Genomics 112 (2020) 3571–3578

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Original Article

# Identification and characterization of trait-specific SNPs using ddRAD sequencing in water buffalo

D.C. Mishra[a,1], Poonam Sikka[b], Sunita Yadav[a,1], Jyotika Bhati[a], S.S. Paul[b], A. Jerome[b], Inderjeet Singh[b], Abhigyan Nath[b], Neeraj Budhlakoti[a], A.R. Rao[a], Anil Rai[a], K.K. Chaturvedi[a,*]

[a] ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India
[b] ICAR-Central Institute for Research on Buffaloes, Hisar, India

ARTICLE INFO

*Keywords:*
Buffalo
ddRAD
SNPs
Annotation
GO terms
Synonymous SNP
Non-synonymous SNP

ABSTRACT

Single Nucleotide Polymorphism (SNP) is one of the important molecular markers widely used in animal breeding program for improvement of any desirable genetic traits. Considering this, the present study was carried out to identify, annotate and analyze the SNPs related to four important traits of buffalo viz. milk volume, age at first calving, post-partum cyclicity and feed conversion efficiency. We identified 246,495, 168,202, 74,136 and 194,747 genome-wide SNPs related to mentioned traits, respectively using ddRAD sequencing technique based on 85 samples of Murrah Buffaloes. Distribution of these SNPs were highest (61.69%) and lowest (1.78%) in intron and exon regions, respectively. Under coding regions, the SNPs for the four traits were further classified as synonymous (4697) and non-synonymous (3827). Moreover, Gene Ontology (GO) terms of identified genes assigned to various traits. These characterized SNPs will enhance the knowledge of cellular mechanism for enhancing productivity of water buffalo through molecular breeding.

## 1. Introduction

In developing countries including India, water buffalo (*Bubalus bubalis*) plays a significant role in agricultural economy by contributing to milk, meat and drought power [1,2]. Recently, world's buffalo population was estimated as 194 million with more than 97% of the total population present in Asia [3]. Globally, India leads in buffalo husbandry followed by Pakistan and China with India producing 70% of the world's total buffalo milk production followed by Pakistan, China, Egypt [4]. In India, buffalo contributes to more than 50% of total milk production and importantly, water buffalo milk possesses higher fat contents than dairy cattle milk [2], specifically saturated fatty acids. Water buffalo can also survive on poor quality roughage, better adapted to harsher environments and are resistant to several bovine tropical diseases [1]. However in buffalo, the average annual milk production is lower than cattle in spite of similar genomes (99.14% similarity). Recently, studies have been carried out in buffalo genome with respect to mining of genes as well as chromosome-level genome assembly depicting various structural variants including SNPs [5,6].

These SNPs are sequence variations genome occurring when a single nucleotide in the genome is altered and they are potential genetic markers [7]. These bi-allelic SNP markers are more stable as compared to other genetic markers viz. simple sequence repeat (SSR), restriction fragment length polymorphism (RFLP), amplified fragment length polymorphism (AFLP) and random amplification of polymorphic dna (RAPD). Identification and characterization of SNP markers related to production and reproduction traits can be used for marker assisted breeding which may pave way for higher productivity of buffaloes [8–10]. Hence, unravelling genome-wide SNPs has an important role in identification of high merit germplasm for ongoing as well future breeding program in buffalo species. Identification of genome-wide SNP markers for the traits viz. milk volume, age at first calving, post-partum cyclicity and feed conversion efficiency requires selective genotyping [11]. Although analysis of large number of genome-wide SNP markers across multiple samples is time-consuming and laborious, sophisticated bioinformatics methods are now available to address these limitations [12]. Also, whole genome sequencing (WGS) to identify variants is costly; however, low input techniques such as Restriction site–associated DNA sequencing (RAD-seq / ddRAD-Seq / ddRAD) [13,14] is available as alternative, which potentially covers more than 40% of the genome [15,16] and are proven useful for selective genotyping based on viable phenotype recording systems. Thus, ddRAD-Seq [17] involves

[*] Corresponding author at: ICAR-Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi 110012, India.
*E-mail address:* kk.chaturvedi@icar.gov.in (K.K. Chaturvedi).
[1] Contributed equally.

https://doi.org/10.1016/j.ygeno.2020.04.012
Received 19 September 2019; Received in revised form 19 March 2020
Available online 19 April 2020

the replacement of fragment shearing by second restriction digestion which will enhance the accuracy of the size-selection and allows combinatorial indexing, thereby skimming through the robust genomes at low cost [12].

Considering this, an attempt was made to generate Murrah buffalo specific ddRAD data based on selectively genotyped animals for important traits i.e. milk volume, age at first calving, post-partum cyclicity and feed conversion efficiency. Based on this investigation, SNPs were identified, characterized, annotated and classified based on their classes. Thus, identification of trait specific SNPs shall be useful for molecular breeding to achieve the faster genetic gain in buffalo species.

## 2. Material and methods

### 2.1. Animal resource and generation of genotyping data

Selection of individuals for implementing improved mating plans is based on specific performance criteria of mates for genetic improvement and productivity enhancement of any breed. These performance criteria are phenotype determinants needed for selective genotyping to map nucleotide markers to achieve higher gain through early selection. At ICAR-Central Institute for Research on Buffaloes, Hisar, Haryana, India, genetic improvement of the Murrah buffalo is carried out through progeny testing for the past two decades and the pedigree and performance records (production and reproduction) are maintained. These specific criteria are phenotype determinants which were used for selective genotyping. Individual animals were selected from unrelated pedigree and extreme performance levels for complex traits viz. milk volume (milk yield), age at first calving, post-partum cyclicity and feed conversion efficiency. Animals for genotyping were selected on the basis of phenotypes ensuring wide variation over sire, parity of animals, best milk yield in whole production span, milk constituents, lactation length, subject to following performance criteria (Table 1).

Based on the parity records, sixty three un-related multi-parous buffaloes were selected from high and low performance categories for the traits under study i.e. milk volume (milk yield) ($n$ = 25), age at first calving ($n$ = 19) and post-partum cyclicity resumption (n = 19). To select animals for high and low feed conversion efficiency [heritability and repeatability as 0.15 and 0.53 reportedly], a feed trial for 3 months was conducted with 42 female heifers/calves (aged 10 to 12 months) under farm management to determine the comparative levels of residue feed intake [RFI] and energy utilizing animals for growth. In this investigation, body weight, feed intake, and actual dry matter intake per unit body weight gain were recorded. Based on this feed trial, it was evident that residual energy intake as promising determinant to select high feed convergent buffaloes. Accordingly, animals ($n$ = 22; High RFI = 11 and low RFI = 11) were identified for selective genotyping w.r.t. feed conversion efficiency trait. Selected genotypes from animals ($n$ = 85) for all 4 traits were bled through veni-puncture to isolate DNA following Phenol-Chloroform method [18] for SNP genotyping. Isolated DNA samples were subjected to next generation sequencing using Hi-Seq Illumina sequencing platform. Reduced representational sequence [Double digestion RAD] analysis was adopted to identify genome wide SNP markers.

