Application of SNP genotyping to understand dairy traits

Ester B. Flores[1*];DVM, MSc, PhD; Jesus Rommel V. Herrera[2,3] DVM, MBA

[1] *Philippine Carabao Center National Headquarters, Muñoz Nueva Ecija Philippines*
[2] *Philippine Carabao Center-University of the Philippines, College, Laguna 4031 Philippines*
[3] *School of Environmental and Rural Science, University of New England, Armidale, NSW, Australia 2350*
**Corresponding Author: esterflrs@yahoo.com*

Abstract

The use of conventional best linear unbiased prediction of estimated breeding values (BLUP EBVs) as a tool for selection in the dairy buffalo breeding program of the Philippines has resulted significant gains in increasing milk production potential of the dairy buffaloes. The availability of dense panel of single nucleotide polymorphism (SNP) markers for buffaloes provides an opportunity to further increase the rate of genetic gain through marker assisted selection (MAS) either by identifying markers significantly associated with milk production traits using genome wide association study (GWAS) or genomic selection (GS). However, methods depend on dense markers to cover the whole genome and for markers to be in linkage disequilibrium (LD) with quantitative trait loci (QTL). The size of the data set also affects the accuracy hence, the need to combine similar breeds into one reference population. As the local dairy buffalo population has not been subjected to either methods, there is a need to estimate the extent of LD in the population and population stratification. A total of nine hundred eighteen (918) buffalo DNA samples were submitted for genotyping. The Axiom analysis suite software was used to analyze the raw data for quality control metrics and generating genotype calls. For downstream analysis, principal component analysis (PCA) based on genomic relationship matrix (GRM) using R was done to visualize the population stratification among the local breeds. For LD estimation, 59 unrelated Bulgarian buffaloes were used. The extent of LD expressed as $\mathbf{r^2}$, was estimated using the R package "synbreed" with a gateway to the PLINK software. To visualize LD decay, the r^2 values were plotted against pairwise distance of increasing interval. The average r^2 for SNP pairs at ~44kb was 0.24. The longest interval with useful LD ($r^2 \geq 0.20$) is at 60-70kb. There is sufficient LD in the local dairy buffalo population for GWAS and GS studies

Key words: Dairy buffaloes, Genomic selection, PCA plot

Introduction

The Bulgarian Murrah Buffalo (BMB) is a riverine type of water buffalo breed developed in Bulgaria from crossing the Mediterranean breed with Murrah buffalo. The development of the breed started from an initial importation of Murrah buffaloes from India in 1962 (Borghese, 2005). The breed was developed by continuous backcrossing of the crossbred females to Murrah bulls for a further three generations before *interse* mating. Offspring of these animals were imported to the Philippines to become base animals in developing a local dairy breed. Subsequently, Brazilian Murrah and Italian Mediterranean buffaloes were also imported.

The breeding program implemented on these buffaloes has been effective in increasing the genetic potential of these animals due to the use of the conventional breeding value

estimation as a tool for selection. The breeding value estimation theory rests on the principle of the *infinitesimal model* that assumes traits are determined by an infinite number of unlinked and additive loci, each with an infinitesimally small effect (Fischer, 1918). There is also more recent evidence that there must be some *finite number of loci* underlying the variation in quantitative traits and that the distribution of the effect of these loci on quantitative traits is such that there are a few genes with large effect, and many of small effect (Ewing & Green, 2000). Advancement in molecular genetics enabled the identification of quantitative trait loci (QTL) affecting economically important traits using DNA markers. One method in identifying these loci is by QTL mapping. This approach utilizes DNA markers associated or linked with variation in quantitative traits. However, with few markers, the power of detection is limited to a few loci with large effects. The availability of dense panel of single nucleotide polymorphisms (SNP) markers increases the power of detections and exploits the linkage disequilibrium (LD) between marker and QTL. LD is the non-random association of alleles between two loci such that genome wide association studies have been done on various livestock species for a variety of traits. However, for most complex traits, large number of QTLs are necessary to explain a substantial proportion of genetic variation such that gains from marker assisted selection using this approach is likely to be small (Hayes & Goddard, 2010). A further development in animal breeding exploiting the existence of LD between causative variants and genetic markers is genomic selection (Meuwissen et al., 2001). The size of the reference population as well as the population stratification is important in accuracy of prediction. The density of markers should be sufficiently high to guarantee that all QTL are in LD with a marker. Measuring the extent of LD is important in determining how dense the markers need to be to be useful for LD mapping and MAS, including genomic selection.

