Identification of single nucleotide polymorphism from Indian *Bubalus bubalis* through targeted sequence capture

A. B. Patel[1,2], R. B. Subramanian[1], H. Padh[1], T. M. Shah[2], A. Mohapatra[2], B. Reddy[2], S. J. Jakhesara[2], P. G. Koringa[2], D. Dash[3] and C. G. Joshi[2,]*

[1]BRD School of Biosciences, Sardar Patel University, V. V. Nagar 388 120, India
[2]Department of Animal Biotechnology, College of Veterinary Science and Animal Husbandry, Anand Agricultural University, Anand 388 001, India
[3]Roche Diagnostics India Pvt Ltd, Kolkata 700 107, India

***Bubalus bubalis* (water buffalo) is an agro-economically important livestock species due to its multipurpose use in India and other Asian countries. The aim of this study was to identify single nucleotide polymorphisms (SNPs) from buffalo genome. Genomic DNA was isolated from 24 blood samples of three Indian buffalo breeds and subjected to targeted pyrosequencing, followed by variant calling and annotation. Target probes for enrichment were designed from exome and 5′ and 3′ untranslated regions of cattle genome. By targeted pyro-sequencing and variant calling from 3.92 Gb data, 923,964 high-quality SNPs were identified. Many SNPs were identified in regulatory regions, leading to conformational changes in factor-binding sites, which play a role in gene expression as in the case of *LPL* gene from low-milk-producing samples. Gene ontology (GO) enrichment and clustering, resulted in the enrichment of GO terms involved in milk production and transport, and fertility-related categories. Around 75% of SNPs were located on cattle quantitative trait loci, supporting trait-wise sample collection approach. Further, PCA analysis from the identified SNPs also supported sample selection strategy based on contrasting trait performance.**

Keywords: Exome, gene ontology, quantitative trait locus, single nucleotide polymorphism.

WATER buffalo (*Bubalus bubalis*) was domesticated approximately 5000 years ago in India to secure supply of milk, meat and power[1]. It has been grouped into (i) swamp, primarily developed for draught purpose and (ii) river buffalo, primarily used for milk production. Among the total of 13 recognized breeds of water buffalo, majority are milch breeds in India and some of them have been listed on a state-level conservation plan by the Ministry of Agriculture, Government of India[2]. As buffalo milk occupies the highest share in Indian dairy sector, the future improvement in traits of economic importance is dependent on genetic variation present within and between breeds. Even though they have an important role in Indian agricultural economy, most of the breeds have not been exploited for their full genetic potential.

Recently, genomic selection in cattle has been adopted globally to accelerate genetic gains[3]. Molecular markers like single nucleotide polymorphisms (SNPs) can play a significant role in livestock improvement through conventional breeding programmes. However, the present genomic resources are limited for river buffalo. Moreover, molecular genetic diversity in river buffalo is explored using cattle-based microsatellite markers[4]. Taking advantage of the availability of fully sequenced cattle genome and other related genomic resources, and given the close evolutionary relationship between cattle and river buffalo. We sequenced the river buffalo genomes on a large scale to detect genetic variants, in particular, identified large-scale SNPs, which may help in the study of river buffalo genomics. Genetic component plays a major role in milk production and other functional traits of dairy animal[5].

The advent of next-generation sequencing has enabled a robust and more cost-effective approach for the identification of high-throughput SNPs. Recently, exome/targeted capture sequencing has been used to analyse disease traits in livestock species because it is efficient and cost-effective[6]. In the present study we carried out targeted sequencing, for discovering variants in and across targeted regions. To the best of our knowledge, there are no earlier studies on targeted (exome) sequencing in river buffalo for high-throughput variant discovery.

Material and methods

Sample collection and genomic DNA extraction

Three river buffalo breeds, viz. Banni, Mehsani and Jafrabadi from Gujarat, India were sampled (8 samples

*For correspondence. (e-mail: cgjoshi@aau.in) per breed, total 24 animals). Blood samples from unrelated animals from the fields with known physiological state for milk production (high and low milk yield) and fertility status (fertile and infertile) were collected (see Supplementary Information; Table S1 online). The gDNA was isolated from blood samples using a Qiagen DNeasy Blood and Tissue kit (Qiagen Corp., CA, USA). DNA was quantified using Qubit® dsDNA BR Assay (Invitrogen Corp., CA, USA) and integrity was confirmed by agarose gel.

Nomenclature of sample: 'BT#', Breed, trait and #: animal laboratory id number. (For example BHP1 indicates Banni high producer animal with laboratory id number 1, BLP1 indicates Banni low producer animal with laboratory id number 1).

Selection of targets and next-generation sequencing

For the intended targets (all coding exons, 3′ UTR and 5′ UTR exons), Baylor Btau_4.6.1/bos-Tau7 genome build associated RefGene tables were downloaded from UCSC Browser and provided to NimbleGen (Roche, Germany) for custom probe design compatible with Roche GS-FLX Titanium chemistry. Rapid library for each sample was prepared from ~1 μg of gDNA separately and multiplexed according to the manufacturer's protocol (Roche) using high-quality DNA. Final libraries were used for setting up hybridization reaction at isothermal temperature of 47°C for 68–72 h in a thermal cycler, with custom-designed probes according to the manufacturer's protocol (NimbleGen, USA). Captured DNA libraries were quantified spectrophotometrically, and evaluated electrophoretically with high sensitivity DNA assay on Agilent Bioanalyzer 2100 (Agilent, USA), and sequenced on GS-FLX Titanium using XLR70 chemistry following the manufacturer's protocol.

Bioinformatics data analysis

The variant identification pipeline consisted of (1) data quality filtering, (2) mapping against cattle genome, (3) post-mapping quality filtering, (4) variant calling and filtering, and (5) variant annotation.

