The Buffalo Genome and the Application of Genomics in Animal Management and Improvement

Daniela IAMARTINOa, John L. WILLIAMSa*, Tad SONSTEGARDb, James REECYc, Curtis Van TASSELLb, Ezequiel L. NICOLAZZIa, Stefano BIFFANIa, Filippo BISCARINIa, Steven SCHROEDERb, Denise A.A. de OLIVEIRAd, Angelo COLETTAe, Jose Fernando GARCIAf, Ahmad ALIg, Luigi RAMUNNOb, Rolando PASQUARIELLOa, Marcela G. DRUMMONDd, Eduardo BASTIANETTOd, Eric FRITZc, James KNOLTESc and the INTERNATIONAL BUFFALO CONSORTIUM

aParco Tecnologico Padano, via Einstein, Lodi 26900 Italy
bARS, USDA, 10300 Baltimore Ave, Beltsville, MD 20705-2350, USA
cIowa State University, Ames IA, dUniversidade Federal De Minas Gerais, Belo Horizonte, Brazil
eUniversity of Naples "Federico II", Naples, Italy,
fSao Paulo State University / UNESP, Brazil
gUniversity of Veterinary and Animal Sciences, Lahore Pakistan
hAssociazione Nazionale Allevatori Specie Bufalina, via Cesare Battisti, Caserta, Italy
*Corresponding email: John.Williams@tecnoparco.org

ABSTRACT
The publication of the human genome sequence in 2001 was a major step forward in knowledge necessary to understand the variations between individuals. For farmed species, genomic sequence information will facilitate the selection of animals optimised to live, and be productive, in particular environments. The availability of cattle genome sequence has allowed the breeding industry to take the first steps towards predicting phenotypes from genotypes by estimating a “genomic breeding value” (gEBV) for bulls using genome-wide DNA markers. The sequencing of the buffalo genome and creation of a panel of DNA markers has created the opportunity to apply molecular selection approaches for this species.

The genomes of several buffalo of different breeds were sequenced and aligned with the bovine genome, which facilitated the identification of millions of sequence variants in the buffalo genomes. Based on frequencies of variants within and among buffalo breeds, and their distribution across the genome compared with the bovine genome, 90,000 putative single nucleotide polymorphisms (SNP) were selected to create an Axiom® Buffalo Genotyping Array 90K. This “SNP Chip” was tested in buffalo populations from Italy and Brazil and found to have at least 75% high quality and polymorphic markers in these populations. The 90K SNP chip was then used to investigate the structure of buffalo populations, and to localise the variations having a major effect on milk production.

Keywords: genetic association study, genomic selection, genome sequence, single nucleotide polymorphisms

Accepted April 10, 2013; Online November 11, 2013.
INTRODUCTION

Genetic improvement of buffalo for traits such as increased milk and meat production, health and efficiency would have a positive impact on agriculture in many regions of the world, and in particular to support poorer economies where buffalo production predominates with respect to cattle. Buffaloes are reared in 129 countries and contribute very significantly to the rural economy, especially in South and South-East Asia (Misra and Tyagi, 2007).

The world population of buffalo is about 158M in comparison with around 1.3B cattle, 1B sheep 500-600M goats. There are two types of domestic water buffalo, the River Buffalo (*Bubalus bubalis*) that are more widely spread globally, and are the predominant type found in the west from India to Europe. The second type, the Swamp Buffalo (*Bubalus carabanensis*), is found more frequently in the east from India to the Philippines.

Systems of buffalo production vary widely in different countries and are governed by several interacting factors, which include the climate, local geography, cropping systems, the size of farms and primary purpose for buffalo production: milk, meat or draught. In Italy, Mediterranean buffalo are kept on large commercial farms under modern intensive systems primarily for producing milk, which is mainly used for *Mozzarella* cheese making. In Latin America Murrah and Jaffarabadi buffalo were originally imported from India, between the 1940s and 1960s. More recently the Mediterranean buffalo has been imported. Today river, swamp and crossbred buffaloes are kept in systems varying from extensive beef production, through rural multi-purpose systems to intensively managed herds for milk production (Borghese, 2005).

The first application of genetics to buffalo improvement was carried out in Italy for the Mediterranean buffalo under the guidance of the Italian Buffalo Breeders Association (ANASB). Mediterranean buffaloes have been exported to many parts of the world. In addition to improved genetics, advanced reproductive technologies including artificial insemination and embryo transfer are used routinely in buffalo husbandry in Italy. At an international level, genetic improvement in production, health and reproductive success would help to sustain buffalo farming and could in part be achieved through molecular assisted genetic improvement.

