GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Water buffalo genome characterization by the Illumina BovineHD BeadChip

R.R.A. Borquis[1], F. Baldi[1], G.M.F. de Camargo[1], D.F. Cardoso[1], D.J.A. Santos[1], N.H. Lugo[1], M. Sargolzaei[2], F.S. Schenkel[2], L.G. Albuquerque[1,3] and H. Tonhati[1,3]

[1]Faculdade de Ciências Agrárias e Veterinárias,
Universidade Estadual Paulista, Jaboticabal, SP, Brasil
[2]Centre for Genetic Improvement of Livestock,
Animal and Poultry Science Department, University of Guelph,
Guelph, Ontario, Canada
[3]Instituto Nacional de Ciência e Tecnologia de Ciência Animal,
Universidade Federal de Viçosa, Viçosa, MG, Brasil

Corresponding author: H. Tonhati
E-mail: tonhati@fcav.unesp.br

Genet. Mol. Res. 13 (2): 4202-4215 (2014)
Received April 16, 2013
Accepted October 29, 2013
Published June 9, 2014
DOI http://dx.doi.org/10.4238/2014.June.9.6

ABSTRACT. To define the best strategies for genomic association studies and genomic selection, it is necessary to determine the extent of linkage disequilibrium (LD) and the genetic structure of the study population. The current study evaluated the transference of genomic information contained in the Illumina BovineHD BeadChip from cattle to buffaloes, and assessed the extent of the LD in buffaloes. Of the 688,593 bovine single nucleotide polymorphism (SNP) that were successfully genotyped from the 384 buffalo samples, only 16,580 markers were polymorphic, and had minor allele frequencies greater than 0.05. A total of 16,580 polymorphic SNPs were identified, which were uniformly distributed throughout the autosomes, because the density and mean distance between markers were similar for all autosomes. The average minor allele frequency for the 16,580 SNPs was 0.23. The overall mean LD for pairs of adjacent markers was 0.29 and 0.71, when measured as for r^2 and |D'|, respectively. The 16,580 polymorphic SNPs were matched to *Bos taurus* chromosome in the current bovine genome assembly (Btau 4.2), and could be utilized in association studies. In conclusion, the Illumina BovineHD BeadChip contains approximately 16,580 polymorphic markers for the water buffalo, which are broadly distributed across the genome. These data could be used in genomic association and genomic selection studies; however, it might be necessary to develop a panel with specific SNP markers for water buffaloes.

Key words: Buffalo; Genetic marker; Linkage disequilibrium;
Illumina BovineHD BeadChip

INTRODUCTION

Water buffaloes (*Bubalus bubalis*) are important farm livestock in many countries of the world, particularly to smallholders. Buffaloes are used as draft animals, in addition to producing meat, horns, leather, and, especially, milk. Buffalo milk contains higher fat, lactose, and protein concentrations compared to cow milk, facilitating the production of high-quality dairy products, such as butter and cheese, among others. Buffalo meat is noted for its tenderness and taste, in addition to its low fat and cholesterol content. In addition, the hides are used to produce high-end leather goods (Roth and Myers, 2004).

Cattle (*Bos taurus*) and water buffalo (*B. bubalis*) belong to the subfamily Bovinae. Cattle were domesticated between 8000 and 10,000 years ago (Bradley and Cunningham, 1998), while buffalo were domesticated at least 7000 years ago (Bibi and Vrba, 2010). Bovines have a large number of single nucleotide polymorphisms (SNPs) in their genome. The current SNP database (dbSNP) of the National Center for Biotechnology Information (NCBI) contains over 9.5 million bovine SNPs; yet, very few buffalo SNPs have been identified (Amaral et al., 2008; Michelizzi et al., 2010, 2011). As buffalo and cattle are closely related, the large quantity of genetic/genomic resources developed for cattle might help with characterizing the buffalo genome, and in the genomic selection of these animals.

Recently developed techniques allowing the breeding genetic value might be obtained from genomic data by marker-assisted selection covering the whole genome, also called genomic selection (Bennewitz et al., 2009; Calus et al., 2009). This type of genomic selection exploits the linkage disequilibrium (LD) between markers; whereby, it is assumed that the effects of the chromosome segments are the same in the entire population, because the markers are in LD with the genes responsible for the expression of the trait of interest (QTL). Therefore, the density of the markers must be high enough to guarantee that all the QTL are in LD with at least a single marker or haplotype of markers. The LD maps provide a basic tool for investigating the genetic base of economically important traits (McKay et al., 2007).

LD between markers has been widely studied in the genome of various species of domesticated animals (McKay et al., 2007; Villa-Angulo et al., 2009; Nagarajan et al., 2009; Bohmanova et al., 2010; Silva et al., 2010). The first step necessary to determine the number of markers required for QTL mapping and genomic selection is the quantification of LD extent in the genome of the target species. It has been suggested (Meuwissen et al., 2001; Calus et al., 2009) that a level of LD (r^2) higher than 0.2 is necessary to achieve a genomic breeding value accuracy approaching 0.8.

This study assessed the possibility of transferring genomic information obtained from the Illumina BovineHD BeadChip (Illumina Inc., San Diego, CA, USA) from cattle to buffalo. In addition, we characterized the bubaline genome with respect to the allele frequencies of the markers, the number of polymorphic markers, the distribution of markers along the genome and the LD between markers.