Raw sequencing data were separated sample-wise in .fasta and .qual files from standard flowgram files (.sff file) using sffinfo command tools (Roche). The generated .fasta and .qual files were combined in .fastq file, which were then quality-filtered based on sequence length ≥40 bp and mean sequence quality score of ≥25 using PRINSEQ[7]. Sequences passing above criteria were used for mapping against *Bos taurus* genome build 4.6.1. using 'bwa-mem' module of Burrows–Wheeler Alignment Tool v 0.7.5a (ref. 8). PCR duplicates were removed from the mapped .bam files using 'MarkDuplicate' module of Picard Tools[9]. Finally, SNPs were identified using the mpileup utility of SAMtools[10].

Next, the identified SNPs were quality-filtered with SNPs by depth (coverage) of ≥5 and SNPs by phred-like consensus quality score of ≥25. From filtered high-quality SNPs, subsets of variants were generated that were present in either high-milk and low-milk production samples, or and shared SNPs (common), likewise subsets for fertility trait samples were generated using vcf-isec, vcf-annotate, vcf-compare; perl modules and scripts of VCFtools (v 0.1.11)[11]. The extracted shared and unique subsets of SNPs were annotated using SnpEff (v3.4)[12]. Further, genes harbouring SNPs were annotated for their involvement in the biological system through gene enrichment functional annotation tool of Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 (ref. 13) with *B. taurus* as background. Based on derived gene ontology (GO) annotation, we analysed genes from lipid-related metabolism pathways using STRING database[14]. The stability, effect and secondary structure of mutated proteins were predicted by I-Mutant2.0 (http://folding.biofold.org/i-mutant/i-mutant2.0.html) and RNAsnp[15]. Protein stability was determined by free energy change value, $\Delta\Delta G$. $\Delta\Delta G < 0$ denotes decreased stability, whereas $\Delta\Delta G > 0$ shows increased stability.

QTL regions were identified from information on cattle QTLs in the Animal QTLdb (Release 23; http://www.animalgenome.org/cgi-bin/QTLdb/BT/index)[16]. QTL locations (Btau_4.6) were downloaded for milk and fertility QTLs. Totally 9180 QTLs were identified in the cattle genome; 1424 and 949 QTLs were associated with milk production and fertility traits respectively. The high-quality SNPs were intersected on milk production and fertility trait QTLs using BEDTools[17].

Relationship among breeds and genetic structure

Pairwise F_{ST} values were calculated on the primary/pooled, SNP dataset in JMP Genomics (SAS, Cary, NC) using Reynolds' distance[18] with significance tested using 10,000 permutations. A neighbour-joining (NJ) cladogram was built using breed allele frequencies calculated from the SNP dataset utilizing the Population Measures in JMP Genomics (SAS, Cary, NC). Bootstrap support from 1000 iterations of the data was used to assess support for the resulting majority rule consensus cladogram. Principal component analysis (PCA) was conducted using Relationship matrix module in JMP Genomics on the SNP set consisting of all 24 individuals.

Results

Performance of target enrichment and sequencing

In-solution custom capture probes designed by NimbleGen consisted of 125,679 exomes, 14,084 3′ UTRs and 16,574 5′ UTRs comprising ~31.19 Mb of genomic

Table 1. Summary of sequencing data for milk production trait samples of Banni, Mehsani and Jafrabadi buffalo breeds

	Banni				Mehsani				Jafrabadi			
	BHP1	BHP2	BLP1	BLP2	MHP1	MHP2	MLP2	MLP5	JHP1	JHP2	JLP1	JLP2
Raw sequences	604,742	629,606	468,655	544,069	544,403	657,400	573,466	437,957	447,055	445,066	632,961	576,284
Total bases (Mb)	229.7	240.7	179.2	206.1	197.3	238.4	213	165.4	172.3	180	249	227.3
Total sequences after filtering	510,789	535,496	396,420	462,671	453,486	548,207	481,812	372,075	376,677	375,909	549,505	500,637
Total bases after filtering (Mb)	207	217.6	160.5	185.7	176.8	213.8	191	149	154.4	161	227.7	207.9
Reads used for variant calling after filtering	413,155	435,727	322,109	369,306	365,464	441,819	391,269	303,783	309,690	307,439	465,751	423,730

Table 2. Summary of sequencing data for fertility trait samples of Banni, Mehsani and Jafrabadi buffalo breeds

	Banni				Mehsani				Jafrabadi			
	BF2	BF3	BI2	BI3	MF1	MF2	MI2	MI4	JH1	JH2	JI1	JI2
Raw sequences	546,122	570,183	512,804	395,870	293,200	456,404	512,104	326,339	493,287	542,546	429,923	420,414
Total bases (Mb)	191.9	202.2	180.5	138.1	116.7	187	202.5	132.8	187.3	205.6	159.8	156.2
Total sequences after filtering	476,206	499,618	437,746	338,092	242,240	393,666	423,102	280,674	431,890	474,577	370,936	364,090
Total bases after filtering (Mb)	175	184.9	162.5	124.3	102.4	168.9	178.1	119.9	171.7	188.4	145.2	142.3
Reads used for variant calling after filtering	476,031	499,433	437,584	337,968	242,214	393,640	423,076	280,648	431,718	474,561	370,920	364,074

sequence. Totally 3.92 Gb sequence data comprising ~12 million reads with median read length of 430 bp were generated from 24 samples of three buffaloes. After quality-filtering of raw reads, ~3.55 Gb data comprising ~10 million high-quality reads were mapped sample-wise against the cattle genome with an average mapping rate of ~98% (Tables 1 and 2). An average of 67.80% targets were sequenced with depth ≥5X. (see Supplementary Information, Table S2 online). To obtain an indirect information about target enrichment, Ts/Tv (transitions/transversions) ratio was calculated. Overall Ts/Tv of ~2.6 was observed for all detected variants. However, slightly higher Ts/Tv ratio was observed when annotation was confined to target regions having comparatively higher sequencing coverage compared to annotation which included off-target regions[19].

Identification of SNPs and annotation from pooled data of milk production and fertility trait samples

The targeted sequencing and SNPs calling resulted in a total 477,996 high-quality SNPs. Based on sequence ontology terms, 18.33% SNPs were located in exons, 44% in introns, while 25.6% and 2.82% SNPs were located in the flanking and intergenic regions of genes (Figure 1).