The methodological workflow is depicted in Fig. 1. Reduced representational sequence (ddRAD sequencing) of genome data, generated over 85 samples were checked for quality parameters (Table S1) followed by trimming of these quality sequences using Trimmomatic [19]. These sequences were further aligned to the available reference genome i.e. *Bubalus bubalis* **(**UOA_WB_1) [6] and indexing has been done by using Bowtie [20]. The resulting aligned SAM files were converted into BAM format from *samtools view* followed by sorting using *samtools sort* [21]. The resultant sorted alignment/map (BAM) files were then exported to STACKS pipeline [22].

The wrapper program ref_map.pl was used to create the catalog of RAD loci for SNP identification with three components viz. *ustacks* (building loci), *cstacks* (creating a catalog of loci) and *sstacks* (matching samples back against the catalog using the standard parameters. First, sequences aligned to the same genomic location were stacked together and merged to form loci. Loci with a sequencing depth of three or more reads per individual were retained and catalogues have been created. SNPs at each locus were selected using a maximum likelihood framework. Population program of stacks pipeline (population.pl) was used to process all the SNP data across the individuals. Haplotypes in all chromosomes with respect to all four traits are also identified by the stacks software. Genetic diversity was estimated through Tassel software by using important genetic parameters viz. major and minor allele frequency, both observed and expected heterozygosity as well as homozygosity in addition to nucleotide diversity (π) [23].

### 2.2. Annotation and functional classification

The output of population program of the STACKS pipeline were in variant calling format (vcf) files. These vcf files were utilized for annotations of SNP effects on gene functions using SnpEff [24]. The Venn diagrams were created using Venny software [25] to depict the overlapping genes. The genes of the variants are converted from the ensemble format into Uniprot gene Id format by using DAVID software [26]. The functional classification of the genes was carried out by using agriGo tool by using FDR value 0.05 [27] based on UniProt gene Id received. The agriGo tool maps genes to function according to GO categories: molecular function, cellular component and biological process. The plot of GO classification was drawn by using Blast2GO [28]. The graphical mapping was done by using CIRCOS (version 0.69) visualization tool [29] for chromosome-wise SNP distribution. This distribution helps in understanding the position of SNPs under synonymous and non-synonymous classes.

## 3. Results

### 3.1. SNP distribution

In this study, total of 683,580 putative SNPs were obtained from 85 animals based on genomic regions variation with respect to four traits of buffalo. Total number of identified haplotypes was 191,318. Trait-wise distribution of genome-wide SNPs were 246,495, 168,202, 74,136

**Table 1**
Details of traits considered for animal selection.

| Traits | Milk volume | Age at first calving | Post-partum cyclicity | Feed conversion efficiency |
|---|---|---|---|---|
| Phenotype determinant | per lactation (kg) | Period between date of birth and first calving (months) | Period between calving & next conception | RFI |
| Higher limit | > 3000 Kg | < 40 months | < 70 days | −0.437 |
| Lower limit | < 1800 Kg | > 55 months | > 200 days | 0.359 |
| Pre-determined/ Experimented | Based on phenotype records | Based on phenotype records | Based on phenotype records | Based on feed trial conducted |
| No. of high performing animals | 11 | 8 | 11 | 11 |
| No. of low performing animals | 14 | 11 | 8 | 11 |

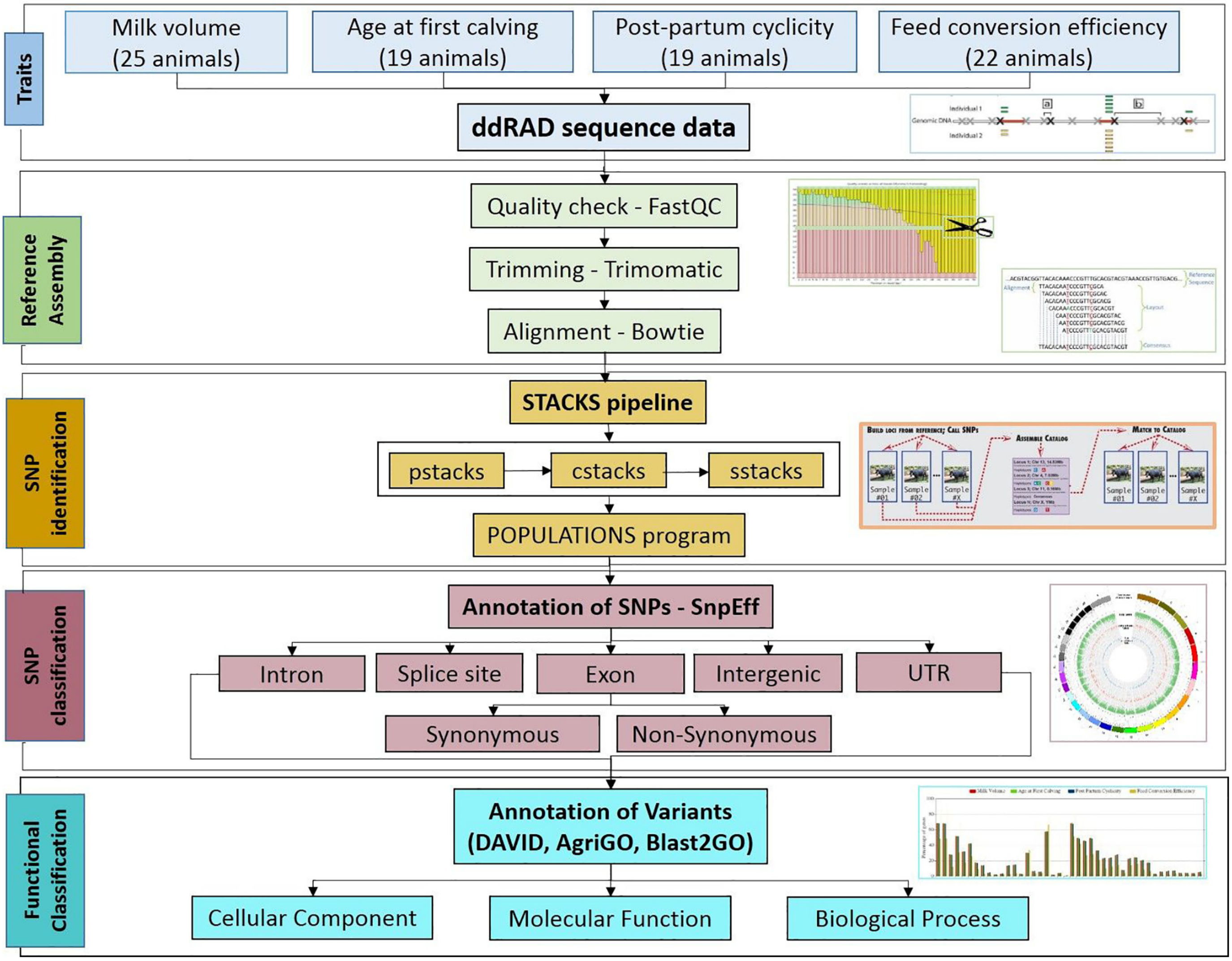

**Fig. 1.** Schematic diagram of workflow.

and 194,747 for milk volume, age at first calving, post-partum cyclicity and feed conversion efficiency, respectively (Table 2). Chromosome-wise distribution of SNP is shown in Fig. S1 for traits under study. Occurrence of SNPs per chromosome varied from 1276 to 17,902, considering autosomes ($n = 24$) and sex chromosomes ($n = 1$) with respect to buffalo genome as reference. Maximum numbers of SNPs were observed on chromosomes 1 and the trait wise count was 17,902, 12,107, 5346 and 14,238 for milk volume, age at first calving, post-partum cyclicity and feed conversion efficiency, respectively. Likewise, minimum number of SNPs were observed on chromosome 24 with 4419, 2935, 1276 and 3376 SNPs attributed for milk volume, age at first calving, post-partum cyclicity and feed conversion efficiency, respectively showing significant difference with respect to SNPs distribution on the chromosome (Fig. S2; Supplementary file S1).