MATERIAL AND METHODS

DNA samples and extraction

The DNA samples were collected from the hair follicles of 384 female water buffaloes (*B. bubalis*) that were born in 2007 and 2008. The buffaloes belonged to 2 important dairy farms in the states of Rio Grande do Norte and São Paulo, Brazil. Pedigree information was available for all animals, with the 384 females being the daughters of 16 sires. The hair follicles were placed in separate envelopes, which were labelled and stored at 4°C until DNA extraction. The laboratory analyses were carried out at the Molecular Genetics Laboratory of Faculdade de Ciências Agrárias e Veterinárias of Universidade Estadual Paulista (UNESP), Jaboticabal Campus, Brazil. The DNA was extracted from the hair follicles by the phenol-chloroform-isoamyl alcohol (PCI) technique. About 40 follicles/animal were placed in a microcentrifuge tube (1.5 mL) and rapidly spun. Then, 500 μL TE-Tween solution (50 mM Tris, 1 mM EDTA, 0.5% Tween 20) was added to each sample, following incubation in a water bath at 65°C for 1.5 h, with periodic agitation. After this period, 2 μL proteinase K/tube (600 μg/μL) was added, and the tubes were incubated at 55°C for a further 6 h, with periodic agitation. Finally, the tubes were incubated at 37°C overnight.

After these procedures, 1 volume of PCI was added to 1 volume of each sample. The tubes were then shaken vigorously for 10 s in an automatic shaker, after which they were centrifuged for 10 min at 12,000 rpm and 23°C, and the supernatant was transferred to a new tube.

The final volume of this phase was approximately 300 μL. Next, the DNA was precipitated with 1/10 sample volume 0.3 M sodium acetate (approximately 30 μL) and ice-cold absolute ethanol (approximately 1 mL). After mixture by inversion, the tubes were placed in a freezer at -80°C for 1 h. The tubes were centrifuged at 4°C for 25 min at 12,000 rpm. The supernatant was discarded, and the remaining DNA was completely dried at room temperature, and then stored in 100 μL Tris HCl EDTA (10:1).

The quantity and quality of the DNA obtained were analyzed in a Nanodrop 1000 spectrophotometer (Thermo Scientific Wilmington, DE, USA, 2008). The DNA is quantified from the level of absorbance, with DNA having a light absorbance peak at a wavelength of 260 nm. Hence, the concentration is measured by the relationship: 1 OD_{260} = 50 μg/mL DNA.

Genotyping and quality control of the genomic data

Genotyping was performed by the Illumina BovineHD BeadChip, utilizing the Infinium® HD assay kit and the Illumina HiScan™ system (Illumina Inc.). The BovineHD BeadChip has 777,962 SNP markers spread through the genome, and an average distance between markers of 3.43 kb. Although the BovineHD BeadChip was developed for bovine, water buffalo chromosomes and *B. taurus* chromosomes are highly homologous (Figure 1). The initial analyses of the images and the genotypes were carried out by the Genome Studio software. A total of 1735 markers were excluded, because their genomic position was not known. Only markers with a call frequency greater than 80% and with heterozygote excess (Het Excess) smaller than -0.70 or greater than 0.70 were considered. The markers that showed low average cluster intensity (AB_R, AA_R or BB_R mean < 0.1; AB_T_mean < 0.2, and AB_T_mean > 0.8), with a GenTrain score <0.30 and cluster separation index <0.13, were excluded from the analysis. The filtering of these criteria led to elimination of 89,369 markers. Similar criteria for filtering genomic data were implemented by Michelizzi et al. (2011), using Illumina BovineSNP50 BeadChip on DNA samples from 10 water buffaloes. In the present study, we only included markers in autosomal chromosomes with minor allele frequency (MAF) greater than 0.05 in the analyses.

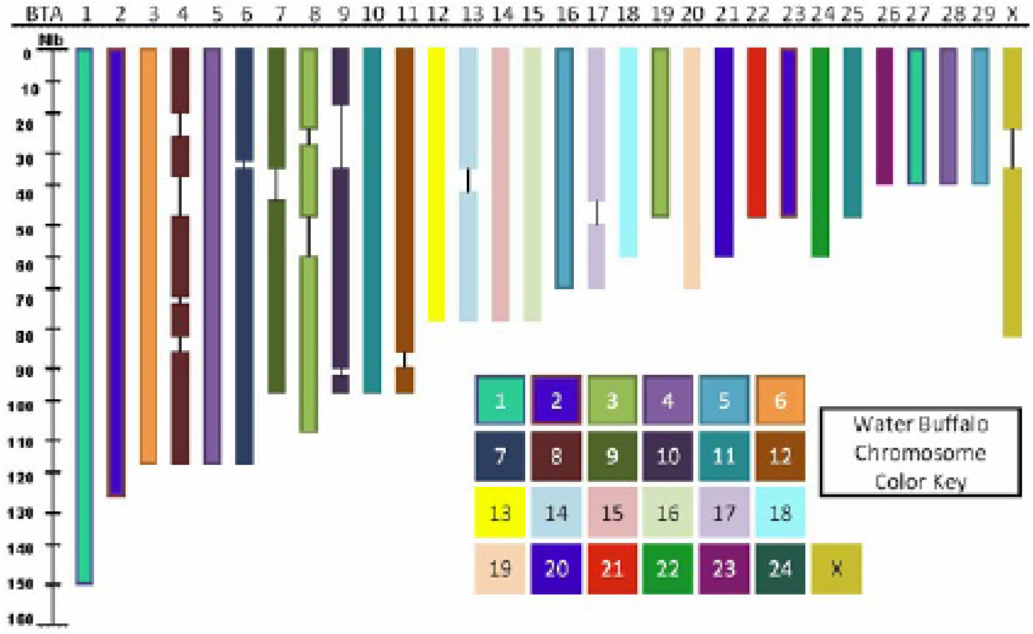
BTA 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 X
Mb
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
1 2 3 4 5 6
7 8 9 10 11 12
13 14 15 16 17 18
19 20 21 22 23 24 X
Water Buffalo Chromosome Color Key

Figure 1. Homology of water buffalo chromosomes and *Bos taurus* chromosomes.