In this study, the development of a genome-wide 90,000 SNP panel is described. The panel has been used for two practical applications on field data: comparing the efficiency of classical pedigree recording with pedigrees constructed from genomic data and a genome-wide association analysis to localise genes having a large effect on milk production.

MATERIALS AND METHODS

**SNP discovery.** Eighty-six river buffalo from 8 breeds (Italian Mediterranean, Murrah, Nili-Ravi, Jaffarabadi, Kundhi, Aza-Kheli, Egyptian and Swamp type from Philippines) were sequenced to a depth of between 5 and 12X by Illumina paired end reads, yielding a total of 470X genome coverage depth. Buffalo sequences were aligned to the bovine UMD3.1 genome using Burrows-Wheeler Alignment tool (BWA; Li and Durbin, 2009). Aligned sequences were processed with SAMtools (SourceForge.net; Li et al., 2009) and Picard tools (SourceForge.net) to format the
data for SNP calling with the unified genotyper of the Genome Analysis ToolKit (GATK, McKenna et al., 2010). Heterozygous SNPs were kept if they had a base pair quality score of Q>10 and did not have another SNP within 10bp. A total of 22,293,567 SNPs were discovered from four river breeds (Mediterranean, Murrah, Jaffarabady and Nili Ravi). The within and across breeds allele frequencies for these SNPs were calculated and the minor allele frequencies (MAF) were determined.

**SNP-chip design.** All SNPs that did not map uniquely, were too close together and had a low probability of being real SNPs were excluded. A total of 5,800,477 high quality SNPs were retained. The spacing of SNPs on the buffalo genome sequence was estimated by mapping them onto the bovine genome, and the probability of a SNP being polymorphic was weighted by breed: Mediterranean 30%; Murrah 30%; Jaffarabadi 20%; Nili-Ravi 20%. This identified 123,040 SNPs for which probe design could be tested. Buffalo sequence was created around these SNPs by aligning the Buffalo sequence reads with the UMD 3.1 bovine sequence, and optimised 71mer probe sequences were identified for each SNP by Affymetrix using proprietary pipelines. The final Axiom® BUFFALO GENOTYPING array design comprises 123,029 probes, which includes probes for 89,988 SNPs, 5,799 QC probes and 1,784 gender calling probes. The Axiom® BUFFALO GENOTYPING array was tested by genotyping a total of 619 Italian Mediterranean buffalo and 282 Brazilian Murrah buffalos samples. The Affymetrix “Powertools” (APT) and “SNPolisher” R packages were used for quality control of the SNP. A total of 76,559 SNP probes validated when considering all the samples (see results for details).

**Genomic relationships on Brazilian dataset.** Classical recorded pedigree was compared with the genomic relationships for the 282 Brazilian Murrah animals by creating a genomic relationship matrix (VanRaden et al., 2008). Of the 78,136 SNPs validated only on Brazilian Murrah samples, monomorphic SNPs and those with MAF<5% were removed (10,556 SNP) leaving 67,580 SNPs which were used to create the G matrix.

**Genome-Wide Association analysis on Italian dataset.** Milk EBVs and lactation records for the 619 Italian Mediterranean buffalo coming from 4 farms in the Lombardy region (Italy) was provided by The Italian Buffalo Breeders Association (ANASB). After quality control to remove replicates, animals with missing data or failing pedigree checks, 529 records were retained. A total of 20,914 probes were discarded because of poor quality on these samples giving a final dataset of 78,137 SNP probes for these samples.

Associations between SNP genotypes and phenotypes were analyzed by fitting all SNP simultaneously using the GRAMMAR procedure (Genome-wide Rapid Association using Mixed Model And Regression: Aulchenko et al. (2007). The procedure involved two steps. First, an additive polygenic model was used to obtain individual environmental residuals using the polygenic function of the GenABEL package (Aulchenko et al., 2007). Then, association between residuals and genetic polymorphisms was tested using least squares methods.

The following model was used to estimate residuals:

\[
LactRecord_{ijkpq} = \mu + Farm_i + CalvYear_j + CalvSeason_k + Calvings_p + Age_q + \text{Polygenic}_r + e_{ijkpq}
\]
where \( \text{LactRecord} \) is a 270 DIM conventional lactation record, \( \mu \) is the general mean, \( \text{Farm} \) is a fixed farm effect (\( i=1,4 \)), \( \text{CalvYear} \) is a fixed effect for calving year (\( j=1,2 \) for pre and post 2010), \( \text{CalvSeason} \) is a fixed effect for season of calving (\( k=1,4 \)), \( \text{Calvings} \) is a fixed effect for the number of calvings (\( p=1,2 \) for primiparous and multiparous animals), \( \text{Age} \) is a covariate for age (in months), \( \text{Polygenic} \) is a polygenic effect for animal \( r \), and \( e \) is the random residual, with \( e \sim N(0, \sigma^2_e) \).