LD is caused by migration, mutation, selection, small population size or other genetic events which the population experiences (Hayes & Daetwyler, 2015) although in livestock, small effective population size generating relatively large amount of LD is generally implicated as the key cause of LD. One measure of LD is **D**, however, **D,** as a measure is highly dependent on allele frequencies and is not suitable for comparing LD at different sites or multiple pairs of loci (Hayes & Daetwyler, 2015). A preferred measure is $\mathbf{r^2}$ which is less dependent on allele frequencies. The extent of LD and also the r^2 decreases as the distance between loci increases (Hayes & Goddard, 2010). In the Holstein cattle, it was reported that the average r^2 when loci are 50 kb apart is 0.35 (Goddard et al. 2006). Thus, to have loci evenly spaced at 50kb apart, 60,000 markers are needed. Cardoso et. al. (2014) estimated the LD in Brazilian buffalo population using 58,585 SNP panel and reported the general mean was 0.29 to $\mathbf{r^2}$ and 0.72 to **|D'|**. This indicated it is possible to use the SNP panel to calculate genomic values for the said population. The objective of this study is to visualize the population stratification of the local buffalo breeds and have a preliminary estimate of the extent of LD of a Bulgarian Murrah buffalo population in the Philippines using a high density SNP panel.

Materials and Methods

A total of nine hundred eighteen (918) buffalo DNA samples from 4 breeds were submitted for genotyping (Affymetrix, Inc., Sta. Clara, California). The Axiom analysis suite software was used to analyze the raw data (".CEL" files) for quality control metrics and generating genotype calls. The configuration used is the Best Practices Workflow mode. Moreover, some default settings for the sample and SNP QC used were: dish QC ≥ 0.82, QC_call_rate ≥ 97, plate QC_percent samples passed ≥ 95, plate_QC_average call rate ≥

98.5, species type is diploid and number of minor allele ≥ 2. Polymorphic SNPs identified by the Analysis Suite were further filtered. Only SNPs with MAF > 0.1, HW p-value > 0, no missing genotypes, and located in autosomes were retained for downstream analysis.

For downstream analysis, principal component analysis (PCA) based on genomic relationship matrix (GRM) using R was done to visualize the population stratification among the local breeds. For LD estimation, 59 unrelated Bulgarian buffaloes were used. No full-sibs and half-sibs were selected in order to maximize the genetic diversity within the sampled population.

LD (r^2) was estimated using the R package "synbreed" (Wimmer et al., 2012) with a gateway to the PLINK software (Purcell et al., 2007). SNP pairs were grouped according to their pairwise distance into 17 intervals ("bins"): 0-10 kb, 10–20 kb, 20–30 kb, 30–40 kb, 40–50 kb,50-60 kb, 70-80 kb, 80-90 kb, 90-100kb, 100–150 kb, 150-200 kb, 200–500 kb, 500 kb–1 Mb, 1–2 Mb, 1-3 Mb, 3–5 Mb and 5–10 Mb. The r^2 for a particular "bin" is the mean r^2 of all SNP pairs in that interval. To visualize LD decay, r^2 values were plotted against pairwise kb distance.

Results

The number of PolyHigh Resolution (PHR) SNPs for the three riverine populations, Bulgarian Murrah (BMB), Brazilian Murrah (BrMB), Italian Mediterranean (ItMB), are 67,810, 66,902 and 57,094, respectively. In the case of the Philippine swamp buffalo, the number of PHR SNPs is only 16,573. Most of its SNPs belong to the Mono High Resolution (37.6%) and the No minor homozygote (29.9%) categories. The high number of PHR SNPs identified in riverine breeds and the low number of PHR SNPs in the swamp(SP) population is due to the design of the SNP chip wherein only SNPs from riverine breeds were included. Among the three riverine breeds, the BMB and BrMB populations have higher PHR SNPs both populations have both Murrah and Mediterranean blood (Borghese, 2005); both of these breeds were included in the design of the SNP chip. Common PHR SNPs among the 3 riverine breeds were 46,445. If the PHR SNPs of the SP population are included, common PHR SNPs are only 10,443. The average inter-marker in autosomes in kb are 39.6, 39.9, 46.3 and 170.0 for the BMB, BrMB, ItMB and SP breeds, respectively (Table 1).

