $r^2$  measure and 861,676 using  $r^2$  corrected by the genomic relationship structure based on SNP are less than the mean of non-syntenic marker pairs.

### Construction model of linkage disequilibrium with distance

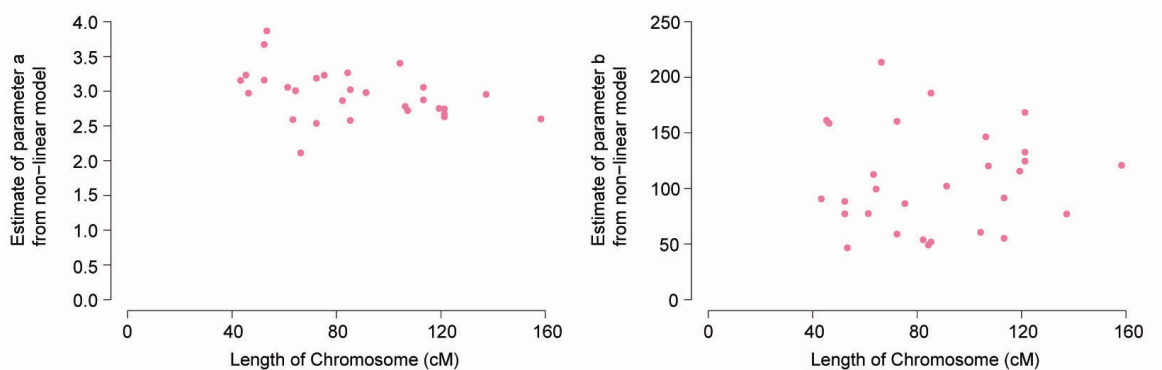
The non-linear regression model of the declining linkage disequilibrium with distance resulted in both parameters, a and b, being significantly different from zero. For parameter estimation using the existing  $r^2$ , the mean estimate and 95% confidence interval by meta-analysis across autosomes for parameters a and b were 2.95 (2.84; 3.07) and 106.32 (92.95; 119.70), respectively. The parameters a and b for  $r^2$  corrected by the genomic relationship structure based on SNP were 2.87 (2.75; 2.99) and 122.28 (107.21; 137.34), respectively. The predicted  $r^2$  from the non-linear regression equation was similar to the mean observed  $r^2$  of regions with massed bins with discrepancies occurring in other regions for both two types of  $r^2$  (Figures S1 and S2). Parameter b showed greater variability between chromosomes than parameter a. No such relationship was observed between the estimated parameters (a, b) and chromosome length (cM) as shown in Figure 3 and 4. This reason for the lack of relationship and interpretation of parameter b in this non-linear regression model as an estimated  $N_e$  is demonstrated in the discussion.



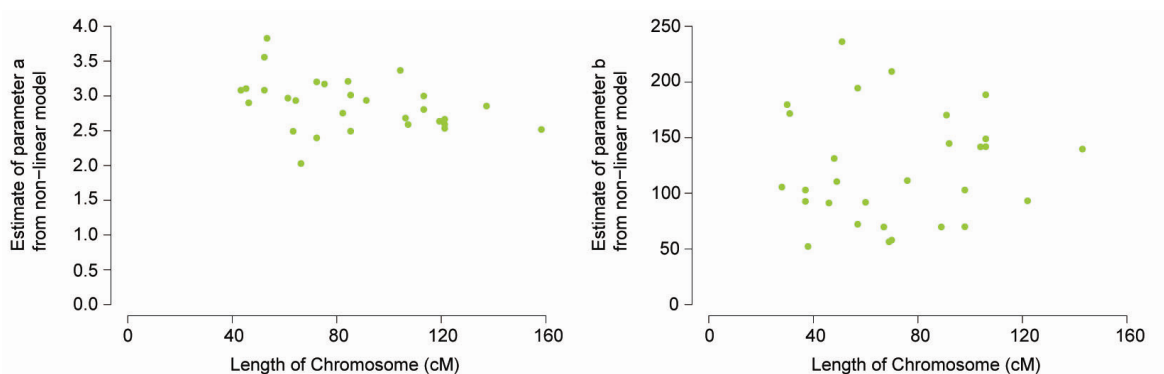
**Figure 2.** Average linkage disequilibrium (solid line) plotted against the median of the distance bin range (Mb). Each hot pink and yellow-green color represents the existing  $r^2$  and the  $r^2$  corrected by the genomic relationship structure based on SNP, respectively. (a) Distances ranged from 0 to 0.5 Mb.  $r^2$  values were averaged using bins of 0.01 Mb and pooled over autosomes. (b) Distances ranged from 0 to 5 Mb.  $r^2$  values were averaged using bins of 0.1 Mb and pooled over autosomes.