LD between the markers

The 2 most common measures used to assess the LD for 2-allele markers are r^2 and |D'| (Hill and Robertson 1966; Valdar et al., 2006; Bohmanova et al., 2010; Hayes and Goddard, 2010). These parameters vary between 0 and 1. |D'| <1 indicates recombination that has occurred between 2 loci, while |D'| = 1 indicates a lack of recombination between the 2 loci. The measure r^2 represents the correlation between 2 loci, and is preferentially used in association studies because there is an inverse relationship between r^2 and the sample size necessary to detect the association between a QTL and a marker. In addition, this measure is much less biased by low allele frequencies compared to |D'| (Pritchard and Przeworski, 2001; Sargolzaei et al., 2008). The LD between 2 SNPs was evaluated using r2 (Hill and Robertson, 1968) and the absolute value of |D'| Lewontin (1964), which was calculated as follows:

$$r^2 = \frac{(D)^2}{(freq.A * freq.a * freq.B * freq.b)}$$

where:

$$D = freq.AB - freq.A * freq.B$$

and:

$$D' = \begin{Bmatrix} \dfrac{D}{\min(freq.A * freq.b, freq.a * freq.B)} & if \quad D < 0 \\ \dfrac{D}{\min(freq.A * freq.B, freq.a * freq.b)} & if \quad D \geq 0 \end{Bmatrix}$$

where *freq.A*, *freq.a*, *freq.B*, and *freq.b* are the frequencies of alleles A, a, B, and b, respectively, and *freq.AB*, *freq.ab*, *freq.Ab*, and *freq.aB* are the frequency of haplotype AB, ab, Ab, and aB in the population. If the 2 loci are independent, the expected frequency of haplotype AB (freq.AB) is calculated as the product between freq.A and freq.B. A value of freq.AB higher or lower than the expected value indicates that these 2 loci in particular tend to assimilate together and in LD. The measures of LD (r2 and |D'|) were calculated for all marker pairs of each chromosome using the SnppldHD software (Sargolzaei M, University of Guelph, Canada). Only maternal haplotypes were considered for the estimation of LD measures (r2 and |D'|). The exclusive use of maternal haplotypes is a common practice in studies estimating LD when the population consists of half-sib families, as was the case here. This is because the pedigree structure leads to the over-representation of paternal haplotypes in the sample, as sires have multiple progenies in the dataset, which might increase the frequency of certain haplotypes, and consequently cause LD to be overestimated (Bohmanova et al., 2010).

Bovine reference genome assembly (Btau_4.2)

The Bovine Genome Sequencing and Analysis Consortium performed genome sequencing and assembly for cattle. The current assembly (Btau_4.2) combined both BAC and whole-genome shotgun sequences, which were then placed on chromosomes by employing different mapping methods. The Consortium estimated that the bovine genome size is ~2.87 Gbp, and that the current assembly covers at least 92% of the genome. Since the SNP map locations are based on the Btau_4.2 assembly, the gene information from the Cow Genome Resources of the NCBI (Bovine Genome Resources) was downloaded and used. This information includes gene symbol, start position, stop position, orientation on the chromosome, and gene description.

SNP density estimation along each chromosome

To estimate the SNP density along the bovine chromosome, we used the nonparametric kernel density method, whereby $x_1, x_2, ..., x_n \sim f$, is considered to be an independent and identically distributed sample of a random variable X, where x_i is the observable location (in bp) of the i^th SNP marker. Therefore, the kernel density estimator is:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \frac{1}{nh}\sum_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}} e^{\frac{(x-x_i)^2}{2h^2}}\right)$$

where h > 0 is a bandwidth smoothing parameter (BW) and *K* is a kernel, considered in the present analysis as a standard Gaussian function with a mean of 0 and a variance of 1.

In the example above, *h* is a free parameter that has a strong influence on the kernel estimates resulting from the density (graphs). The kernel density graphs show similar patterns, with BW = 1 M or less. Therefore, we decided to present the results obtained with BW = 0.05 M.

We plotted the call rate graphs function of x_i to obtain a position on the map referring to SNP i, where y_i is a call rate measure for *i* = 1, 2, ..*n*. We applied the LOWESS (locally weighted scatterplot smoothing) technique to describe the relationship between x_i and y_i, as follows:

$$y_i = g(x_i) + \varepsilon_i$$

where g is a smoothing function and ε_i is a random variable with a mean of 0 and a constant scale. According to Michelizzi et al. (2011), the LOWESS is a non-parametric curve- or function-fitting technique (Cleveland, 1979, 1981), in which the fitting at point x is carried out using only the points in the neighborhood of x. Thus, this method has more detailed assumptions about the form of the relationship in comparison with parametric methods, and allows the relationship to be described closer to its true form, as demonstrated by the data (Michelizzi et al., 2011).