Identification of SNPs and annotation from fertility trait data subset

In the fertility trait group data, a total of 540,414 high-quality SNPs were identified (Table 3). Based on sequence oncology (SO) term assignment, 17% SNPs were located in the exonic regions, 44.9% in the intronic regions, 26.1% in the flanking regions and 2.89% SNPs in intergenic regions. Totally 0.26% (2090) SNPs had high impact, leading to either loss or gain of start/stop site. In order to detect potential SNPs involved in fertility, we searched across all detected SNPs of the fertility subset and found various SNPs in genes like *SATT5A*, *FAF2*, *PGR*, *FSHR* and *PAPPA2*, which are reported to be associated with the fertility trait.

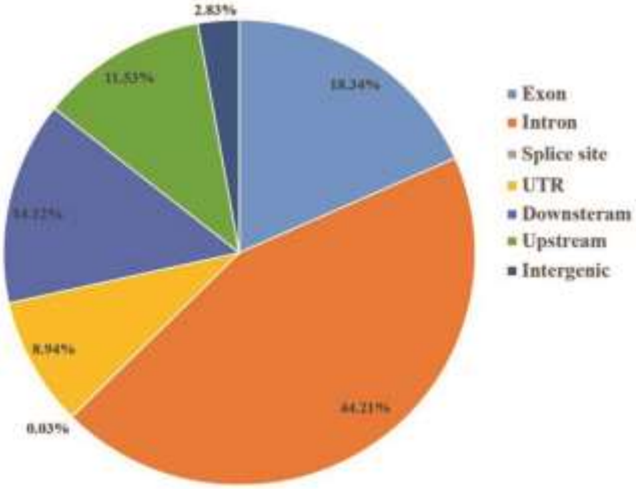

Figure 1. Distribution of single nucleotide polymorphisms (SNPs) identified from pooled data. SNPs were identified by sequence comparison of 24 animals across genomic regions after annotating variant file with SnpEff. We found 126,816 exonic SNPs, 305,740 intronic SNPs, 172 SNPs in splice site, 19,552 intergenic SNPs, 61,816 SNPs in untranslated region (UTR) and 177,359 SNPs in 5 kb flanking regions each, upstream and downstream.

GO terms enrichment for fertility trait samples

For fertility trait, gene ontology (GO) terms under biological processes were related to processes like nuclear division, mitosis, organelle fission, M-phase mitotic cell cycle, cell-cycle process, ATP biosynthetic process, nucleotide metabolic and biosynthetic processes, phosphorylation and phosphate-related metabolic processes, positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic processes and nitrogen metabolic processes. In case of cellular component, GO terms related to cytoskeleton, nucleus, nucleoplasm, intracellular, cell-substrate adherens junction membrane, protein complex biogenesis and assembly, extracellular region and protein transport were enriched. For molecular function, subsets related to ATPase activity, hydrolase activity, transport associated activity, nucleotide-binding, metal ion-binding and helicase activity were clustered, which are observed to be highly active molecular process at specific stage during reproduction (see Supplementary Information, Table S3 online).

Distribution of identified SNPs from fertility trait samples on cattle fertility trait QTLs

In fertility trait samples, chromosome-wise analysis identified SNPs located within fertility QTLs, where highest and lowest proportion of SNPs were observed from chromosome 6 (81.79%) and chromosome X (2.46%) respectively (Figure 2).

Identification of SNPs and annotation of milk production data

After filtering variants based on coverage and phred base quality score, 383,550 high-quality SNPs were identified (Table 3). Sequence ontology annotation revealed that the SNPs were distributed in exonic (10.14%), intronic (55.11%), regulatory (27.25%) and intergenic (7.42%) regions.

Table 3. Summary of single nucleotide polymorphisms (SNPs) identification from milk production and fertility trait group samples

Group	SNPs*	High quality SNPs**	Non-synonymous changes
Total pooled data	9,986,501	477,996	48,309
Fertility group	7,442,920	540,414	50,237
Milk production group	5,734,242	383,550	42,157

*SNPs in various groups were identified against cattle as reference. **SNPs were filtered based on ≥25 phred quality score and ≥5X sequencing depth.

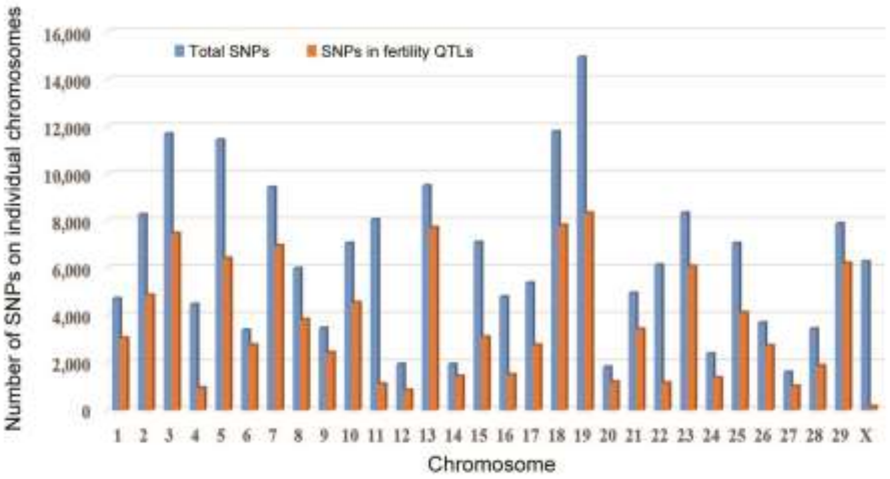

Figure 2. Chromosome-wise distribution of SNPs on cattle fertility QTLs. Distribution of identified SNPs from fertility trait samples on QTLs of fertility traits was found by intersecting variant file with QTL coordinates (in Mb) using BEDTools. QTLs of fertility traits were downloaded from cattle QTLdb (http://www.animalgenome.org/cgi-bin/QTLdb/BT/browse).