### Ancestral effective population size estimation

We observed rapid increase in  $N_e$  over the past 10 generations with the values increasing fivefold (close to 500) by 10 generations ago (Figure 5). From the past 10 generations to the past 100 generations,  $N_e$  increased slowly. Our results suggest that a continuous increase in  $N_e$  occurred over the last 100,000 generations (Figure S3). Overall, the tendencies for the two different  $r^2$  were similar, but  $N_e$  based on  $r^2$  corrected by the genomic relationship structure based on SNP was slightly higher than  $N_e$  based on the existing  $r^2$ . In Figure 5, although the recent difference between the two estimated measures was small, the difference in value increased with time.



**Figure 3.** Parameter estimates from equation (3) plotted against chromosome length (cM) according to the bovine linkage map using the existing  $r^2$  (Arias et al., 2009). (a) Estimates of parameter a plotted against chromosome length (cM). (b) Estimates of parameter b plotted against chromosome length (cM).

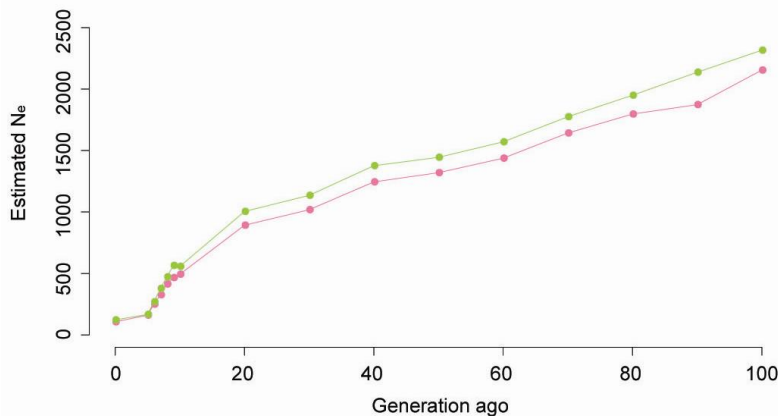


**Figure 4.** Parameter estimates from equation (3) plotted against chromosome length (cM) according to the bovine linkage map using the  $r^2$  corrected by the genomic relationship structure based on single nucleotides (Arias et al., 2009). (a) Estimates of parameter a plotted against chromosome length (cM). (b) Estimates of parameter b plotted against chromosome length (cM).

**Estimation of effective population size based on genomic relationship per chromosome**

Linkage disequilibrium corrected by the genomic relationship structure per chromosome declined with increasing distance between marker pairs, as shown in Figure S4a and b. Over the first four bins, the decline was greater than two previous results. The mean  $r^2$  corrected by

the genomic relationship structure per chromosome with distance between marker pairs was less than the two previous mean but the change pattern was similar. (Figure S5) In total, 23,783 of marker pairs were in complete linkage disequilibrium; 12,220 of these were adjacent pairs. For parameter estimation using  $r^2$  corrected by the genomic relationship structure per chromosome, the mean estimate



**Figure 5.** Average estimated effective population size plotted against generations in the past, truncated at 100 generations. Each hot pink and yellow-green color dot represents the use of the existing  $r^2$  and the  $r^2$  corrected by the genomic relationship structure based on SNP, respectively.

and 95% confidence interval by meta-analysis across autosomes for parameters a and b were 2.11 (2.04; 2.18) and 361.73 (335; 339.47), respectively. Parameter b was approximately three times greater than the two previous values for parameter b. The decline in linkage disequilibrium was almost linear with log-transformed distance and the decline in linkage disequilibrium with log-transformed distance was better fitted with predicted values than that of the other two. No such relationship was observed between parameters (a, b) and chromosome length (cM) in  $r^2$  corrected by the genomic relationship structure per chromosome as shown in Figure S6. Similar to previous patterns, we observed a rapid increase in  $N_e$  over the past 10 generations and then a slow increase up to 100 generations ago. However, the values increased fivefold (close to 1,500) at 10 generation than two kinds of  $r^2$  (Figure S7). From 10 generation ago, our results suggest that continuous increase in  $N_e$  has occurred over the 100,000 generations. (Figure S8) Overall, these tendencies were similar with that of the previous two but  $N_e$  based on  $r^2$  corrected by the genomic relationship structure per chromosome was greater than the two previous kinds of  $r^2$  measures. Although recent difference with two estimated measures was big, their difference in value decreased with generations ago.