The polygenic effect was included to account for genetic sub-structure, as higher or lower degree of genomic relationship between animals can have a direct impact on estimates, increasing false positives and negatives.

**RESULTS**

**Axiom Genotyping.** All of the 901 samples belonging to Italian Mediterranean and Brazilian Murrah breeds genotyped using the Axiom array were female, hence the gender calling probes were not considered in the analyses. All plates had an initial call rate of 97% or greater and were considered to have passed the Affymetrix Quality Control (QC) parameters. Of the 89,988 SNP probes, 67,330 (74.8%) were polymorphic (PolyHighResolution) and 9,229 (10.3%) gave high quality signals but were monomorphic (MonoHighResolution) in the two breeds tested, but may be polymorphic in other breeds. Less than 0.1% (83 probes) gave spurious signals with variable intensity (Variable Intensity Non-hybridizing Oligo, VINO cluster), which most likely arise from variations within the target probe sequence that were not identified in the sequence set used for the probe design. About 1.7% of probes (1,494) were missing one of the homozygous genotypes (NoMinorHom). Only 4.1% (3,668) of the probes had a call rate below the threshold (0.15) and 9.1% were rejected for low quality genotypes. Considering the PolyHighResolution SNPs, the average sample call rate was 99.75% and the average sample reproducibility comparing the 26 replicate samples included in the sample set was 99.96%, demonstrating a high quality of the genotyping results.

**Identification of pedigree errors.** Recorded pedigrees for all populations contain errors, even when good systems to verify data are in place, with errors of up to 25% estimated for cattle populations (Matukumalli et al., 2009). A more accurate measure of relationship may be obtained considering the fraction of DNA in common between individuals, based on alleles shared at given loci. Genomic relationship measures are useful in selection and parentage testing (Dodds et al., 2005) and have been used to manage genetic diversity (Caballero and Toro, 2002).

The pedigree information on the Brazilian samples was available only as far as the sire and dam. Seven bulls in the sample had more than 10 daughters and were used to evaluate possible pedigree errors using genomic data. Clustering the daughters using a genomic relationship produced tight homogeneous groups of daughters for 5 of these bulls, while the daughters of the remaining 2 bulls fell into dispersed heterogeneous clusters. The heterogeneous clusters may be an effect of mendelian sampling, but are more likely to be caused by errors in the paternal assignment for the 2 bulls. Identification of potential pedigree errors to clean data sets prior to analyses is necessary in association studies to help reduce false positive associations, and to improve the accuracy of estimation of standard as well as genomic breeding values.
Genome-wide Association Study (GWAS) for milk production traits. Population structure was assessed using multi-dimensional scaling, which showed clear population structure clusters depending on the farm of origin. Clusters for Farms 1,859 and 26,225 overlapped, while clusters for Farm 71,801 and 61,207 were independent, showing only minimal overlap with each other and the other two farms (Figure 1). This genetic sub-structure is most likely due to a sire effect: while Farms 1,859 and 26,225 predominantly use bulls available for artificial insemination, and therefore have bulls in common, Farms 71,801 and 61,207 use mainly natural service bulls and only occasionally use the AI bulls in common with the other farms.

Heritability of the trait based on a classical animal model and recorded pedigree (additive relationship matrix) was 0.38, while it was 0.45 when using the genomic relationships matrix. This is similar to the results in cattle for the same trait (Interbull, 2009).

As the draft genomic sequence of the buffalo is currently based on the bovine genome, the results shown assumed the order of loci on the bovine genome, and reported chromosomes are bovine equivalents. Moreover, as the draft buffalo genome is not yet annotated, alignment with the bovine genome also facilitated the use of the bovine information to explore genes under the peaks identified.

Loci associated with milk production (genome-wide significance \(P<10^{-04}\)) were found on the chromosomes corresponding with bovine chromosomes 4, 11 and 19. Suggestive associations were also found on chromosomes 10, 14 and 23 (see Figure 2). The SNPs most significantly associated with milk yield are shown in Table 1.