The synonymous (4697) and non-synonymous (3827) SNP identified for all traits is shown in Fig. S3. Trait-wise distribution of SNPs classified as synonymous and non-synonymous SNPs were mapped on water buffalo genome using CIRCOS map (Fig. 2). It depicts the comparative view of chromosome wise distribution of identified SNPs (total, synonymous and non-synonymous) in various traits. More synonymous SNPs were observed than non-synonymous SNPs in all the traits, except post-partum cyclicity. Furthermore, region wise location of SNPs were higher in intronic region (61.69%) followed by intergenic (17.71%), upstream (8.88%), downstream (8.70%) and exons (1.78%). It was also deduced that low occurrence of SNPs was observed in other regions (3'UTR, 5'UTR, Splice Site Region Splice Site Donor and Splice Site Acceptor) (Fig. 3).

In addition, SNPs identified over different genomic regions were classified into four impact classes viz. 99.01% in modifier (upstream, downstream, intergenic, UTR), 0.53% as low (synonymous coding/start/stop, start gained), 0.41% as moderate (non-synonymous) and 0.05% as high (affecting splice-sites, stop and start codons) class impact classifier (Fig. S4; Table S2).

### 3.2. Functional classification of SNPs

Substitution is a functional class of mutations which alters the function of genome during cellular processes. Silent, mis-sense and non-sense are the three types of single base-pair substitution mutation that occur in any genome. In this study, the mis-sense mutations were found highest in post-partum cyclicity trait (79.31%) followed by milk volume (75.37%), feed conversion efficiency (73.01%) and age at first calving (70.80%). But, silent mutations were high in age at first calving (24.63%), followed by milk volume (21.77%), feed conversion efficiency (21.45%) and post-partum cyclicity (17.30%). The occurrence of non-sense mutations were higher in feed conversion efficiency (5.54%) followed by age at first calving (4.57%), post-partum cyclicity (3.39%) and milk volume (2.86%) (Fig. 4).

**Table 2**
Total number of SNP and haplotypes identified for the economic traits under study.

| Traits | Total number of raw reads | Number of SNPs | Ratio of Total number of raw reads and Number of SNPs | Number of haplotypes |
|---|---|---|---|---|
| Milk volume | 40,458,845 | 246,495 | 164.14 | 65,693 |
| Age at first calving | 28,338,931 | 168,202 | 168.48 | 49,422 |
| Post-partum cyclicity | 26,372,976 | 74,136 | 355.74 | 27,538 |
| Feed conversion efficiency | 32,049,707 | 194,747 | 164.57 | 48,665 |

### 3.3. Homozygous and heterozygous SNP count per sample

In case of bi-allelic SNPs, if both alleles have same SNPs they are termed homozygous, and if different they are known as heterozygous SNPs. The frequency of homozygous and heterozygous SNPs per sample in four different traits is shown in Fig. S5. Milk volume trait had the maximum number of homozygous (839619) and heterozygous SNPs (138743). In contrast, post-partum cyclicity and feed conversion efficiency trait showed minimum number of homozygous (253058) and heterozygous SNPs (31249). Occurrence of heterozygous SNPs and their proportion differentiate the low and high performing animals with respect to different traits as evident in this study (Table S3).

### 3.4. Structural and functional annotation of identified SNPs

Structural annotation of identified SNPs resulted in 40251 annotated and 136821 unannotated genes. Annotated genes were further classified into various functional classes (Supplementary file S2) and a Venn diagram was plotted to depict the common genes belonging to all four traits under study (**Fig. 5**). Based on the analysis, 5149 genes were common in all four traits and 1899 genes found common in milk volume, age at first calving and feed conversion efficiency traits. There are