GO enrichment analysis for milk production data subset

GO analysis reflected the involvement of genes in milk production and translocation (see Supplementary Information; Table S4 online). We found that various GO categories involved in different pathways and processes like lipid metabolism, carbohydrate metabolism processes, cellular development and protection and response to stimulus, biosynthesis energy generation; synthesis, transfer and secretion of biomolecules; ion-binding and homeostasis; cellular development and proliferation were found to be associated with milk globule formation, along with other milk components formation and secretion[20,21]. Under all the three categories, GO terms reflected the involvement of genes in milk production and translocation. Most of identified SNPs were located in genes that have been reported to be potentially associated with economically important traits for other mammal species.

Distribution of identified SNPs from milk production trait data on cattle milk QTLs

Figure 3 shows the distribution of high-quality SNPs across the genome for each milk QTL. In milk fat category, 197,083 high-quality SNPs were located in QTLs spans. Chromosome-wise, highest proportion of SNPs was located on chromosome 10 (87.83%), and lowest on chromosome 25 (0.24%). Similarly, for milk protein and milk yield traits QTLs, the highest number of SNPs was located on chromosomes 6 (88.01%) and 5 (88.43%), while the lowest number of SNPs was located on chromosomes 24 (1.42%) and 19 (0.54%).

Among-breed relationship

An NJ tree was constructed on the basis of the Nei's genetic distances with relatively high bootstrap values (Figure 4). The Banni and Mehsani breeds clustered closely together, whereas Jafrabadi breed clustered differently. We identified three clusters: Banni and Mehsani (cluster I), Jafrabadi (cluster II) and cattle (cluster III). This grouping pattern was further supported by PCA analysis, which was used to study possible genetic relationships among these buffalo breeds. The first principal component (PC) explains 98.5% of the observed genetic variation, and the second and third PCs resolve 0.1% and 0.1% of this variation respectively. Together, these three PCs account for 98.7% of the total genetic variation. A PCA plot for the three breeds (Figure 5) revealed clustering

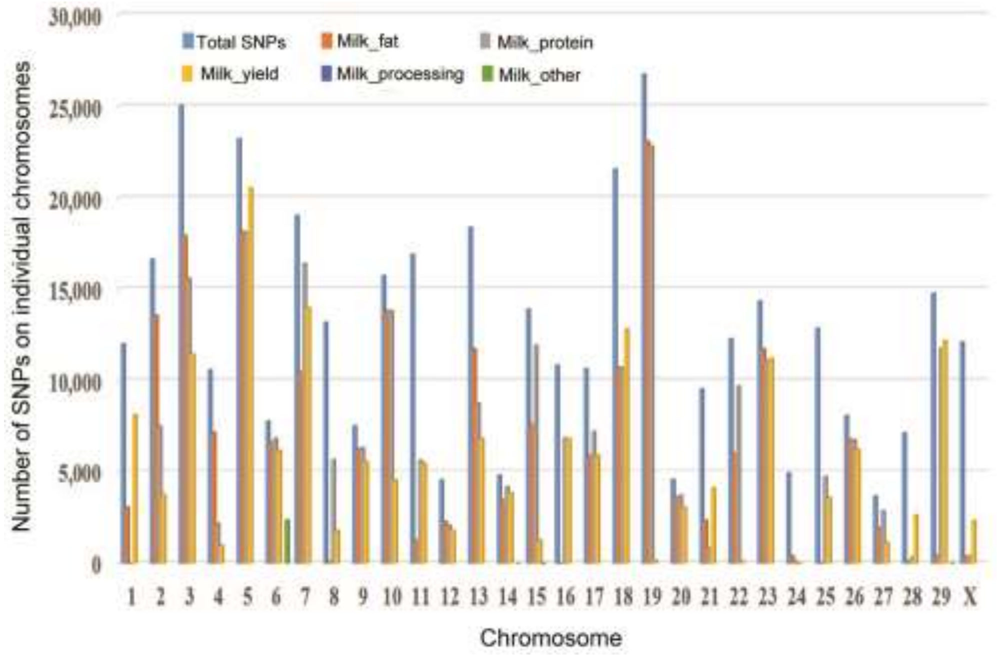

Figure 3. Chromosome-wise distribution of SNPs on cattle milk QTLs. Distribution of identified SNPs from milk production trait samples on QTLs of various milk traits was found by intersecting variant file with QTL coordinates (in Mb) using BEDTools. QTLs of milk traits were downloaded from cattle QTLdb (http://www.animalgenome.org/cgi-bin/QTLdb/BT/browse).