## DISCUSSION

This study provides an overview of linkage disequilibrium in dairy cattle in Korea using a high density SNP panel. Validation work by Khatkar (2008) on their cattle data suggested that the  $r^2$  correlation between 100 and 1,000 samples is 0.94 (Khatkar et al., 2008). Thus, we can obtain an unbiased picture of linkage disequilibrium (3,385,801 pairs) in our population of 96 dairy cattle. The pattern of decline of linkage disequilibrium in this population is consistent with that reported by Flury (2010) in 128 Swiss cattle (Flury et al., 2010) with existing  $r^2$  remaining higher than approximately 0.3 for distance up to 0.05 Mb. The linkage disequilibrium observed is higher at short distances and more extensive than that observed in human populations (Shifman et al., 2003). Overall,  $r^2$  values corrected by the genomic relationship structure based on SNP were slightly lower than the existing  $r^2$  because the correction process prevented the overestimation of linkage disequilibrium.

Our population consisted of individuals from several genetic origins, which produces linkage disequilibrium between unlinked loci because of differences in allele frequencies. Consequently, such a structured sample can lead to a biased estimate of linkage disequilibrium, which may increase false positives rates. Therefore  $r^2$  was corrected for by inferring  $N_e$  to take into account the non-independence of loci due to population differentiations (Yu

et al., 2005). Not corrected  $r^2$  can be a biased estimation of linkage disequilibrium, which could lead to estimated genetic parameters that reduce the power of analysis (Mangin et al., 2012). We used an R package ("LDcorSV") to calculate  $r^2$  corrected by the genomic relationship structure based on SNP, which was slightly smaller than the existing  $r^2$ . This was expected, as we speculated that the calculated  $r^2$  corrected by the genomic relationship structure based on SNP was more accurate. Thus we can infer an accurate  $N_e$  of dairy cattle in Korea.

The mean value of  $r^2$  between non-syntenic markers was 0.014248, which provides an approximation of the linkage disequilibrium that can be expected by chance. The value observed here is higher than the value observed by Khatkar (2008) in a sample of over 1,500 cattle (0.0032) (Khatkar et al., 2008). Sample size and genetic sampling (drift) affect the mean of non-syntenic  $r^2$  values, and hence the mean may be expected to decrease with an increase in sample size. The larger non-syntenic value observed by Khatkar (2008) may be more affected by large populations. Since the sample size was smaller in this study than the other studies, the larger non-syntenic values of our dataset are reasonable. We used marker pairs with more than mean of  $r^2$  for non-syntenic marker pairs as a standard for estimating linkage disequilibrium. Many low  $r^2$  can cause overestimated  $N_e$  more than expected. Therefore we decided to use marker pairs more than mean of non-syntenic marker pairs for inferring  $N_e$ .

Using Sved's (1971) formula for the expected  $r^2$ , a non-linear regression model was fitted to the data to describe the relationship between linkage distance and linkage disequilibrium. Our estimate of parameter a supports an alternative version of Sved's (1971) equation (Sved, 1971), derived by Tenesa (2007), which accounts for mutation and puts  $a = 2$  (Tenesa et al., 2007). While estimating parameters, the initial value of parameter a was two with this approach. The estimated parameter a ranged from 2 to 3. For the heterogeneity of variance of the observed  $r^2$ , variance of  $r^2$  declined with increasing distances between markers, which may have impacted our results. A significant negative relationship between chromosome length (cM) and estimates of parameter b from the non-linear model have been observed (Corbin et al., 2010), while others have observed a positive relationship in domestic livestock species (Khatkar et al., 2008; Muir et al., 2008). In this study, all marker pairs were calculated in each bin so  $r^2$  was not affected by chromosome length. Thus, we could not observe a relationship between chromosome length (cM) and estimates of b.

Our estimate of b represents an estimated  $N_e$  assuming a constant population size. However, this assumption is difficult to maintain, b represents a conceptual average of  $N_e$  over the period inferred from the range of marker pairs

distances (Toosi et al., 2010). Two measures of  $r^2$  resulted in two different estimates of  $N_e$ .  $N_e$  based on the existing measure of  $r^2$  is about 106 and  $N_e$  based on  $r^2$  corrected by the genomic relationship structure based on SNP was about 122. Assuming that  $r^2$  corrected by the genomic relationship structure based on SNP was more accurate than the existing  $r^2$ , we predict that  $N_e$  of dairy cattle in Korea is about 122. Figure 5, S3, S7, and S8 show historical  $N_e$  assuming a linear population following Hayes et al. (2003). The observed pattern shows a rapid increase in  $N_e$  up to around 10 generations ago. Several explanations exist for this pattern including bottlenecks associated with domestication, selection and breed formation, and endangerment of the breed. Therefore, it is useful to consider our results in the context of the demographic history of the dairy cattle in Korea. The reliability of this method depends both on the technical implementation and approached used in a previous study approach (Corbin et al., 2010).