In this analysis, we started LOWESS with a local polynomial (a fit of the *k* - *NN* type) adjusted by least squares, and then used robust methods to obtain the final fit. First a polynomial regression was fitted in a neighborhood of x. This is equivalent to finding $\beta \in R^{P+1}$, which minimizes:

$$\frac{1}{n}\sum W_{ki}(x)\left(y_i - \sum_{j=0}^{p}\beta_j x_j\right)^2,$$

where $W_{ki}(x)$ denotes *k* - *NN* weights. Afterwards, the residuals $\hat{\varepsilon}_i$ and the scale of the parameter $\hat{\sigma} = median(\hat{\varepsilon}_i)$ were computed, and the robustness of the weights was defined as:

$$\delta_i = K(\hat{\varepsilon}_i / 6\hat{\sigma})$$

where:

$$K(z) = \begin{cases} \frac{15}{16}(1-z)^2, & \text{if } |z| \leq 1 \\ \text{otherwise} & 0 \end{cases}$$

Finally, the regression analysis fits the polynomial in the first equation, but with weights $\delta_i W_{ki}(x)$.

A notable characteristic of the procedure above is that it is not necessary to define a global function of any particular form to fit the model to the data. The fit is performed locally using only a segment of the data. Mathematically, "locally" is defined as within the distance represented by the largest integer not greater than $f\,x\,n$, where *f* is the smoothest extension (Michelizzi et al., 2011). The value of *f* gives the proportion of points on the graph that influence the smoothing for each value. In general, higher values of *f* means a smoother curve. Therefore, a good choice of *f* is the largest possible value that minimizes the variability of the smoothed points without distorting the pattern of the data. For the pattern in this study, *f* = 0.001 was empirically chosen, as in Michelizzi et al. (2011).

RESULTS AND DISCUSSION

The call rate for all samples, which indicates the quality of the genotyping, ranged from 54 to 90% (mean = 85%). Usually, with bovine samples, call rates of greater than 98% are expected. The low call rate obtained in the present study probably indicates the presence of differences in the hybridization of the samples from some animals with the primers, leading to possible problems in amplifying the markers. Among the 688,593 bovine SNPs that were successfully genotyped, only 16,580 markers were polymorphic and presented a minor allele frequency greater than 0.05, allowing their use in association studies and genomic selection. After filtering the data, the number of markers with fixed alleles was 26,042 SNPs. There were a large number of markers with a minor allele frequency smaller than 0.05 (645,971 SNPs), not including the fixed alleles.

The 16,580 SNPs were uniformly distributed among the autosomes, because the density and mean distance between markers was similar for all of the autosomes (Table 1). After filtering the SNP data, a considerable proportion of the SNPs had MAF values of below 0.20 (Figure 2). Similar results were reported by Michelizzi et al. (2011), who found 926 polymorphic SNP markers in water buffaloes using the Illumina BovineSNP50 BeadChip. The authors also reported that 386 markers of the total 926 SNPs had allele frequencies of lower than 0.10, and that these markers are probably too rare for use in buffalo studies.

Table 1. Summary of the polymorphic SNP markers analyzed for each autosome (BTA).

BTA	Size (Mb)	SNPs (N)	Mean MAF
1	158.11	1543	0.14 ± 0.13
2	136.30	1216	0.15 ± 0.13
3	120.67	1051	0.16 ± 0.13
4	120.44	1126	0.16 ± 0.13
5	120.267	949	0.16 ± 0.13
6	118.99	1195	0.14 ± 0.13
7	111.26	982	0.16 ± 0.14
8	112.70	969	0.16 ± 0.13
9	104.89	1016	0.15 ± 0.14
10	102.99	909	0.16 ± 0.14
11	106.93	961	0.16 ± 0.13
12	90.65	878	0.15 ± 0.13
13	84.09	639	0.17 ± 0.13
14	83.13	773	0.15 ± 0.14
15	84.81	748	0.17 ± 0.13
16	81.14	736	0.16 ± 0.13
17	74.99	656	0.15 ± 0.13
18	65.25	541	0.18 ± 0.13
19	62.90	472	0.17 ± 0.13
20	71.54	717	0.16 ± 0.13
21	68.87	661	0.16 ± 0.13
22	61.05	512	0.19 ± 0.14
23	51.32	446	0.18 ± 0.13
24	62.05	568	0.17 ± 0.14
25	42.42	408	0.19 ± 0.14
26	50.44	453	0.15 ± 0.13
27	45.11	452	0.17 ± 0.13
28	46.13	441	0.16 ± 0.14
29	49.94	463	0.15 ± 0.13

MAF = minor allele frequency.

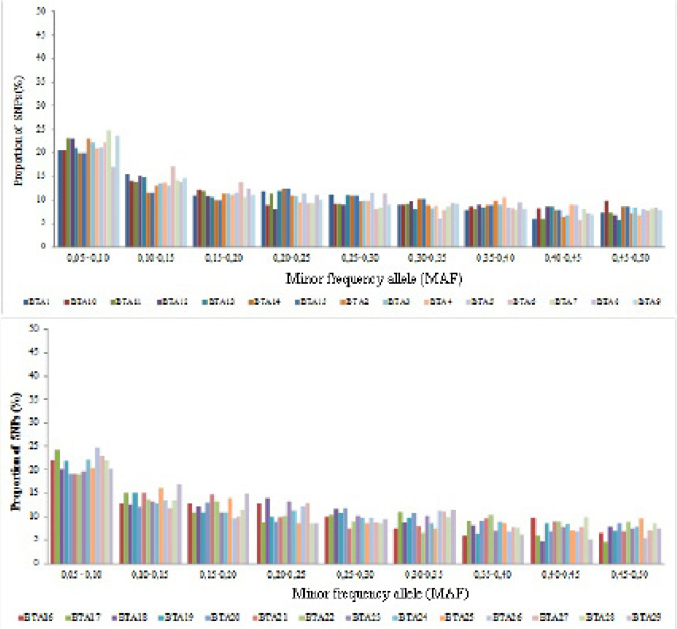

Figure 2. Mean proportions of single nucleotide polymorphism (SNPs) for various minor allele frequencies (MAF) by chromosome (intervals do not included the upper limit).