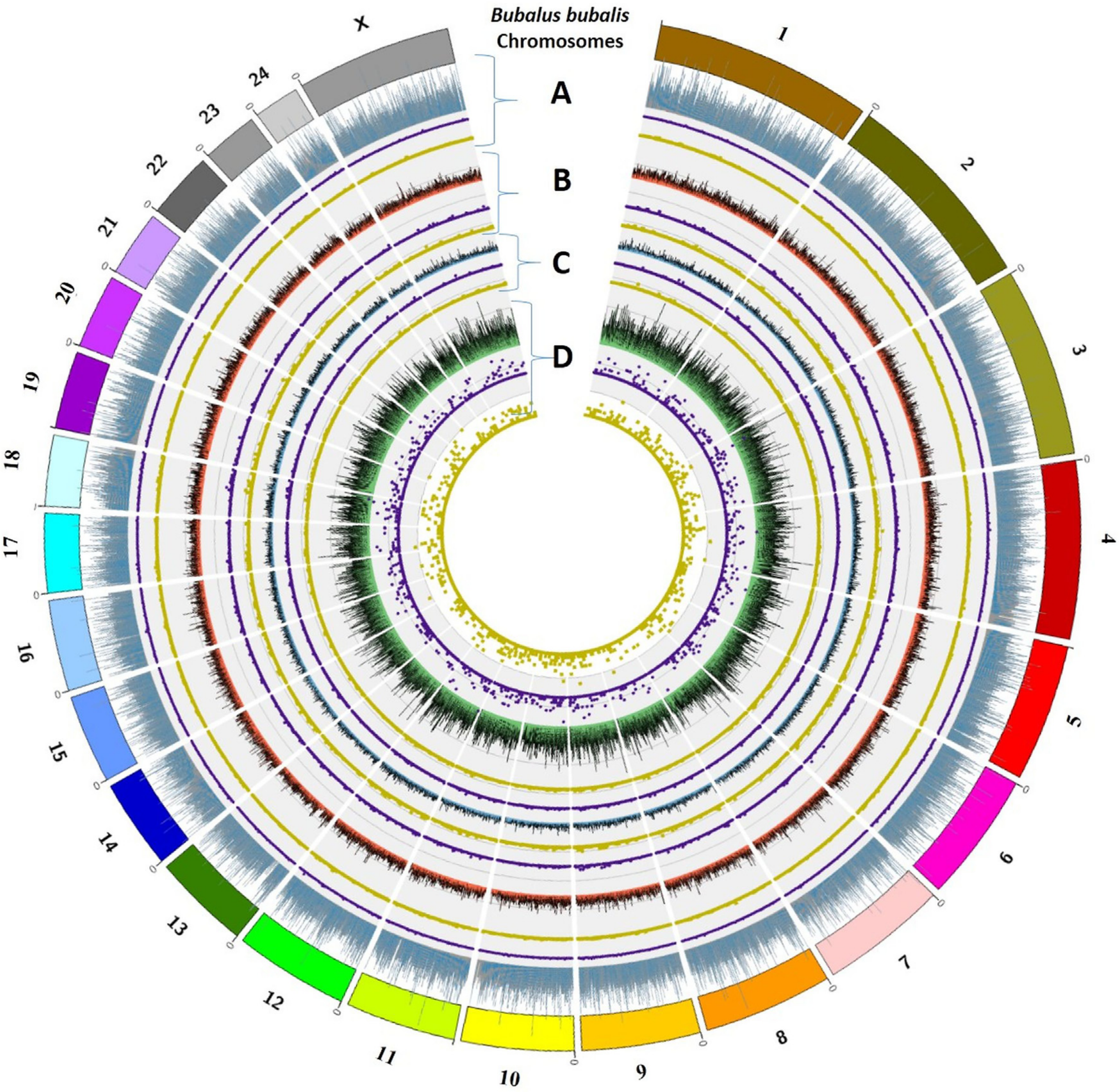

**Fig. 2.** Trait-wise SNP distribution of total, synonymous and non-synonymous (shown in different colors and types. Outer circle depicts total SNPs, purple and scattered circle, synonymous and yellow and scattered rectangle depict non-synonymous SNPs) for four different traits: milk volume, age at first calving, post-partum cyclicity and feed conversion efficiency shown as A, B, C and D respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

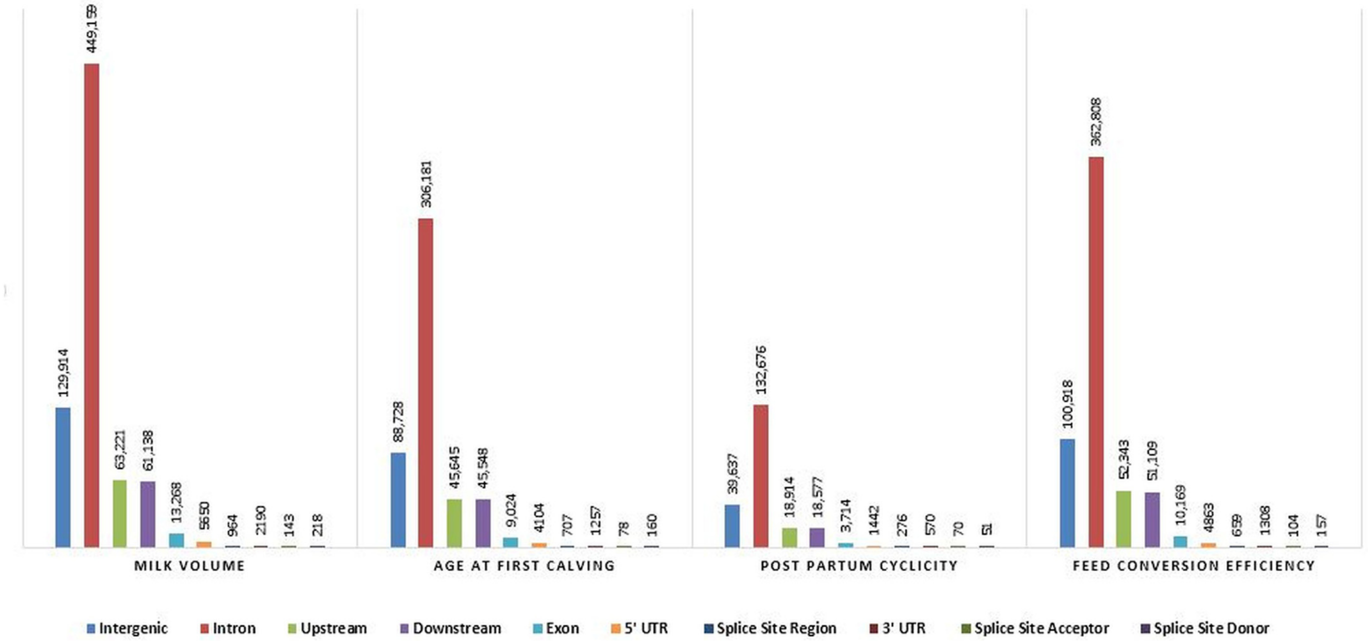

**Fig. 3.** Number of SNP (Variant) effects w.r.t. to different genomic region (intergenic, introns, exons, untranslated region (5' UTR and 3' UTR), splice site] for milk volume, age at first calving, post-partum cyclicity, and feed conversion efficiency based on their position in the annotated *Bubalus bubalis* genome.

1099 genes common between milk volume and feed conversion efficiency trait, while 819 were between milk volume and age at first calving trait. There were 1504, 765, 226 and 709 unique genes found in milk volume, age at first calving, post-partum cyclicity and feed conversion efficiency trait, respectively. Gene Ontology (GO) terms of identified genes assigned to milk volume, for age at first calving, for post-partum cyclicity and for feed conversion efficiency traits were 2405, 1356, 589 and 1725, respectively (Supplementary file S3) traits. Significant GO terms (FDR value ≤ 0.05), classified as cellular components, molecular functions and biological processes through GO