To the best of our knowledge, this is a novel study on  $N_e$  estimation based on  $r^2$  corrected by the genomic relationship structure per chromosome resulting in an underestimation of inbreeding and thus an overestimation of the current population size. The estimates for recent  $N_e$  for dairy cattle in Korea based on  $r^2$  corrected by the genomic relationship structure based on SNP was around 120 individuals in contrast to estimates in the range of 361 using  $r^2$  corrected by the genomic relationship structure per chromosome. Interesting thing is that recent  $N_e$  for Swiss Eringer breed using pedigree information covering 15 generations (the range of 110 to 321) is three times higher than the linkage disequilibrium-based estimates for recent  $N_e$  (around 100 individuals). Thus, we guess that the number of genomic relationship matrix that covers generation depends on how we establish the genomic relationship structure. These differences are important for inferring  $N_e$ .

The extent of linkage disequilibrium in a population can be used to estimate the SNP density required for genome-wide association studies to be effective, as well as providing some indication of genomic selection. This has generated thresholds for useful linkage disequilibrium described as the proportion of QTL (quantitative trait locus) variance explained by a marker (Zhao et al., 2005). The consensus is that an average  $r^2 > 0.3$  will permit reasonable sample sizes for genome-wide association studies (Ardlie et al., 2002; Du et al., 2007; Khatkar et al., 2008). In this data set, markers 200 kb apart achieved an average linkage disequilibrium of  $r^2 = 0.303$  excluding marker pairs less than the mean of non-syntenic marker pairs. However, marker pairs with  $r^2 = 1$ , which represent the high variability of  $r^2$  values at small distances are typically excluded in genomic selection. This is likely due to underestimation of the actual number of SNP needed.

Genomic selection appears to be effective at lower average  $r^2$  with simulation results demonstrating accuracies of up to 0.65 with an average  $r^2$  between adjacent markers as low as 0.2 and a trait heritability of 0.1 (Calus et al., 2008). Deterministic equations demonstrates that the accuracy of genomic selection can be expressed as a function of the effective number of loci in a population (Daetwyler et al., 2010). The effective number of loci in a population relates to the number of independent chromosome segments and assumes a random mating population. Our dataset covered an effective number of loci. However, because the estimated  $N_e$  was greater than our sample size, we required more preparation to predict the potential accuracy of genomic selection in dairy cattle populations.

We used dense SNP genotype data to characterize linkage disequilibrium and infer the ancestral  $N_e$  for a sample of dairy cattle. In the population studied, linkage disequilibrium extended for long distances, reaching baseline levels at more than 5 Mb. From the decay in linkage disequilibrium with genetic distance, we inferred the ancestral  $N_e$  and observed a recent rapid increase in  $N_e$  which reached approximately 500, 10 generations ago followed by a decrease until the present time. The final results were that we used correction by the genomic relationship structure to ensure accurate derivation of  $N_e$ , resulting in 122 individuals.

## ACKNOWLEDGEMENTS

The work was supported by Project (PJ0092602013) of the National Livestock Research Institute of Rural Development Administration, Republic of Korea.

## REFERENCES

- Ardlie, K. G., L. Kruglyak, and M. Seielstad. 2002. Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.* 3:299-309.
- Arias, J. A., M. Keehan, P. Fisher, W. Coppieters, and R. Spelman. 2009. A high density linkage map of the bovine genome. *BMC Genet.* 10:18.
- Calus, M., A. de Roos, and R. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. *Genetics* 178:553-561.
- Corbin, L., S. Bishop, J. Swinburne, M. Vaudin, S. Blott, and J. Woolliams. 2010. The impact of method on the estimated effective population size of a Thoroughbred population using genotype data. *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production.*
- Corbin, L., S. Blott, J. Swinburne, M. Vaudin, S. Bishop, and J. Woolliams. 2010. Linkage disequilibrium and historical effective population size in the Thoroughbred horse. *Anim. Genet.* 41:8-15.
- Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams. 2010. The impact of genetic architecture on