Table 1. Number of autosomal polymorphic SNPs in four water buffalo populations in the Philippines.

Population	Total no. of Polymorphic SNPs	Polymorphic SNPs with MAF≥ 0.01 and HW p-value ≥ 0.01	Autosomal Polymorphic SNPs	Average inter-marker distance in autosomes (kb)
Bulgarian Murrah	67,810	64,561 (71.7)	63,252 (70.3)	39.6
Brazilian Murrah	66,902	65,298 (72.6)	62,709 (69.7)	39.9
Italian Mediterranean	57,094	56,106 (62.3)	53,917 (59.9)	46.3
Philippine Swamp	16,573	14,920 (16.6)	14,578 (16.2)	170

MAF - minor allele frequency, Hardy-Weinberg

Figure 1 shows a Principal Component Analysis (PCA) plot generated based on GRM using only the PHR SNPs of the 4 Philippine buffalo breeds. The first principal component (PC1), with a variance of 89.5%, splits the data into the riverine type on the left side and the swamp type (SP1) on the right. The variation due to PC2 is only 2.45%. BMB and BrMB

samples overlapped one another, since the 2 breeds have both Murrah and Mediterranean blood.

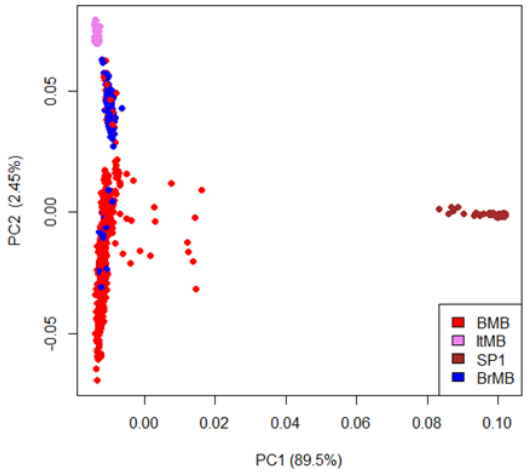
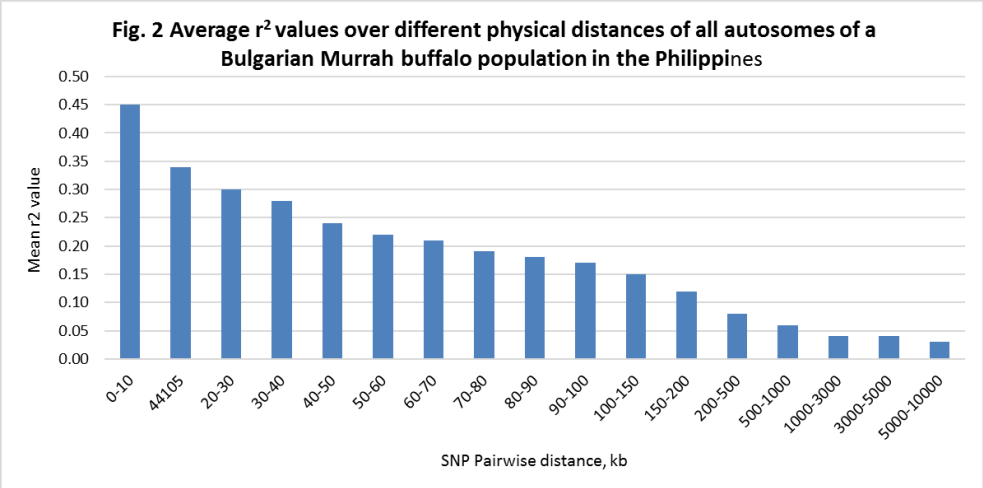


Figure 1. PCA plot among animals belonging to 3 riverine buffalo populations and 1 swamp buffalo population (n=902). BMB =Bulgarian Murrah, BrmB= Brazilian Murrah, ItMB= Italian Meditteranean, SP1= Philippine Swamp. PCA plot generated using R scripts.