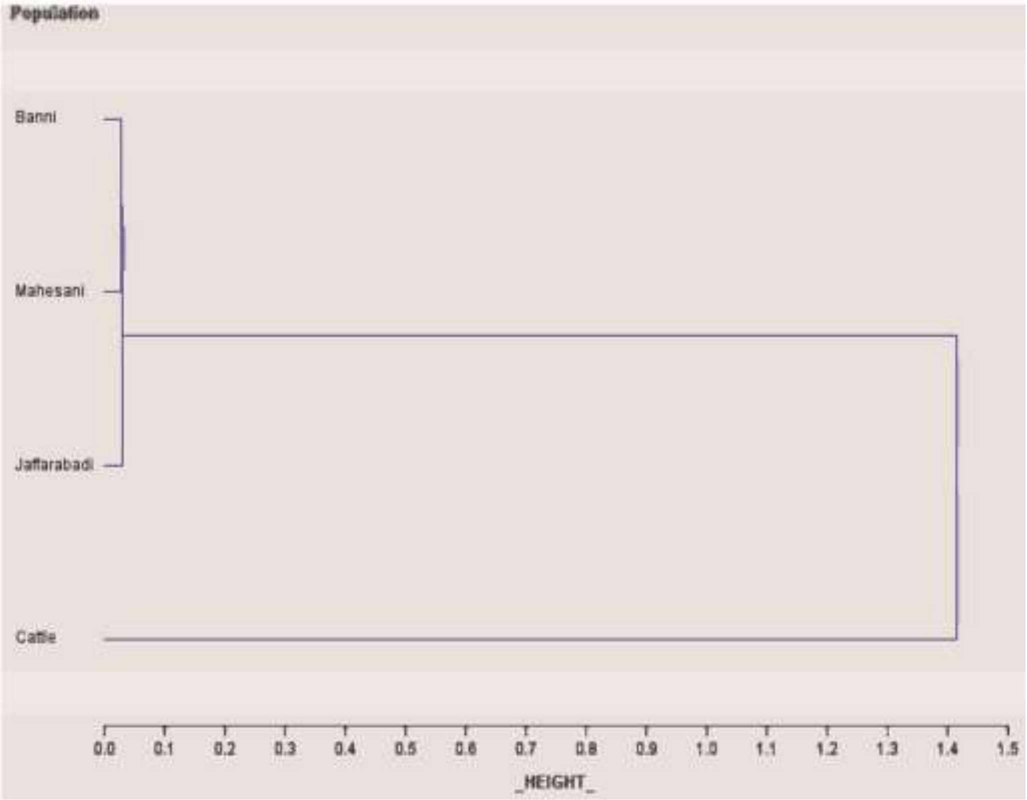

Figure 4. Nei’s genetic distance-based neighbour joining tree calculated from SNP frequencies in 24 individuals from three different buffalo breeds. Cattle were considered out group with reference to alleles.

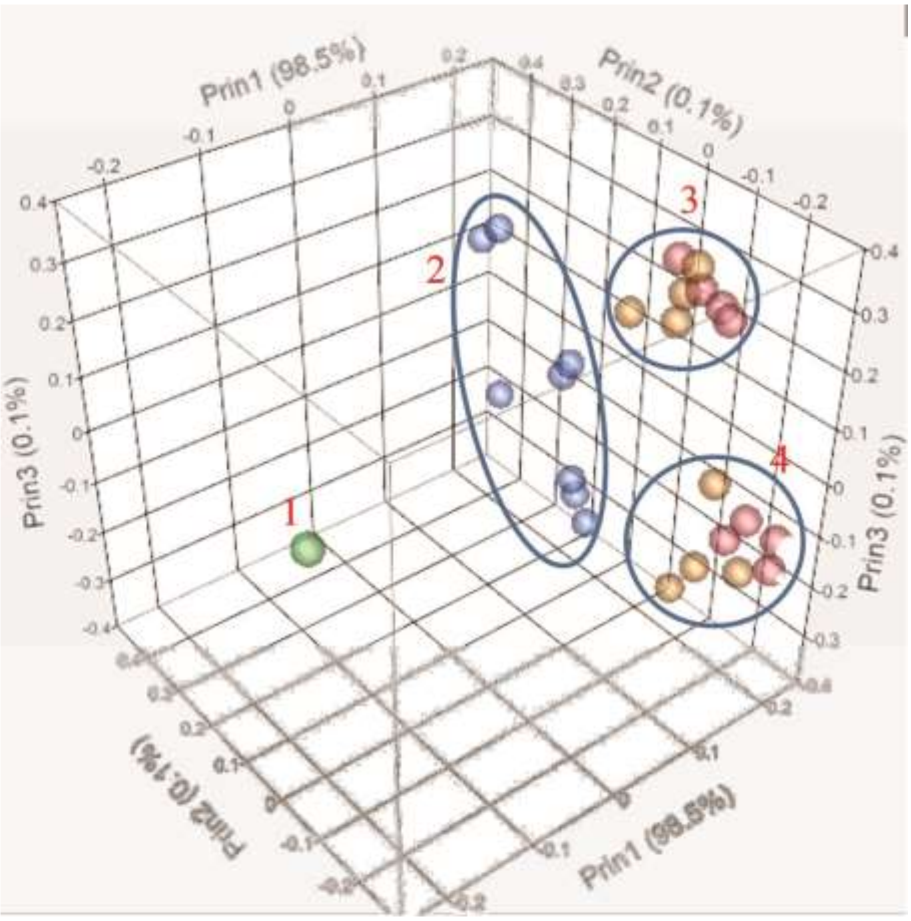

Figure 5. Principal component analysis capturing clear difference in milk production and fertility trait animals of Banni (pink colour), mehsani (golden colour) and Jafrabadi (blue colour) buffalo breeds. Banni and Mehsani clustered together (clusters 3 and 4) compared to Jafrabadi (cluster 2). High and low producers (cluster 3), and fertile and infertile clustered together (cluster 4). Cattle (green colour) were considered out group.

of Banni (pink colour) and Mehsani (golden colour) into one group, while Jafrabadi (blue colour) formed a different cluster, as observed in an NJ tree. The PCA analysis also revealed that milk production and fertility trait SNPs clustered separately.

Discussion

In the present study, we generated 3.55 Gb sequence data by targeted pyro-sequencing. The data were mapped against cattle genome assembly[22] with overall mapping rate of ~98%. Mapping rate was higher compared to that reported in an earlier study[23], mainly due to experiment design, wherein we have targeted coding regions which are conserved compared to other parts of the genome, followed by detection of SNPs. Furthermore, unmapped PCR duplicates and multiple mapped reads were removed for downstream analysis to reduce computational time and also to mitigate the effect of PCR amplification bias that might be introduced during library preparation during pre- and post-hybridization amplification steps. Enrichment efficiency was calculated as a quality-control step for checking specificity of probes, which was in concordance with an earlier study where enrichment increased with increase in sequencing coverage[24]. Along with the targeted region, we also detected SNPs in non-target regions which might have important roles in gene functions as reported earlier[25]. For all SNPs detected in the target region Ts/Tv ratio was ~2.6, which was higher than that observed in whole-genome resequencing studies and comparable to other exome sequencing studies[25]. Likewise, as we increased sequencing coverage, SNPs were confined to targeted regions only, and *Ts*/*Tv* ratio increased to 3–4 as observed in previous studies[26].