The mean MAF value observed in this study (0.23) was slightly higher compared to that reported by Matukumalli et al. (2009) for Nelore cattle (0.19) using the Illumina BovineSNP50K BeadChip and by the Bovine Hapmap Consortium for Nelore cattle (0.20) (Gibbs et al., 2009). According to information contained in the datasheet of the BovineHD BeadChip (www.illumina.com/Documents/products/datasheets/datasheet_bovineHD.pdf) of 10 samples from outgroup species (including the water buffalo *B. bubalis*, yak *Bos grunniens*, and guar *Bos gaurus*), more than 167,000 polymorphic SNPs were obtained, with an average MAF lower than that reported in the present study (0.06). Khatkar et al. (2008) showed that the MAF limit affects the distribution and extension of the LD, with this finding being important for association studies and for genomic selection.

The overall mean LD values for the marker pairs were 0.28 and 0.70 for r^2 and |D'|, respectively. The extent of LD between markers determines the number of markers (density of markers) required to apply genome-wide association studies and genomic selection (Hayes and Goddard, 2010). Calus et al. (2007) demonstrated that when the mean r^2 between adjacent SNPs was >0.2, accurate genomic breeding values could be predicted. Therefore, the average LD (r^2) between adjacent markers in the buffalo genome is higher than the minimum value that is required to implement genomic selection.

The LD means of the autosomes ranged from 0.07 to 0.34 for r^2 and from 0.30 to 0.80 for |D'| (Table 2). For Gyr cattle, Silva et al. (2010) found a smaller range of LD values, which ranged from 0.17 to 0.24 for r^2 and from 0.60 to 0.72 for |D'|. More recently, Espigolan et al. (2012) evaluated the extent of the LD in 48 Nelore cattle by high-density SNP marker panels, and obtained general mean LD values between marker pairs of 0.18 and 0.55 for r^2 and |D'|, respectively.

In the present study, we estimated lower LD levels on chromosomes BTA1 and BTA29. Arias et al. (2009) recorded high variability in the recombination rates of autosomes, which, among other factors, leads to considerable diversity in the LD pattern of different genome regions. However, the results obtained for BTA1 and BTA29 in the present study probably arose as a consequence of a sampling problem, because the number of markers, their density, and the mean MAF and MAF proportion were not different compared to the other autosomes studied.

Few studies have reported LD estimates between genetic markers in buffaloes. In an analysis of LD among 525 microsatellite loci, Nagarajan et al. (2009) reported that a larger number of marker pairs exhibited significant LD for buffalo markers compared to bovine cattle markers. The higher LD levels obtained in the present study compared to other studies investigating cattle (taurine and zebuine) may be explained by the history of the introduction of buffaloes and the evolution of buffalo herds in Brazil. According to Hayes and Goddard (2010), the extent of genome-wide LD is largely determined by the effective population size, both past and present. The most recent entry of buffaloes in Brazil occurred in 1962, when 5 males and 12 females were imported (Santana et al., 2011). Therefore, the national herd of today originates from just a small number of animals, indicating that the effective population size might be small and the level of inbreeding high (Santana et al., 2011). There is an inverse relationship between the effective population size and its LD level (Hayes et al., 2003). In a study of dairy buffalo in Brazil, Santana et al. (2011) recorded a very small population effective size (40.10 ± 1.27).

The 16,580 polymorphic SNPs, which might be utilized in association studies, were matched to *Bos taurus* chromosomes (BTAs) in the current bovine genome assembly (Btau 4.2). The count of known bovine genes in each BTA ranged from 304 genes on BTA27 to 1463 genes on BTA3. In contrast, the number of polymorphic SNPs ranged from 304 on BTA27 to 992 on

BTA 3 (Figures 3 and 4). There are potentially more genes than polymorphic SNP markers on several BTAs (Figure 4). As a consequence, not all genes potentially have SNPs in LD (Figure 4).

Table 2. Linkage disequilibrium (r^2 and |D'|) between adjacent synthetic SNP for each autosome (BTA).

BTA	Mean r^2 ± SD[1]	Mean |D'| ± SD
1	0.15 ± 0.28	0.46 ± 0.37
2	0.30 ± 0.36	0.71 ± 0.33
3	0.29 ± 0.34	0.71 ± 0.33
4	0.28 ± 0.34	0.73 ± 0.33
5	0.28 ± 0.35	0.71 ± 0.82
6	0.32 ± 0.35	0.73 ± 0.32
7	0.31 ± 0.35	0.74 ± 0.31
8	0.27 ± 0.33	0.70 ± 0.33
9	0.31 ± 0.35	0.73 ± 0.32
10	0.28 ± 0.34	0.72 ± 0.32
11	0.32 ± 0.34	0.76 ± 0.30
12	0.30 ± 0.34	0.75 ± 0.28
13	0.34 ± 0.33	0.75 ± 0.31
14	0.28 ± 0.33	0.69 ± 0.35
15	0.30 ± 0.34	0.71 ± 0.34
16	0.32 ± 0.35	0.77 ± 0.29
17	0.30 ± 0.36	0.74 ± 0.31
18	0.31 ± 0.34	0.75 ± 0.30
19	0.29 ± 0.32	0.74 ± 0.32
20	0.30 ± 0.34	0.72 ± 0.33
21	0.33 ± 0.34	0.74 ± 0.34
22	0.32 ± 0.35	0.77 ± 0.301
23	0.23 ± 0.31	0.68 ± 0.32
24	0.29 ± 0.33	0.76 ± 0.30
25	0.25 ± 0.30	0.74 ± 0.32
26	0.31 ± 0.34	0.80 ± 0.28
27	0.32 ± 0.36	0.71 ± 0.32
28	0.24 ± 0.33	0.74 ± 0.31
29	0.07 ± 0.18	0.30 ± 0.30