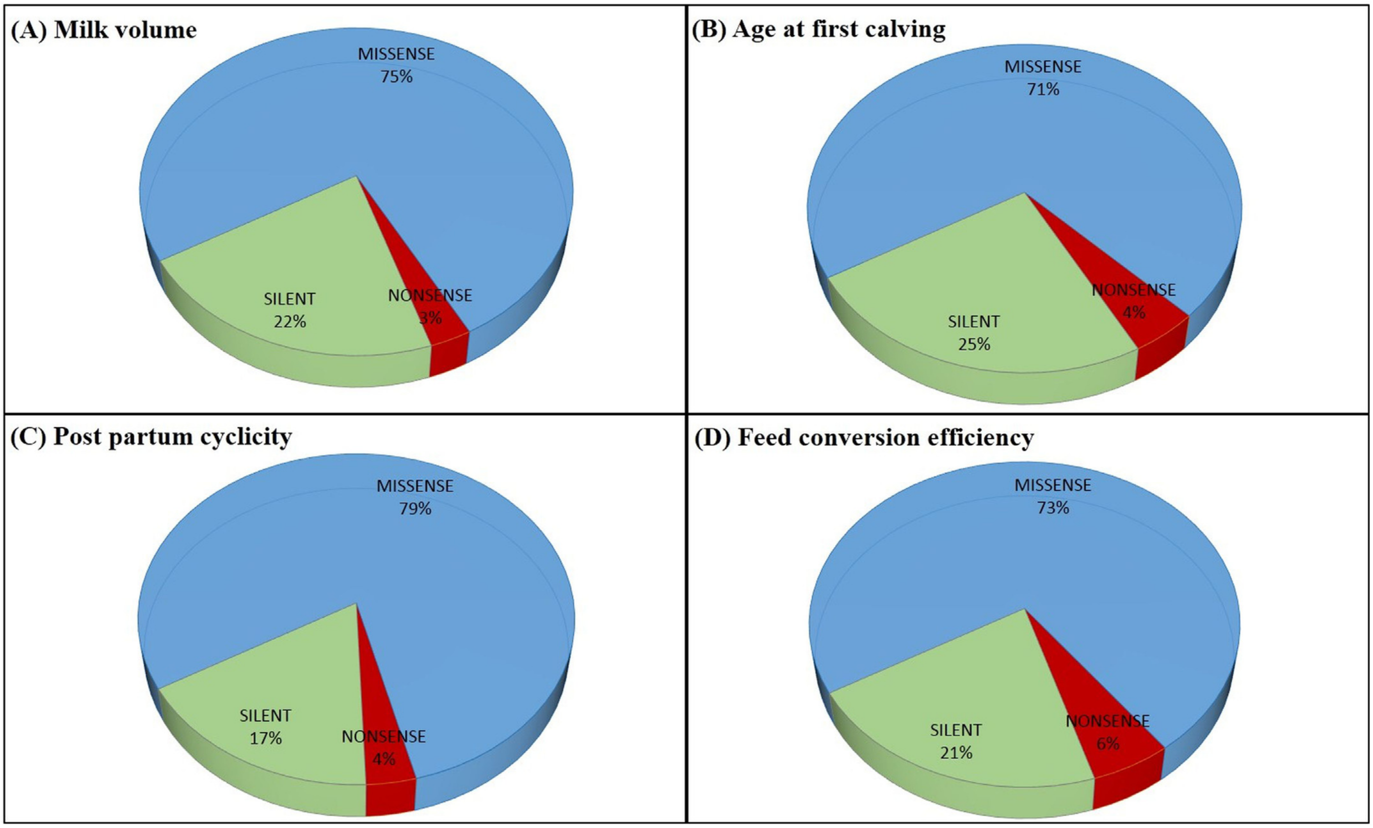

**Fig. 4.** Number of SNP effects by functional class. These SNP effects were categorized by functional class as missense, nonsense and silent substitution w.r.t. milk volume, age at first calving, post-partum cyclicity and feed conversion efficiency traits depicted as (A), (B), (C) and (D) respectively.

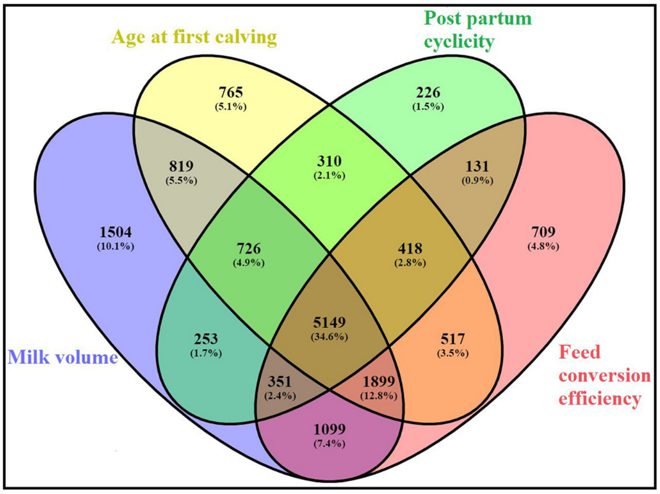

**Fig. 5.** Venn diagram showing different types of trait specific overlapping annotated genes.

enrichment analysis is shown in Fig. 6.

Protein binding (GO:0005515) was the most represented GO term in buffalo molecular functions, whereas the most abundant biological processes were organic substance metabolic process (GO:0006082), primary metabolic process (GO:0030258) and cellular metabolic process (GO:0044237). Majority of the cellular components were found in organelle (GO:0043226) and intracellular organelle (GO:0043229) (Fig. 6; Supplementary file S3).

## 4. Discussion

Selection based on pedigree has played an important role for selective breed improvement for domestic animals including buffalo. In this study, we identified and characterized genome-wide SNPs with respect to four selected traits viz. milk volume, age at first calving, post-partum cyclicity and feed conversion efficiency which contributes to the overall life-time productivity in buffalo. Based on this investigation, number of SNPs identified from each trait was comparable (Table 2); however, maximum number of the SNPs is found in milk volume trait followed by age at first calving and feed conversion efficiency traits. An earlier study in Murrah buffalo ($n = 4$) by restriction enzyme method [9] resulted in the identification of lower number of SNPs (168048) as compared to this study which could be attributed to higher gene pool of animals ($n = 85$) in this study and hints that the accuracy of identified SNP with the increase in sample size [30]. Furthermore, use of STACKS software as compared to PRINSEQ [31] for filtering and identification of the SNPs as well as haplotypes could have contributed to difference in SNPs identified. STACKS software is the most suitable tool for the RAD and ddRAD sequencing data analysis which checks the mean quality score across windows (by default 15 bp); whereas PRINSEQ [31] calculates quality score across the whole reads resulting in higher read loss. In this study, higher level of homozygosity emerged out of higher mis-sense SNP effects identified with post-partum cyclicity trait indicate that the negative selection pressure is delegated by non-synonymous base substitution to reduce genetic variation. It is also evident that more allele sharing among individual genotypes normalizes the genome towards natural selection, especially, with respect to post-partum cyclicity trait thus improving productive life [32].

Further, characterization of related genes with the study traits showed that the SNPs were associated with non-coding RNAs like snoRNA, miRNA, snRNA and few protein coding elements (Supplementary file S2). The spliceosomal machinery i.e. pre-RNA splicing [33] of eukaryotes might play a catalytic role by acting as regulatory switches of cellular processes at modifier levels as observed in cattle [34]. The present study shows that variant effects in non-coding region is maximum (61.69%) as these are located at RNA-producing regions (non-coding transcripts exons and non-coding transcripts) and considerably influence several biological activities viz. post-transcriptional processing, translational initiation and gene expression through the alteration of RNA folding, RNA-binding proteins recognition motif and microRNA binding. Functional class of SNPs observed for the traits, were identified on genes encoding for polymerases, zinc-fingers, solute carrier, trans-membrane protein, protein kinases, and defense related proteins. These genes play active role in governing milk volume, age at first calving, post-partum cyclicity and feed conversion efficiency traits [5,35]. In addition, these genes govern several cellular functions of gene regulation, including genome stability through poly-adenylation due to spliceosomal components as they encode for enzymes viz. lipase, nuclease, kinase, carrier proteins and polymerases [36].