In the case of fertility trait, variants in genes related with fertility and reproduction have been identified. We have detected variants in 3′ UTRs, exons and introns of signal transducer and activator of transcription 5A (*STAT5A*). The gene mediates its action via peptide hormones (like progesterone and growth hormone) and cytokines in target cells and found to be associated with fertilization and embryonic survival rate[27]. Additionally, fibroblast growth factor 2 (*FGF2*) and progesterone receptor (*PGR*) were found to be involved in fertilization, development and survival of fertilized embryo[28–30], while growth hormone receptor (GHR), follicle stimulating hormone receptor (FSHR), Leptin (LEP) and pregnancy-associated plasma protein A2 (PAPPA2) were found to be associated with ovulation, calving interval, survival and growth rate of cattle, and perinatal mortality[31–35]. Further, all these proteins, except *FGF*2, *FSHR* and *PAAPPA2* have direct interaction with each other. For example, interaction of *STAT*5A and *PGR* on β-casein promoter leads to repression of transcriptional activity during pregnancy establishment and early embryonic survival and development[31,36]. Hence mutations with high impact in these genes will affect embryonic survival and development. GO terms enrichment under different GO categories like cell division, cellular and tissue morphology, energy metabolism and metabolic exchange processes of various biomolecules movement, reflect involvement of genes under these processes in fertilization and further supports their involvement in embryo development[37]. Similar GO terms enrichment for reproduction-related genes in humans and rodents was observed in an earlier study[38].

For milk production dataset, most of identified SNPs were located in genes that have been reported to be potentially related with economically important trait for other mammal species. Under all the three categories, GO terms reflect the involvement of genes in milk production and translocation. Under all the three categories, GO terms reflect the involvement of genes in milk production and translocation. For example, genes like growth hormone (*GH*), and leptin (*LEPI*) are associated with growth and development[39–41]; *BTN1A1* and *XDH* are involved in milk fat droplet formation[42,43]. We detected two synonymous (T > C, A > G) SNPs in exons 6 and 10 of the *LPL* gene from low-milk production group, leading to change in secondary structure (see Supplementary Information, Figure S1 online). Due to change in secondary structure, there was increase in $\Delta\Delta G$ for mutant compared to wild type (–95.80 to –94.10 kcal/mol), which in turn might affect the properties of transcription factor binding pockets, leading to change in the gene expression[44–47]. In our study, *LPL* gene expression was decreased, whereas under usual conditions, it is found to be highly expressed in mammary cells during lactation and also associated with traits like carcass trait and visceral fat deposition in cattle which is important for supplying fatty acids to mammary gland by hydrolysis of triglycerides from very-low density lipoproteins[42,43,48,49]. A non-synonymous SNP (p.R260Q) was detected in exon 5 of *ACSL3* gene from high group samples. The *ACSL3* gene is important for maintaining energy balance in the organism, as it plays a vital role in the metabolism of fatty acids by catalysing formation of acly-coA as well as for catabolism of fatty acids via β-oxidation[50,51], and has been reported as expressed during various lactation stages[52]. Based on I-mutant 2 prediction, there is minor change in free energy ($\Delta\Delta G$ = –0.21). Moreover, wild and mutant amino acids (p.R260Q) are polar/hydrophilic in nature and belong to the same amino acid group, this might be reason for very little change in free energy. Also changing amino acid might not have been exposed to the outer surface, which could have damaging effect on the protein[53]. The UTR regions on either side of the gene play an important role in expression, modulation and transport of mRNA from the nucleus along with efficient translation of mRNA[54,55]. Nine variants were detected in the UTR genes involved in fatty acid metabolism (see Supplementary Information; Table S5 online), like fatty acid binding proteins (*FABP*), acyl-CoA synthetases (*ACSL*), ATP-binding cassette, sub-family G (*ABCG*)[56,57]. We identified variants in isoforms 2 of *ABCG* gene, which have an important role in milk yield and composition, as in case of *ABCG2* isoform, where single nucleotide variation is known to affect milk yield and composition in cattle[58]. We have also identified variants in genic regions of genes like *FASN*, *ACACA*, *FADS1*, *DGAT*, *GPAM*, *LPIN1*, *BTN1A1* and *XDH*, which are involved in fatty acid synthesis and milk droplet formation[59–63].

Thus, the present study provides evidence that our criteria for selecting individuals based on contrasting performance for milk production and fertility traits helped identify more number of relevant variants compared to single selection criterion.

Accession number: All sequencing data have been deposited at NCBI SRA database with accession number SRA246917 of bio-project number PRJNA278493.

Competing interests: The authors declare that they have no competing interests.

1. Kierstein, G., Vallinoto, M., Silva, A., Schneider, M. P., Iannuzzi, L. and Brenig, B., Analysis of mitochondrial D-loop region casts new light on domestic water buffalo (*Bubalus bubalis*) phylogeny. *Mol. Phylogenet. Evol*., 2004, **30**, 308–324.
2. Annual Report, Department of Animal Husbandry, Dairying and Fisheries, 2015–16.
3. Dalton, R., No bull: genes for better milk. *Nature*, 2009, **457**, 369.
4. Womack, J. E., Advances in livestock genomics: opening the barn door. *Genome Res*., 2005, **15**, 1699–1705.
5. Cole, J. B. *et al*., Distribution and location of genetic effects for dairy traits. *J. Dairy Sci*., 2009, **92**, 2931–2946.
6. Hirano, T. *et al*., Mapping and exome sequencing identifies a mutation in the IARS gene as the cause of hereditary perinatal weak calf syndrome. *PLOS ONE*, 2013, **8**, e64036.
7. Schmieder, R. and Edwards, R., Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 2011, **27**, 863–864.
8. Li, H. and Durbin, R., Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 2010, **26**, 589–595.
9. Picard: Java-based command-line utilities that manipulate Sam files; http://broadinstitute.Github.lo/picard/
10. Li, H. *et al*., The sequence alignment/map format and samtools. *Bioinformatics*, 2009, **25**, 2078–2079.