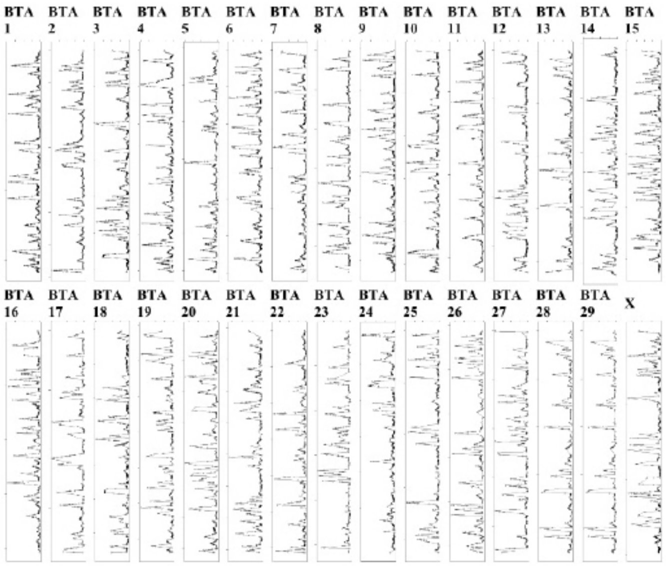

Figure 3. Call frequency plots of bovine SNPs in water buffalo samples. The chromosome size can be seen in Figure 5. The six-scale marks on the top of each chromosome represent 0 (left most), 25, 50, 75, and 100% (right most) of call frequency.

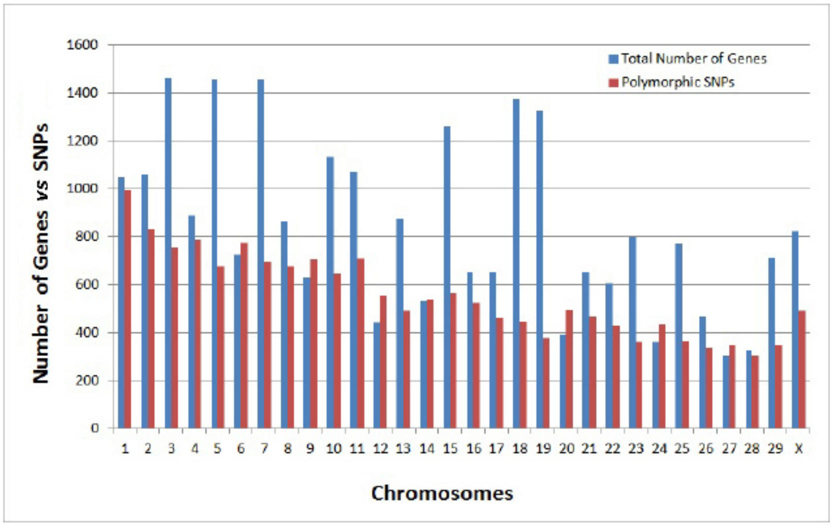

Figure 4. Distribution of polymorphic SNPs in the water buffalo and of known genes on each bovine chromosome based on NCBI gene information.

This result differs to those reported by Michelizzi et al. (2011), mainly because the authors used all of the SNPs of the BovineSNP50 BeadChip, rather than just the polymorphic SNPs. The analyses indicate that 5471 (33%) of the polymorphic SNPs are intragenic (located within genes) and that 11,109 (67%) are intergenic (located between genes) (Figure 5). Therefore, more than 13,000 genes in the bovine genome do not have intragenic SNPs, while the remaining genes have between 1 and 42 intragenic SNPs, which are polymorphic in buffalo. In comparison, each intergenic region is covered by at least one SNP.

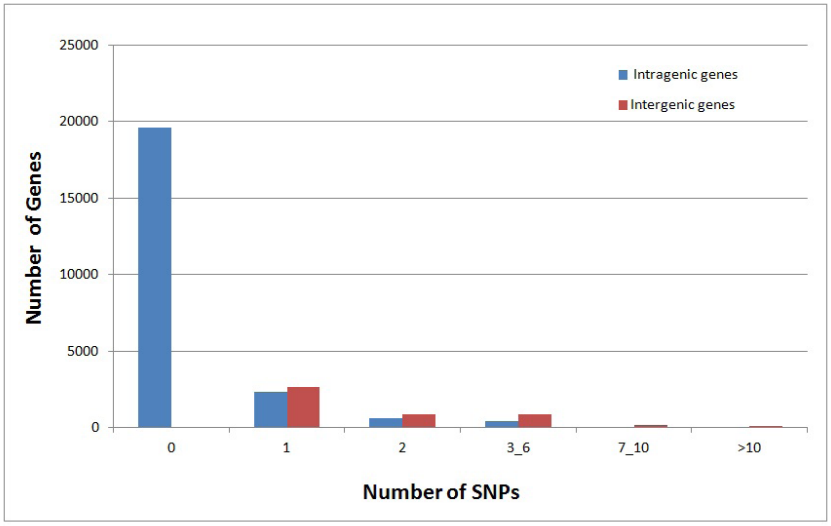

Figure 5. Distribution of known genes and the number of intragenic and intergenic SNPs in the BovineHD BeadChip, which are polymorphic in water buffalo.