For LD estimation, of the data set used, only 57,325 polymorphic SNPs passed the above filtering criteria and were included in the final analysis. This number of SNPs is very similar to what Cardoso et al (2014) found i.e., 58,585 SNPs, that were used for the study of LD in Brazilian milk buffaloes. These 57,325 SNPs cover 2.5015 Mb of the genome, the average inter-marker interval is 43.6 kb, the largest gap between SNPs (2,556,222 bp) is located in chromosome 12 and the smallest gap (8 bp) is located in chromosome 15. Nevertheless, the LD decay shown in Figure 2 shows the average r^2 for SNP pairs at ~44kb distance was at least 0.24. To obtain an average spacing of 50 kb requires 51,124.44 evenly spaced markers. The number of SNP markers that passed the criteria and was used in this study is greater than that number. The longest interval with useful LD (r2 ≥0.20) is at 60-70kb. This is similar to what Cardoso et al. (2014) found in Brazilian milk buffaloes. The similarity in the results obtained for the Brazilian and Bulgarian buffalo population is not surprising considering both breeds have Indian Murrah and Mediterranean blood. However, this is lower than what was reported in dairy cattle but this could be due to difference in effective population size. The extensive use of artificial insemination in dairy cattle compared to the riverine buffalo population is probably a reason for the former to have a smaller effective population and generates a relatively larger LD.



Conclusions

With the LD level estimated for markers separated by less than 70kb being $r^2 \geq 0.20$ and the average inter-marker distance of the 57,325 SNPs being 43.6kb, there is sufficient evidence to show that the extent of LD in the dairy buffalo population may be suitable for marker assisted selection. The Axiom 90k Buffalo Genotyping Array is a suitable tool for GWAS and GS studies.

Funding acknowledgements

This research is funded by the Philippine Council for Agriculture, Aquatic, and Natural Resources Research and Development- Department of Science and Technology (PCAARRD-DOST).

References

Borghese A. 2005. Buffalo Production and Research. FAO Regional Office for Europe Inter-Regional Cooperative Research Network on Buffalo (ESCORENA). http://www.fao.org/3/a-ah847e.pdf

Cardoso D F, Aspilcueta-Borquis R R, Santos D J A, Hurtado-Lugo N A, de Camargo, G. M. F. Scalez, D. C. B. de Albuquerque, L. G. Tonhati H. 2010. Study of Linkage Disequilibrium in Brazilian Dairy Buffaloes Proceedings, 10th World Congress of Genetics Applied to Livestock Production. Vancouver, BC, Canada. Aug. 17-22, 2014.

Ewing B, Green P. 2000. Analysis of expressed sequence tags indicates 35,000 human genes. Nat Genet 25, 232-4.

Fischer R A. 1918 The correlation between relatives: the supposition of Mendelian inheritance. Trans Royal Soc Edin 52, 399.

Goddard M E, Hayes B, McPartlan H, Chamberlain A J. 2006. Can the same genetic markers be used in multiple breeds? 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Brazil, 8 pp.22–14.

Hayes B, Daetwyler H. 2015. Course note: Genomic Selection. Armidale Animal Breeding Summer Course. February 2015. Armidale New South Wales, Australia

Hayes B. Goddard M. 2010. Genome-wide association and genomic selection in animal breeding. Genome. 53 (11): 876-883

Hayes B J, Visscher P M, McPartlan H C,Goddard M E. 2000. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res*. 13,635–643.

Hill W G. 1981 Estimation of effective population size from data on linkage disequilibrium. Genetics Research 38, 209-16.

Meuwissen T H, Hayes B J, Goddard M E. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157(4):1819–29.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81.

R: A Language and Environment for Statistical Computing. http://www.Rproject.org.

Wimmer V, Albrecht T, Auinger H J, Schoen C C. 2012. synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics, 28: 2086-2087