11. Danecek, P. *et al.*, The variant call format and VCFtools. *Bioinformatics*, 2011, **27**, 2156–2158.
12. Cingolani, P. *et al.*, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, 2012, **6**, 80–92.
13. Huang da, W., Sherman, B. T. and Lempicki, R. A., Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Proto.*, 2009, **4**, 44–57.
14. Szklarczyk, D. *et al.*, The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*., 2011, **39**, D561–D568.
15. Sabarinathan, R., Tafer, H., Seemann, S. E., Hofacker, I. L., Stadler, P. F. and Gorodkin, J., The RNAsnp web server: predicting SNP effects on local RNA secondary structure. *Nucleic Acids Res*., 2013, **41**, W475–W479.
16. Hu, Z.-L., Park, C. A., Wu, X.-L. and Reecy, J. M., Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. *Nucleic Acids Res*., 2013, **41**, D871–D879.
17. Quinlan, A. R. and Hall, I. M., BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 2010, **26**, 841–842.
18. Reynolds, J., Weir, B. S. and Cockerham, C. C., Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics*, 1983, **105**, 767–779.
19. Krzywinski, M. *et al.*, Circos: an information aesthetic for comparative genomics. *Genome Res*., 2009, **19**, 1639–1645.
20. Yang, W. C. *et al.*, Polymorphisms in the 5′ upstream region of the FSH receptor gene, and their association with superovulation traits in Chinese Holstein cows. *Anim. Reprod. Sci*., 2010, **119**, 172–177.
21. Yang, J., Jiang, J., Liu, X., Wang, H., Guo, G., Zhang, Q. and Jiang, L., Differential expression of genes in milk of dairy cattle during lactation. *Anim. Genet*., 2015.
22. Elsik, C. G., Tellam, R. L. and Worley, K. C., The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*, 2009, **324**, 522–528.
23. Jin, X. *et al.*, An effort to use human-based exome capture methods to analyze chimpanzee and macaque exomes. *PLOS ONE*, 2012, **7**, e40637.
24. Clark, M. J., Chen, R., Lam, H. Y., Karczewski, K. J., Euskirchen, G., Butte, A. J. and Snyder, M., Performance comparison of exome DNA sequencing technologies. *Nature Biotechnol*., 2011, **29**, 908–914.
25. Guo, Y. *et al.*, Exome sequencing generates high quality data in non-target regions. *BMC Genomics*, 2012, **13**, 194.
26. Bainbridge, M. N. *et al.*, Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome Biol*., 2011, **12**, R68.
27. Cochran, S. D., Cole, J. B., Null, D. J. and Hansen, P. J., Discovery of single nucleotide polymorphisms in candidate genes associated with fertility and production traits in Holstein cattle. *BMC Genet*., 2013, **14**, 49.
28. Khatib, H., Monson, R. L., Schutzkus, V., Kohl, D. M., Rosa, G. J. and Rutledge, J. J., Mutations in the STAT5A gene are associated with embryonic survival and milk composition in cattle. *J. Dairy Sci*., 2008, **91**, 784–793.
29. Khatib, H., Maltecca, C., Monson, R. L., Schutzkus, V., Wang, X. and Rutledge, J. J., The fibroblast growth factor 2 gene is associated with embryonic mortality in cattle. *J. Anim. Sci*., 2008, **86**, 2063–2067.
30. Wang, X., Schutzkus, V., Huang, W., Rosa, G. J. and Khatib, H., Analysis of segregation distortion and association of the bovine FGF2 with fertilization rate and early embryonic survival. *Anim. Genet*., 2009, **40**, 722–728.
31. Driver, A. M., Huang, W., Gajic, S., Monson, R. L., Rosa, G. J. and Khatib, H., Short communication: effects of the progesterone receptor variants on fertility traits in cattle. *J. Dairy Sci*., 2009, **92**, 4082–4085.
32. Nyegaard, M. *et al.*, Lack of functional pregnancy-associated plasma protein – a (PAPPA) compromises mouse ovarian steroidogenesis and female fertility. *Biol. Reprod*., 2010, **82**, 1129–1138.
33. Wickramasinghe, S., Rincon, G. and Medrano, J. F., Variants in the pregnancy-associated plasma protein-a2 gene on *Bos taurus* autosome 16 are associated with daughter calving ease and productive life in Holstein cattle. *J. Dairy Sci*., 2011, **94**, 1552–1558.
34. Luna-Nevarez, P. *et al.*, Single nucleotide polymorphisms in the growth hormone-insulin-like growth factor axis in straightbred and crossbred Angus, Brahman, and Romosinuano heifers: population, genetic analyses and association of genotypes with reproductive phenotypes. *J. Anim. Sci*., 2011, **89**, 926–934.
35. Brickell, J. S., Pollott, G. E., Clempson, A. M., Otter, N. and Wathes, D. C., Polymorphisms in the bovine leptin gene associated with perinatal mortality in Holstein–Friesian heifers. *J. Dairy Sci*., 2010, **93**, 340–347.
36. Liu, X., Robinson, G. W., Wagner, K.-U., Garrett, L., Wynshaw-Boris, A. and Hennighausen, L., STAT5A is mandatory for adult mammary gland development and lactogenesis. *Genes Dev*., 1997, **11**, 179–186.
37. Killeen, A. P., Morris, D. G., Kenny, D. A., Mullen, M. P., Diskin, M. G. and Waters, S. M., Global gene expression in endometrium of high and low fertility heifers during the mid-luteal phase of the estrous cycle. *BMC Genomics*, 2014, **15**, 234.
38. Chalmel, F. *et al.*, The conserved transcriptome in human and rodent male gametogenesis. *Proc. Natl. Acad. Sci. USA*, 2007, **104**, 8346–8351.
39. Thomas, M., Enns, R., Shirley, K., Garcia, M., Garrett, A. and Silver, G., Associations of DNA polymorphisms in growth hormone and its transcriptional regulators with growth and carcass traits in two populations of Brangus bulls. *Genet. Mol. Res*., 2007, **6**, 222–237.
40. Ferraz, J. *et al.*, Association of single nucleotide polymorphisms with carcass traits in Nellore cattle. *Genet. Mol. Res*., 2009, **8**, 1360–1366.
41. Nkrumah, J. D., Li, C., Yu, J., Hansen, C., Keisler, D. H. and Moore, S. S., Polymorphisms in the bovine leptin promoter associated with serum leptin concentration, growth, feed intake, feeding behavior, and measures of carcass merit. *J. Anim. Sci*., 2005, **83**, 20–28.
42. Bionaz, M. and Loor, J. J., Gene networks driving bovine milk fat synthesis during the lactation cycle. *BMC Genomics*, 2008, **9**, 366.
43. Bionaz, M. and Loor, J. J., ACSL1, AGPAT6, FABP3, LPIN1, and Slc27a6 are the most abundant isoforms in bovine mammary tissue and their expression is affected by stage of lactation. *J. Nutr*., 2008, **138**, 1019–1024.
44. Pedersen, J. S., Forsberg, R., Meyer, I. M. and Hein, J., An evolutionary model for protein-coding regions with conserved RNA structure. *Mol. Biol. Evol.*, 2004, **21**, 1913–1922.
45. Tian, E. *et al.*, Allelic mutations in noncoding genomic sequences construct novel transcription factor binding sites that promote gene overexpression. *Genes, Chromosomes Cancer*, 2015, **54**, 692–701.
46. Yamazaki, J. *et al.*, Tet2 mutations affect non-CPG island DNA methylation at enhancers and transcription factor binding sites in chronic myelomonocytic leukemia. *Cancer Res*., 2015.
47. Buratti, E. and Baralle, F. E., Influence of RNA secondary structure on the pre-mRNA splicing process. *Mol. Cell. Biol*., 2004, **24**, 10505–10514.