The SNP density graphs for each chromosome, assuming BW = 0.05 M (50 kb), are presented in Figure 6. All of the chromosomes had regions that were rich in SNP coverage along them, except for chromosome X. This result indicates that the polymorphic markers found in the Illumina BovineHD BeadChip are well distributed in all autosomes, enabling this chip to be used for buffalo studies.

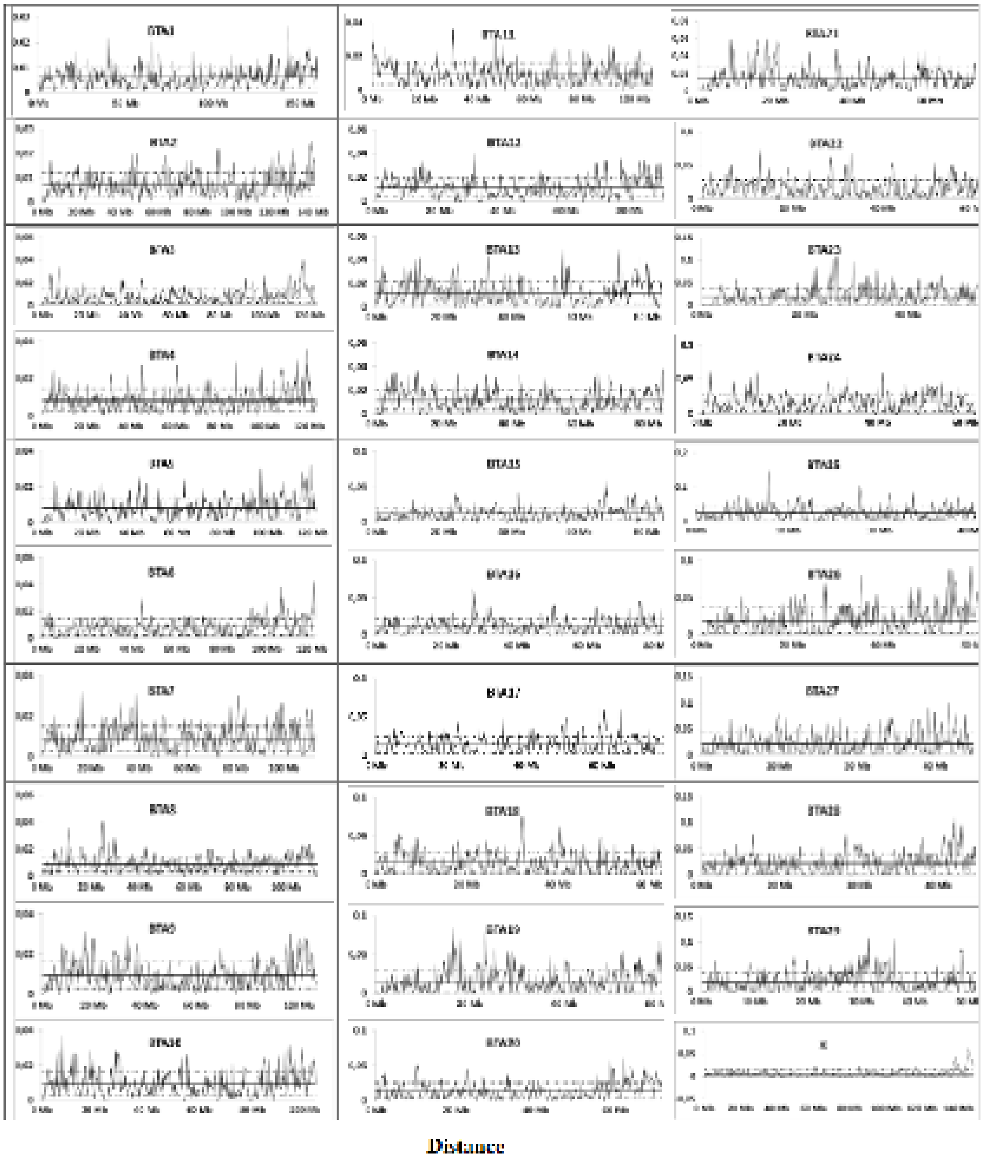

Figure 6. Kernel density plots for SNPs on the Illumina BovineHD BeadChip, which are polymorphic in water buffalo by chromosome. Dotted lines indicate 1 SD above and below the chromosome mean.

The number of polymorphic SNPs in the buffalo genes found in the BovineHD BeadChip and their distribution on the chromosomes might provide information about conserved region of the genome. Such information could be used to explain evolutionary processes and provide insights about species biodiversity, phylogeny, and adaptation to environmental changes. Therefore, the results of this study allow water buffalo genes with potential economic importance to be identified based on the known functions of these genes in cattle. Many of these SNPs have been investigated in cattle, and provide information about the potential functions of SNPs that are transferable to buffalo. Knowledge about these SNPs could potentially be applied to genetic improvement programs with the aim to increase the productivity and reproductive efficiency of buffalo, and to select animals that are more resistant to certain diseases. The results also indicate that strategies for mapping and genomic selection based on LD might be effective for the genetic improvement of water buffalo. However, the development of a panel with specific SNP markers for water buffalo might be required to determine the exact position of the SNPs.