Genes affected by SNP variants, based on estimated FDR ($< 0.05$), showed that key genes [inositol 1,4,5-trisphosphate receptor type 2

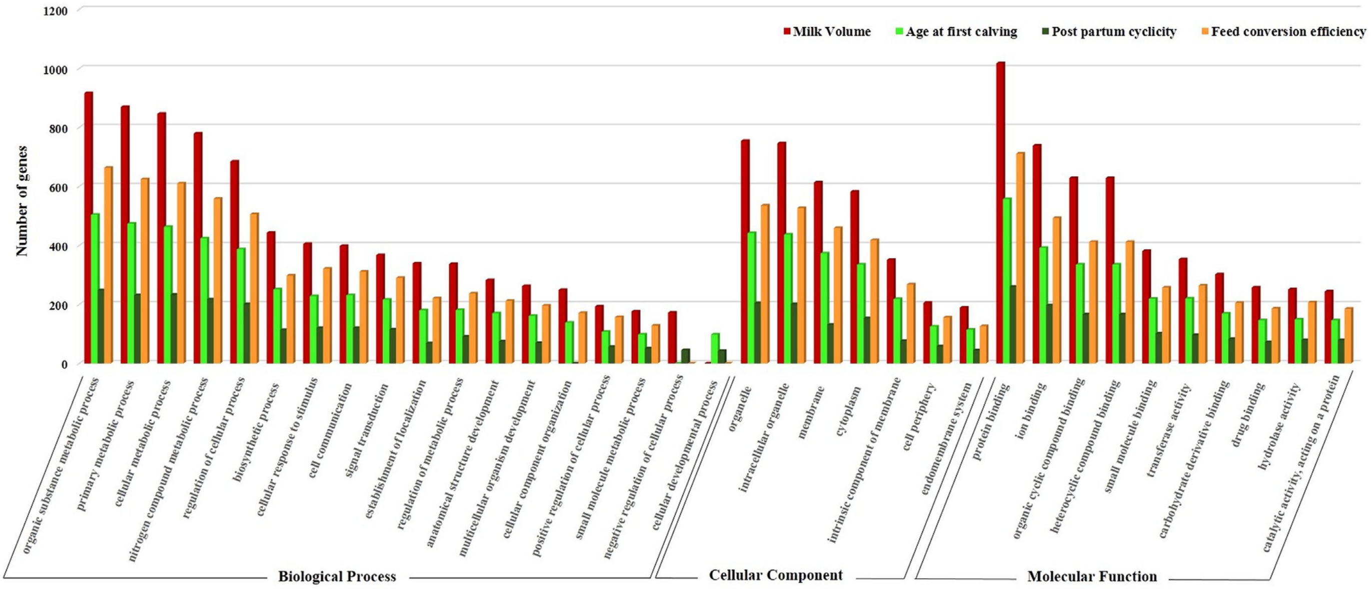

**Fig. 6.** Classification of identified genes in cellular component, molecular function and biological processes.

(ITPR2) and branched chain amino acid transaminase 1 (BCAT1)] were associated with all traits (Supplementary file S2) [37]. Our findings showed that the candidate genes [phospholipase C epsilon 1(PLCE1), protein kinase C epsilon (PRKCE), growth hormone receptor (GHR) and disco interacting protein 2 homolog A (DIP2A)] associated with milk yield and fat harboured higher number of SNPs [9] (Supplementary file S2). In addition, other SNP associated genes [contactin associated protein 2 (CNTNAP2), discs large MAGUK scaffold protein 2 (DLG2), mono-ADP ribosylhydrolase 2 (MACROD2), kazrin periplakin interacting protein (KAZN), CD14 molecule (CD14) and BCL2 like 12 (BCL2L12)] related with milk volume were involved in histone modification, epidermal differentiation, cell adhesion, cytoskeletal organization hinting their role in milk synthesis [38]. Also, it is noteworthy that we failed to identify milk synthesis governing genes [mediator complex subunit 28 (MED28) and vasoactive intestinal peptide receptor 2 (VIPR2)] as reported in cattle [37,39] which opens new avenue for investigation of plausible alternate pathways of milk synthesis in buffalo species.

Pertinent to age at first calving, the key responsive genes were RNA binding Fox 1 (RBFox1), cell adhesion molecule and neurexin receptors (CNTNAP2), transmembrane protein3 (TENM3), discs large MAGUK scaffold protein 2 (DLG2), ATP/GTP binding cytosolic carboxy-peptidase (AGBL4), neuraxin-3-beta (NRXN3), protein tyrosine phosphatase family-signaling molecule (PTPRD), dystrophin protein (DMD) and RNA binding protein (RBMS3) (Supplementary file S3). These genes play dominant role in cellular growth, differentiation, mitotic cycle, adhesion, DNA replication, RNA transcription and apoptosis [40].

Likewise, the most common genes associated with fertility traits viz. age at first calving and post-partum cyclicity include SRY-box transcription factor 5/6 (SOX5/SOX6), calcium voltage-gated channel auxiliary subunit beta 2 (CACNB2), zinc finger protein 521(ZNF521), progesterone receptor (PGR), GRAM domain containing 1B (GRAMD1B), calcium voltage-gated channel subunit alpha1 D (CACNA1D), glutamate ionotropic receptor AMPA type subunit 3 (GRIA3) and thymocyte selection associated high mobility group box (TOX) identified through gene enrichment analysis with FDR $<$ 0.05. Associated SNPs count with respect to age at first calving and post-partum cyclicity were 109 and 25, respectively located as modifier, downstream/upstream gene and intron effect (Supplementary file S2). It is evident that the above mentioned genes govern fertility and the identified SNPs (88%) were located in the non-coding regions hinting their possible role in gene regulatory processes i.e. transcription to post-translation modifications [9]. Other linked genes with post-partum cyclicity were myosin heavy chain 14 (MYH14), cell adhesion molecule and neurexin receptors (CNTNAP2), mono-ADP ribosylhydrolase 2 (MACROD2) and kazrin periplakin interacting protein (KAZN) which play predominant role in growth and immunity of cattle [41].

This study also identified key genes [dystrophin (DMD), discs large MAGUK scaffold protein 2 (DLG2) and propionyl-CoA carboxylase subunit alpha (PCCA)] associated with feed conversion efficiency. SNPs were detected in the key genes [non-SMC condensin I complex subunit G (NCAPG), ligand dependent nuclear receptor corepressor like (LCORL) and regulatory associated protein of MTOR complex 1 (RPTOR)] which regulate cell growth, energy homeostasis, apoptosis and immune response in young heifers, thereby suggesting their role in feeding efficiency [42–44]. In contrast to previous reports, this study failed to identify genes [leucine rich repeats and IQ motif containing 3 (LRRIQ3), CXADR like membrane protein (CLMP), protein phosphatase 2 scaffold subunit A alpha (PPP2R1A), aldehyde oxidase 1 (AOX1) and insulin-like growth factor binding protein 6 (IGFBP6)] reported for feed conversion efficiency in cattle [45–48] which needs further investigation.