48. Rudolph, M. C. *et al.*, Metabolic regulation in the lactating mammary gland: a lipid synthesizing machine. *Physiol. Genomics*, 2007, **28**, 323–336.
49. Ding, X. *et al.*, A novel single nucleotide polymorphism in exon 7 of LPL gene and its association with carcass traits and visceral fat deposition in yak (*Bos grunniens*) steers. *Mol. Biol. Rep*., 2012, **39**, 669–673.
50. Van Horn, C. G., Caviglia, J. M., Li, L. O., Wang, S., Granger, D. A. and Coleman, R. A., Characterization of recombinant long-chain rat acyl-CoA synthetase isoforms 3 and 6: identification of a novel variant of isoform 6. *Biochemistry*, 2005, **44**, 1635–1642.
51. Mashek, D. G. and Coleman, R. A., Cellular fatty acid uptake: the contribution of metabolism. *Curr. Opin. Lipidol*., 2006, **17**, 274–278.
52. Mercade, A. *et al.*, Characterization of the porcine acyl-CoA synthetase long-chain 4 gene and its association with growth and meat quality traits. *Anim. Genet.*, 2006, **37**, 219–224.
53. Schwehm, J. M., Kristyanne, E. S., Biggers, C. C. and Stites, W. E., Stability effects of increasing the hydrophobicity of solvent-exposed side chains in staphylococcal nuclease. *Biochemistry*, 1998, **37**, 6939–6948.
54. Manjithaya, R. R. and Dighe, R. R., The 3′ untranslated region of bovine follicle-stimulating hormone β messenger RNA downregulates reporter expression: involvement of Au-rich elements and transfactors. *Biol. Reprod*., 2004, **71**, 1158–1166.
55. Rao, Y. S., Wang, Z. F., Chai, X. W., Nie, Q. H. and Zhang, X. Q., Relationship between 5′ UTR length and gene expression pattern in chicken. *Genetica*, 2013, **141**, 311–318.
56. Cohen-Zinder, M. *et al.*, Identification of a missense mutation in the bovine *ABCG2* gene with a major effect on the qtl on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Res*., 2005, **15**, 936–944.
57. Robenek, H. *et al.*, Butyrophilin controls milk fat globule secretion. *Proc. Natl. Acad. Sci. USA*, 2006, **103**, 10385–10390.
58. Gilchrist, E. J., Sidebottom, C. H., Koh, C. S., Macinnes, T., Sharpe, A. G. and Haughn, G. W., A mutant *Brassica napus* (canola) population for the identification of new genetic diversity via tilling and next generation sequencing. *PLOS ONE*, 2013, **8**, e84303.
59. Pannier, L., Mullen, A. M., Hamill, R. M., Stapleton, P. C. and Sweeney, T., Association analysis of single nucleotide polymorphisms in DGAT1, TG and FABP4 genes and intramuscular fat in crossbred *Bos taurus* cattle. *Meat Sci*., 2010, **85**, 515–518.
60. Macciotta, N. P. *et al.*, Association between a polymorphism at the stearoyl CoA desaturase locus and milk production traits in Italian Holsteins. *J. Dairy Sci*., 2008, **91**, 3184–3189.
61. Huang, W., Penagaricano, F., Ahmad, K. R., Lucey, J. A., Weigel, K. A. and Khatib, H., Association between milk protein gene variants and protein composition traits in dairy cattle. *J. Dairy Sci*., 2012, **95**, 440–449.
62. Greene, E. A. *et al.*, Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in arabidopsis. *Genetics*, 2003, **164**, 731–740.
63. Maningat, P. D. *et al.*, Gene expression in the human mammary epithelium during lactation: the milk fat globule transcriptome. *Physiol. Genomics*, 2009, **37**, 12–22.

ACKNOWLEDGEMENTS. This study was funded by the Department of Biotechnology, Ministry of Science and Technology, Government of India (BT/PR13605/AAQ/01/431/2010).

Received 28 June 2016; revised accepted 9 November 2016

doi: 10.18520/cs/v112/i06/1230-1239

Erratum

Should Indian researchers pay to get their work published?

Muthu Madhan, Siva Shankar Kimidi, Subbiah Gunasekaran and Subbiah Arunachalam
[*Curr. Sci.*, 2017, **112**(4), 703–713]

We inadvertently missed to present the total number of papers written by Indian researchers in the first sentence under the Discussion section (page 707). Please read the sentence as ‘Over 14.4% (or 37,122) of the 256,822 papers from India as seen from SCIE have been published in OA journals’.

We regret the error.

– Authors