The most convincing association with milk yield data was found on chromosome 11, with several significant SNPs (see Figure 2 and Table 1). In particular, three of the significant SNPs (AX-85041172, 85125077, 85114201) are located near two gene dense regions. The closest gene downstream is gene coding for the apolipoprotein B (APOB), a protein involved in the early lactation lipid metabolism in cow (Gruffat et al., 1997). This is probably the best functional candidate for the chromosome 11 QTL. Growth differentiation factor 7 (GDF7), and U4 small nuclear RNA genes are also present on this region. Upstream of these significant SNP are kelch-like gene 29 (KLHL29) and the U6 non-coding small nuclear RNA. The other two SNPs with strong association located on chromosome 11 (AX-85080229 and AX-85093842) are close the neurexin-1-alpha precursor gene (NRX1) and located upstream a dense gene region.

The most significant SNP on chromosome 19 (AX-85140457) is close to the short chain dehydrogenase/reductase SDR family member 11 gene (DHSR11), which codes for one of the enzymes involved in the metabolism steroid hormones, prostaglandins, retinoids and lipids.

The SNP on chromosome 4 (AX-85143079) is located downstream the Collagen alpha-2 gene (COL1A2) which is associated with milk ability and more precisely the kinetics of milk emission in sheep (Dhorne-Pollet et al., 2012).

DISCUSSIONS

The data reported here describes the creation of a genome wide SNP panel for buffalo, which has been tested to analyse pedigree structure of a buffalo
population and to carry out a GWAS analysis for milk yield in buffalo. The GWAS identified SNPs strongly associated with milk yield on chromosomes 11 and 19. Further investigation of these regions with additional data is necessary to confirm the associations and fine-map the QTL regions.

In the future, breeding programs will be driven primarily by high-density SNP data (Eggen, 2012). Genomic information will be used both to identify the causal mutations to select for superior variants at loci having a major effect on desired phenotypes, or used to calculate estimated genomic breeding values to reduce the selection interval, thus boosting genetic improvement. The work presented here takes the first step towards these goals for the buffalo.

ACKNOWLEDGEMENTS

This study was partly funded by the ASTIL-International Cooperation project Italy-Brazil “BuffaloSNP” (Regione Lombardia, Italy). Sequence data for SNP discovery was contributed by the Italian Buffalo Genome Consortium and International Buffalo Genome Consortium.

REFERENCES


Interbull routine genetic evaluation for dairy production traits. [http://www-interbull.slu.se/eval/aug09.html]


Table 1. The 9 most significantly associated SNPs.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chromosome</th>
<th>Position bp</th>
<th>N</th>
<th>effB</th>
<th>se_effB</th>
<th>Pc1df</th>
</tr>
</thead>
<tbody>
<tr>
<td>AX-85080229</td>
<td>11</td>
<td>34107047</td>
<td>529</td>
<td>-111.53</td>
<td>28.16</td>
<td>7.245e-06</td>
</tr>
<tr>
<td>AX-85114201</td>
<td>11</td>
<td>76258496</td>
<td>529</td>
<td>-94.28</td>
<td>24.95</td>
<td>1.863e-05</td>
</tr>
<tr>
<td>AX-85125077</td>
<td>11</td>
<td>75854702</td>
<td>529</td>
<td>106.17</td>
<td>28.23</td>
<td>2.042e-05</td>
</tr>
<tr>
<td>AX-85143079</td>
<td>4</td>
<td>11676240</td>
<td>529</td>
<td>157.72</td>
<td>42.06</td>
<td>2.163e-05</td>
</tr>
<tr>
<td>AX-85093842</td>
<td>11</td>
<td>32955361</td>
<td>529</td>
<td>-86.06</td>
<td>23.49</td>
<td>2.163e-05</td>
</tr>
<tr>
<td>AX-85048470</td>
<td>23</td>
<td>6051649</td>
<td>529</td>
<td>303.50</td>
<td>83.96</td>
<td>4.227e-05</td>
</tr>
<tr>
<td>AX-85140457</td>
<td>19</td>
<td>13338605</td>
<td>528</td>
<td>-132.79</td>
<td>37.09</td>
<td>5.009e-05</td>
</tr>
<tr>
<td>AX-85041172</td>
<td>11</td>
<td>75787638</td>
<td>528</td>
<td>87.01</td>
<td>24.43</td>
<td>5.493e-05</td>
</tr>
<tr>
<td>AX-85086756</td>
<td>29</td>
<td>9205713</td>
<td>528</td>
<td>-96.70</td>
<td>27.46</td>
<td>6.637e-05</td>
</tr>
</tbody>
</table>

Figure 1. MDS plot representing the structure of the four farms used in the study.
Figure 2. Manhattan plot GWAS-Lactation.