Enriched GO terms identified in this study, were related to milk production, reproduction, immune response and resistance/susceptibility to infectious diseases. Expectedly, the present results confirmed that the individuals under study experienced selective pressure for these specific traits. This is ascertained by the appearance of overlapping genomic regions within buffalo population due to selection and adaptability in addition to their temporal regulation of expression. With respect to milk volume, several GO terms were related to glycoprotein, fatty acid, glycerolipid and sterol biosynthesis, in addition to other biological process viz. hexose metabolic processes, oxidative stress, calcium transport, divalent metal ion transport, calcium channel activity, acetyl transferase activity and mRNA processing, thus confirming their significant role in lactogenesis. With respect to physiological processes, the identified genes were classified under common GO terms, depicting regulation biological quality (GO:0065008), apoptosis (GO:0016265), cellular component organization (GO:0006915), cell enzyme activity or gene expression to stimulus (GO:0051716), molecular functions' regulation (GO:0016043) and cell death (GO:0008219), related to age at first calving and feed conversion efficiency indicating the higher level of organelle reorganization and cell growth required by the study traits. With milk volume and feed conversion efficiency trait, prominent GO term involved were in response to stimulus (GO:0050896), regulation to biological quality (GO:0065008) and molecular function (GO:0065009). Macromolecular localization process (GO:0033036) was prominent in milk volume and age at first calving traits (Supplementary file S3). Likewise, GO terms identified with respect to post-partum cyclicity trait were multicellular organismal process (GO:0032501), multicellular organism development related to zygote settlement etc. (GO:0007275), developmental processes (GO:0032502), response to biological metabolic changes (GO:0042221), regulation of phosphorylation (GO:0042325) and regulation of phosphorus metabolic processes (GO:0051174). These GO terms were involved in active energy reorganization, mRNA processing, transcription activator / regulation, intracellular signaling cascade, regulation of macromolecules biosynthesis and cellular component assembly. It was notable that the genes classified under cellular organizational functions regulations (GO:0016043) was identified in all four traits and were involved in channeling different biological processes viz. chemical synaptic transmission, ion trans-membrane transport and regulation of membrane potential signal transduction and milk production [40,49]. In addition, significant GO terms (FDR < 0.05) found associated exclusively with age at first calving trait, were relating to humoral immune response (GO:005089), establishment and maintenance of cellular component location (GO:0051179). The GO term (GO:0065007) encoding condensin complex subunit 2 (Q3MHQ) gene was associated with feed efficiency trait. Interestingly, this gene is related to metabolic processes involved in feed conversion efficiency as it converts interphase chromatin into mitotic chromosome condensation, thus regulating cell division and thereby enhancing growth in cattle [50,51].

In conclusion, this is the first study to document the identification and functional classifications of genome-wide SNPs along with molecular functions and processes with respect to four important traits viz. milk volume, age at first calving, post-partum cyclicity resumption and feed conversion efficiency in buffalo. The information deduced will provide a novel avenue for identification of superior germplasm in buffalo breeding program.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2020.04.012.

## Ethics statement

Animals ($n$ = 85) were selected for selective genotyping of buffaloes, based on performance phenotype recorded at ICAR-Central Institute for research on Buffaloes (ICAR-CIRB) Hisar, Haryana, India. Genotype data was generated for selected genotypes through outsourced services hired by the institute. Animals were maintained under Institute's farm management and experiment design was approved by the Institute animal ethics committee (IAEC).

## Funding

The financial support for this study has been provided by Network Project on Agricultural Bioinformatics and Computational Biology under Centre for Agricultural Bioinformatics Scheme, ICAR-IASRI, Indian Council for Agricultural Research (ICAR), New Delhi, India.

## Declaration of Competing Interest

The authors declare that they have no competing interests.

## Acknowledgement

We acknowledge the support from Network Project on Agricultural Bioinformatics and Computational Biology under Centre for Agricultural Bioinformatics Scheme, ICAR-IASRI, Indian Council for Agricultural Research (ICAR), New Delhi, India.

## References

[1] H. Warriach, D. McGill, R. Bush, P. Wynn, K. Chohan, A review of recent developments in buffalo reproduction—a review, Asian Australas. J. Anim. Sci. 28 (2015) 451.

[2] S. Niranjan, S. Goyal, P. Dubey, N. Kumari, S. Mishra, M. Mukesh, et al., Genetic diversity analysis of buffalo fatty acid synthase (FASN) gene and its differential expression among bovines, Gene. 575 (2016) 506–512.

[3] F. Faostat, Agriculture Organization of the United Nations Statistics Division. Economic and Social Development Department, Rome, Italy, http://faostat3. fao. org/home/E. Accessed (2016), p. 12.

[4] C. Khedkar, S. Kalyankar, S.S. Deosarkar, Buffalo milk, Encyclopedia of Food and Health, 2016, pp. 522–528.

[5] G. De Camargo, R.R. Aspilcueta-Borquis, M. Fortes, R. Porto-Neto, D.F. Cardoso, D. Santos, et al., Prospecting major genes in dairy buffaloes, BMC Genomics 16 (2015) 872.

[6] W.Y. Low, R. Tearle, D.M. Bickhart, B.D. Rosen, S.B. Kingan, T. Swale, et al., Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity, Nat. Commun. 10 (2019) 260.

[7] H.L. Stickney, J. Schmutz, I.G. Woods, C.C. Holtzer, M.C. Dickson, P.D. Kelly, et al., Rapid mapping of zebrafish mutations with SNPs and oligonucleotide microarrays, Genome Res. 12 (2002) 1929–1934.

[8] L.K. Matukumalli, C.T. Lawley, R.D. Schnabel, J.F. Taylor, M.F. Allan, M.P. Heaton, et al., Development and characterization of a high density SNP genotyping assay for cattle, PLoS One 4 (2009) e5350.

[9] T. Surya, M. Vineeth, J. Sivalingam, M. Tantia, S. Dixit, S. Niranjan, et al., Genomewide identification and annotation of SNPs in *Bubalus bubalis*, Genomics 111 (2019) 1695–1698.

[10] A. Jerome, J. Bhati, D.C. Mishra, K.K. Chaturvedi, A.R. Rao, A. Rai, et al., MicroRNA-related markers associated with corpus luteum tropism in buffalo (Bubalus bubalis), Genomics. 112 (2020) 108–113.

[11] P. Yodklaew, S. Koonawootrittriron, M.A. Elzo, T. Suwanasopee, T. Laodim, Genome-wide association study for lactation characteristics, milk yield and age at first calving in a Thai multibreed dairy cattle population, Agriculture and Nat. Res. 51 (2017) 223–230.

[12] R.A. Arafa, M.T. Rakha, N.E.K. Soliman, O.M. Moussa, S.M. Kamel, K. Shirasawa, Rapid identification of candidate genes for resistance to tomato late blight disease using next-generation sequencing technologies, PLoS One 12 (2017) e0189951.

[13] N.A. Baird, P.D. Etter, T.S. Atwood, M.C. Currey, A.L. Shiver, Z.A. Lewis, et al., Rapid SNP discovery and genetic mapping using sequenced RAD markers, PLoS One 3 (2008) e3376.

[14] J.W. Davey, T. Cezard, P. Fuentes-Utrilla, C. Eland, K. Gharbi, M.L. Blaxter, Special features of RAD sequencing data: implications for genotyping, Mol. Ecol. 22 (2013) 3151–3164.

[15] K. Shirasawa, H. Hirakawa, DNA marker applications to molecular genetics and genomics in tomato, Breed. Sci. 63 (2013) 21–30.

[16] M. Víquez-Zamora, B. Vosman, H. van de Geest, A. Bovy, R.G. Visser, R. Finkers, et al., Tomato breeding in the genomics era: insights from a SNP array, BMC Genomics 14 (2013) 354.

[17] K. Shirasawa, H. Hirakawa, S. Isobe, Analytical workflow of double-digest restriction site-associated DNA sequencing based on empirical and in silico optimization in tomato, DNA Res. 23 (2016) 145–153.