CONCLUSION

This study showed that the BovineHD BeadChip has approximately 16,580 polymorphic markers that are evenly distributed on the Murrah buffalo genome, and could be used in association and genomic selection studies.

The level of linkage disequilibrium between SNP markers in this study indicated that the BovineHD BeadChip could potentially be used in association and genomic selection studies of buffalo; however, it might be necessary to develop a panel with specific SNP markers for water buffaloes.

ACKNOWLEDGMENTS

Research supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, São Paulo, SP, Brazil) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brasília, DF, Brazil).

REFERENCES

Amaral ME, Grant JR, Riggs PK, Stafuzza NB, et al. (2008). A first generation whole genome RH map of the river buffalo with comparison to domestic cattle. *BMC Genomics* 9: 631.

Arias JA, Keehan M, Fisher P, Coppieters W, et al. (2009). A high density linkage map of the bovine genome. *BMC Genet.* 10: 18.

Bennewitz J, Solberg T and Meuwissen T (2009). Genomic breeding value estimation using nonparametric additive regression models. *Genet. Sel. Evol.* 41: 20.

Bibi F and Vrba ES (2010). Unraveling bovin phylogeny: accomplishments and challenges. *BMC Biol.* 8: 50.

Bohmanova J, Sargolzaei M and Schenkel FS (2010). Characteristics of linkage disequilibrium in North American Holsteins. *BMC Genomics* 11: 421.

Bradley DG and Cunningham EP (1998). Genetic Aspects of Domestication. In: The Genetics of Cattle (Fries and Ruvinski, eds.). CAB International, Oxon.

Calus MP, de Roos SP and Veerkamp RF (2009). Estimating genomic breeding values from the QTL-MAS Workshop data using a single SNP and haplotype/IBD approach. *BMC Proc.* 3 (Suppl 1): S10.

Cleveland WS (1979). Robust locally weighted regression and smoothing scatterplots. *J. Am. Statist. Assoc.* 74: 829-836.

Cleveland WS (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *Am. Stat.* 35: 34.

Espigolan R, Baldi F, Souza FRP, Gordo DM, et al. (2012). Whole Genome Linkage Disequilibrium in Nellore Cattle. In: IX Simpósio Brasileiro de Melhoramento Animal, João Pessoa.

Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, et al. (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 324: 528-532.

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

Hayes BJ, Visscher PM, McPartlan HC and Goddard ME (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* 13: 635-643.

Hill WG and Robertson A (1966). The effect of linkage on limits to artificial selection. *Genet. Res.* 8: 269-294.

Hill WG and Robertson A (1968). Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38: 26-23.

Khatkar MS, Nicholas FW, Collins AR, Zenger KR, et al. (2008). Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. *BMC Genomics* 9: 187.

Lewontin RC (1964). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49 1: 49-67.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, et al. (2009). Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One* 4: e5350.

McKay SD, Schnabel RD, Murdoch BM, Matukumalli LK, et al. (2007). Whole genome linkage disequilibrium maps in cattle. *BMC Genet.* 8: 74.

Meuwissen TH, Hayes BJ and Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157: 1819-1829.

Michelizzi VN, Dodson MV, Pan Z, Amaral ME, et al. (2010). Water buffalo genome science comes of age. *Int. J. Biol. Sci.* 6: 333-349.

Michelizzi VN, Wu X, Dodson MV, Micha JJ, et al. (2011). A global view of 54,001 single nucleotide polymorphisms (SNPs) on the Illumina BovineSNP50 BeadChip and their transferability to water buffalo. *Int. J. Biol. Sci.* 7: 18-27.

Nagarajan M, Kumar N, Nishanth G, Haribaskar R, et al. (2009). Microsatellite markers of water buffalo, *Bubalus bubalis* - development, characterisation and linkage disequilibrium studies. *BMC Genet.* 10: 68.

Pritchard JK and Przeworski M (2001). Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* 69: 1-14.

Roth J and Myers P (2004). *Bubalus bubalis*. Avaliable at [http://animaldiversity.ummz.umich.edu/site/accounts/information/Bubalus_bubalis.html]. Accessed April 20, 2012.

Santana ML Jr, Aspilcueta-Borquis RR, Bignardi AB, Albuquerque LG, et al. (2011). Population structure and effects of inbreeding on milk yield and quality of Murrah buffaloes. *J. Dairy Sci.* 94: 5204-5211.

Sargolzaei M, Schenkel FS, Jansen GB and Schaeffer LR (2008). Extent of linkage disequilibrium in Holstein cattle in North America. *J. Dairy Sci.* 91: 2106-2117.

Silva CR, Neves HHR, Queiroz S, Sena J, et al. (2010). Extent of Linkage Disequilibrium in Brazilian Gyr Dairy Cattle Based on Genotypes of AI Sires for Dense SNP Markers. In: 9th World Congress on Genetics Applied to Livestock Production Proceedings of the 9th World Congress on Genetics Applied to Livestock Production, Leipzig.

Valdar W, Solberg LC, Gauguier D, Burnett S, et al. (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.* 38: 879-887.

Villa-Angulo R, Matukumalli LK, Gill CA, Choi J, et al. (2009). High-resolution haplotype block structure in the cattle genome. *BMC Genet.* 10: 19.