- genome-wide evaluation methods. *Genetics* 185:1021-1031.
- De Roos, A., B. J. Hayes, R. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics* 179:1503-1512.
- DerSimonian, R. and N. Laird. 1986. Meta-analysis in clinical trials. *Control. Clin. Trials* 7:177-188.
- Du, F.-X., A. C. Clutter, and M. M. Lohuis. 2007. Characterizing linkage disequilibrium in pig populations. *Int. J. Biol. Sci.* 3: 166-178.
- Falconer, D. S., T. F. Mackay, and R. Frankham. 1996. Introduction to Quantitative Genetics (4th edn). *Trends in Genetics* 12:280.
- Flury, C., M. Tapio, T. Sonstegard, C. Drögemüller, T. Leeb, H. Simianer, O. Hanotte, and S. Rieder. 2010. Effective population size of an indigenous Swiss cattle breed estimated from linkage disequilibrium. *J. Anim. Breed. Genet.* 127:339-347.
- Goddard, K. A., P. J. Hopkins, J. M. Hall, and J. S. Witte. 2000. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am. J. Hum. Genet.* 66:216-234.
- Hayes, B., P. Bowman, A. Chamberlain, and M. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.* 92:433-443.
- Hayes, B. and M. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819-1829.
- Hayes, B. J., P. M. Visscher, H. C. McPartlan, and M. E. Goddard. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* 13:635-643.
- Hill, W. and A. Robertson. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38:226-231.
- Huang, W., B. Kirkpatrick, G. Rosa, and H. Khatib. 2010. A genome-wide association study using selective DNA pooling identifies candidate markers for fertility in Holstein cattle. *Anim. Genet.* 41:570-578.
- Jiang, L., J. Liu, D. Sun, P. Ma, X. Ding, Y. Yu, and Q. Zhang. 2010. Genome wide association studies for milk production traits in Chinese Holstein population. *PLoS One* 5(10):e13661.
- Khatkar, M. S., F. W. Nicholas, A. R. Collins, K. R. Zenger, J. A. Cavanagh, W. Barris, R. D. Schnabel, J. F. Taylor, and H. W. Raadsma. 2008. Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. *BMC Genomics* 9:187.
- Mangin, B., A. Siberchicot, S. Nicolas, A. Doligez, P. This, and C. Cierco-Ayrolles. 2012. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity* 108:285-291.
- Muir, W., G. Wong, Y. Zhang, J. Wang, M. Groenen, R. Crooijmans, H.-J. Megens, H. Zhang, J. McKay, and S. McLeod. 2008. Review of the initial validation and characterization of a 3K chicken SNP array. *World's Poult. Sci. J.* 64:219-226.
- Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, and M. J. Daly. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81: 559-575.
- Qanbari, S., E. Pimentel, J. Tetens, G. Thaller, P. Lichtner, A. Sharifi, and H. Simianer. 2010. The pattern of linkage disequilibrium in German Holstein cattle. *Anim. Genet.* 41: 346-356.
- Sahana, G., B. Gulbrandsen, C. Bendixen, and M. Lund. 2010. Genome-wide association mapping for female fertility traits in Danish and Swedish Holstein cattle. *Anim. Genet.* 41:579-588.
- Schaeffer, L. 2006. Strategy for applying genome-wide selection in dairy cattle. *J. Anim. Breed. Genet.* 123:218-223.
- Shifman, S., J. Kuypers, M. Kokoris, B. Yakir, and A. Darvasi. 2003. Linkage disequilibrium patterns of the human genome across populations. *Hum. Mol. Genet.* 12:771-776.
- Sved, J. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol.* 2:125-141.
- Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke, M. E. Goddard, and P. M. Visscher. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Res.* 17:520-526.
- Thevenon, S., G.-K. Dayo, S. Sylla, I. Sidibe, D. Berthier, H. Legros, D. Boichard, A. Eggen, and M. Gautier. 2007. The extent of linkage disequilibrium in a large cattle population of western Africa and its consequences for association studies. *Anim. Genet.* 38:277-286.
- Toosi, A., R. Fernando, and J. Dekkers. 2010. Genomic selection in admixed and crossbred populations. *J. Anim. Sci.* 88:32-46.
- 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.
- Wang, J. 2005. Estimation of effective population sizes from data on genetic markers. *Phil. Trans. R. Soc. B: Biol. Sci.* 360:1395-1409.
- Yu, J., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, and J. B. Holland. 2005. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38:203-208.
- Zhao, H., D. Nettleton, M. Soller, and J. Dekkers. 2005. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. *Genet. Res.* 86:77-87.