[18] J. Sambrook, D.W. Russell, Molecular cloning: a laboratory manual (3-volume set), Immunol. 49 (2001) 895–909.

[19] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics. 30 (2014) 2114–2120.

[20] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with bowtie 2, Nat. Methods 9 (2012) 357.

[21] H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, Bioinformatics. 27 (2011) 2987–2993.

[22] J. Catchen, P.A. Hohenlohe, S. Bassham, A. Amores, W.A. Cresko, Stacks: an analysis tool set for population genomics, Mol. Ecol. 22 (2013) 3124–3140.

[23] P.J. Bradbury, Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, E.S. Buckler, TASSEL: software for association mapping of complex traits in diverse samples, Bioinformatics. 23 (2007) 2633–2635.

[24] P. Cingolani, A. Platts, L.L. Wang, M. Coon, T. Nguyen, L. Wang, 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, 6 (2012), pp. 80–92.

[25] Oliveros J.C. VENNY. An interactive tool for comparing lists with venn's diagrams. https://bioinfogp.cnb.csic.es/tools/venny/index.html. 2007.

[26] D.W. Huang, B.T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, et al., DAVID bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists, Nucleic Acids Res. 35 (2007) W169–W175.

[27] T. Tian, Y. Liu, H. Yan, Q. You, X. Yi, Z. Du, et al., agriGO v2. 0: a GO analysis toolkit for the agricultural community, 2017 update, Nucleic Acids Res. 45 (2017) W122–W129.

[28] A. Conesa, S. Götz, Blast2GO: a comprehensive suite for functional analysis in plant genomics, Int. J. Plant Genom. 2008 (2008) 619832.

[29] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, et al., Circos: an information aesthetic for comparative genomics, Genome Res. 19 (2009) 1639–1645.

[30] M. Fumagalli, Assessing the effect of sequencing depth and sample size in population genetics inferences, PLoS One 8 (2013) e79667.

[31] R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets, Bioinformatics. 27 (2011) 863–864.

[32] K. Norrgard, J. Schultz, Using SNP data to examine human phenotypic differences, Nat Educ. 1 (2008) 85.

[33] J. Karijolich, Y.-T. Yu, Spliceosomal snRNA modifications and their function, RNA Biol. 7 (2010) 192–204.

[34] R. Singh, V. Junghare, S. Hazra, U. Singh, G.S. Sengar, T. Raja, et al., Database on spermatozoa transcriptogram of catagorised Frieswal crossbred (Holstein Friesian X Sahiwal) bulls, Theriogenology. 129 (2019) 130–145.

[35] M. Vercouteren, J. Bittar, P. Pinedo, C. Risco, J. Santos, A. Vieira-Neto, et al., Factors associated with early cyclicity in postpartum dairy cows, J. Dairy Sci. 98 (2015) 229–239.

[36] H. Cai, Y. Zhou, W. Jia, B. Zhang, X. Lan, C. Lei, et al., Effects of SNPs and alternative splicing within HGF gene on its expression patterns in Qinchuan cattle, J. Anim. Sci. and Biotechnol. 6 (2015) 55.

[37] X. Zheng, Z. Ju, J. Wang, Q. Li, J. Huang, A. Zhang, et al., Single nucleotide polymorphisms, haplotypes and combined genotypes of LAP3 gene in bovine and their association with milk production traits, Mol. Biol. Rep. 38 (2011) 4053–4061.

[38] C. Du, T. Deng, Y. Zhou, T. Ye, Z. Zhou, S. Zhang, et al., Systematic analyses for candidate genes of milk production traits in water buffalo (Bubalus Bubalis), Anim. Genet. 50 (2019) 207–216.

[39] S. Capomaccio, M. Milanesi, L. Bomba, K. Cappelli, E.L. Nicolazzi, J.L. Williams, et al., Searching new signals for production traits through gene-based association analysis in three Italian cattle breeds, Anim. Genet. 46 (2015) 361–370.

[40] H. Fan, Y. Wu, X. Qi, J. Zhang, J. Li, X. Gao, et al., Genome-wide detection of selective signatures in Simmental cattle, J. Appl. Genet. 55 (2014) 343–351.

[41] M.K. Abo-Ismail, L.F. Brito, S.P. Miller, M. Sargolzaei, D.A. Grossi, S.S. Moore, et al., Genome-wide association studies and genomic prediction of breeding values for calving performance and body conformation traits in Holstein cattle, Genet. Sel. Evol. 49 (2017) 82.

[42] N. Sasago, T. Abe, H. Sakuma, T. Kojima, Y. Uemoto, Genome-wide association study for carcass traits, fatty acid composition, chemical composition, sugar, and the effects of related candidate genes in Japanese black cattle, Anim. Sci. J. 88 (2017) 33–44.

[43] C. Sun, C. Southard, D.B. Witonsky, R. Kittler, A. Di Rienzo, Allele-specific down-regulation of RPTOR expression induced by retinoids contributes to climate adaptations, PLoS Genetics (2010) 6.

[44] K. Setoguchi, T. Watanabe, R. Weikard, E. Albrecht, C. Kühn, A. Kinoshita, et al., The SNP c. 1326T > G in the non-SMC condensin I complex, subunit G (NCAPG) gene encoding a p. Ile442Met variant is associated with an increase in body frame size at puberty in cattle, Animal Genetics 42 (2011) 650–655.

[45] M. Abo-Ismail, M. Kelly, E. Squires, K. Swanson, S. Bauck, S. Miller, Identification of single nucleotide polymorphisms in genes involved in digestive and metabolic processes associated with feed efficiency and performance traits in beef cattle, J. Anim. Sci. 91 (2013) 2512–2529.

[46] V. Prakash, T.K. Bhattacharya, B. Jyotsana, O. Pandey, Molecular cloning, characterization, polymorphism, and association study of the interleukin-2 gene in Indian Crossbred cattle, Biochem. Genet. 49 (2011) 638–644.

[47] G. Sahana, J.K. Höglund, B. Guldbrandtsen, M.S. Lund, Loci associated with adult stature also affect calf birth survival in cattle, BMC Genet. 16 (2015) 47.

[48] D. Duarte, C.J. Newbold, E. Detmann, F. Silva, P. Freitas, R. Veroneze, et al., Genome-wide association studies pathway-based meta-analysis for residual feed intake in beef cattle, Anim. Genet. 50 (2019) 150–153.

[49] S. Nayeri, P. Stothard, Tissues, metabolic pathways and genes of key importance in lactating dairy cattle, Springer Science Reviews. 4 (2016) 49–77.

[50] P. Widmann, A. Reverter, R. Weikard, K. Suhre, H.M. Hammon, E. Albrecht, et al., Systems biology analysis merging phenotype, metabolomic and genomic data identifies Non-SMC Condensin I Complex, Subunit G (NCAPG) and cellular maintenance processes as major contributors to genetic variability in bovine feed efficiency, PLoS One 10 (2015) e0124574.

[51] H.D. Daetwyler, A. Capitan, H. Pausch, P. Stothard, R. Van Binsbergen, R.F. Brøndum, et al., Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle, Nat. Genet. 46 (2014) 858.