https://doi.org/10.56093/ijans.v90i11.111494

Indian Journal of Animal Sciences **90** (11): 1479–1484, November 2020/Article

An information system on genomic elements and predicted protein structures of buffalo (*Bubalus bubalis*)

AMIT KAIRI[1], TANMAYA KUMAR SAHU[1] and ATMAKURI RAMAKRISHNA RAO[2]✉

ICAR–Indian Agricultural Statistics Research Institute. New Delhi 110 012 India

Received: 3 May 2019; Accepted: 29 March 2020

ABSTRACT

Among the livestock species, buffalo remained as an integral part of the Indian rural economy. With the advent of genome sequencing technologies, it became possible to sequence the whole genome of Murrah buffalo. Also, significant amount of information on different genomic elements of buffalo is available at National Centre for Biotechnology Information (NCBI). However, the positions of these elements on the genome are not fully known. In addition, the 3D structures of buffalo proteins are not available and also there exist no browser to visualize important genic elements on buffalo genome. Hence, a study was taken up to develop a web-based information system having information on genomic elements, protein 3-D structures and genome browser. Initially, information on nucleotide and protein sequences were retrieved from NCBI and parsed suitably. Later, the protein structures were predicted, validated, refined and stabilized *in silico*. An Information System on Buffalo Genome (ISBG) with 3-tier architecture was developed containing the sequence and structural information. ISBG contains complete coding sequences (CDS), Mitochondrial DNAs, 1k upstream regions and Untranslated Regions (UTRs) of buffalo genome. The buffalo genes were also mapped onto the genome. The results revealed that maximum number of genes were found distributed on chromosome 4 followed by chromosome 18, which can also be visualized from the developed genome browser. ISBG can be accessed at http://cabgrid.res.in:8080/bgis. The proposed information system helps animal breeders and biotechnologist in animal improvement.

Keywords: Buffalo genome, Genome browser, Genomic elements, Information system, Protein Structure, Web-interface

Buffalo is an important animal species contributing significantly to the world's rural economy for many years as a source of milk, hide and draft (Nanda *et al.* 2003). Buffalo milk is commercially more viable than cow milk from the viewpoint of manufacture of fat-based and Solids-not-fat (SNF) based milk products. Tantia *et al.* (2011) provided a whole-genome sequence assembly of single female Murrah buffalo and a large amount of nucleotide sequence, protein sequence and structure information of buffalo species is available at NCBI and other public domain resources. Such information plays a crucial role to understand the functionality of genes. Another important aspect of genomic studies is the annotation and mapping of sequence information of genomic elements, viz. CDS, UTR, Exon, Intron, and Promoters on genome through a browser. The most popular genome browsers, in general, are Generic Genome Brower (Stein *et al.* 2002), SynBrowse (Pan *et al.* 2005), Light Weight Genome Viewer (Faith *et al.* 2007), JBrowser (Mitchell *et al.* 2009) and CisGenome Browser (Jiang *et al.* 2010). In the past, little efforts were made to consolidate the available genome sequence information of buffalo species. For example, Tantia *et al*. (2011) integrated the buffalo genome assembly and made it publicly available through a genome browser. Such information could be used for exploring uniqueness of buffalo species and permit improvement in production traits. Moreover, prediction of 3D structure of all available buffalo proteins could be of immense use while improving production traits. Keeping the above in view, the present study is taken up with the objectives (i) to design and populate a database on genomic elements, protein sequences and their structural information of buffalo, and (ii) to develop visual interface for annotation of genomic elements on genome through development of a genome browser.

Present address: [1]Centre For Agricultural Bioinformatics (CABin) ICAR-IASRI, New Delhi 110 012; [2]Centre of Agricultural Bioinformatics (CABin), PUSA, Library Avenue, New Delhi 110 012. ✉Corresponding author e-mail: ar.rao@icar.gov.in

MATERIALS AND METHODS

The nucleotide sequence information of genomic elements of buffalo species was collected from the NCBI by using a developed in-house BioPERL script and the assembled whole genome sequence information of Murrah buffalo was collected from http://webapp.cabgrid.res.in/buffsatdb/chr/.

Parsing of sequence information: The nucleotide sequence information of *Bubalus bubalis* available at NCBI was parsed through a developed BioPERL script. There were 6,20,623 sequences present in NCBI related to *Bubalus bubalis*. Another BioPERL script was written to categorize the nucleotide sequences information into different categories, viz. Compete CDS, Exonic Region, Intronic Region, Promoter Region, Partial CDS, Mitochondrial Sequence, Satellite sequence, Minisatellite tagged sequence and other short read sequences. The categorized nucleotide sequences were then stored in CSV format along with their Accession No. and description. The translated protein sequences were also retrieved form NCBI. The protein sequences, their IDs and annotations were parsed and stored in CSV format.

Protein structure prediction: As the 3D structures of the retrieved/stored protein sequences were not available, the same were predicted by submitting them to Phyre^2 (Protein Homology/Analogy Recognition Engine) server. Phyre^2 (Kelley *et al.* 2015) uses the Fold recognition technique to predict the protein structure. For that, it uses a library of known protein structures taken from the Structural Classification of Protein (SCOP) database and Protein Data Bank (PDB). Five iterations of PSI-Blast were done by Phyre^2 to construct the profile and to accumulate the remote and close homologs. After the construction of profile, three independent structure prediction programs, viz. Psi-Pred (McGuffin *et al.* 2000); SSPro (Pollastri *et al.* 2002) and JNet (Cole *et al.* 2008) were used to predict the secondary structure (α helix, β strand and coil) of the submitted proteins. The above mentioned programs provide a confidence value at each position of the submitted sequence for α helix, β strand and coil. The average confidence value of Psi-Pred, SSPro and JNet were taken and a final, consensus prediction for each targeted protein was calculated. A program Dispored (Ward *et al.* 2004) was also run by Phyre^2. These generated profiles and secondary structures were then scanned against the fold library using a profile-profile alignment exclusively developed for Phyre^2 tool (Bernnett-Lovsey *et al.* 2008). Ranking of alignments was done based on the score returned by the Bernnett-Lovsey alignment algorithm. These scores were fitted to an extreme value distribution to generate E-value by Phyre^2. The top ten highest scoring alignments were then used to construct full 3D models of the query.

For validation of the predicted structures, Ramachandran Plots were obtained by online utility Rampage server (Bakker *et al.* 2002;http://mordred.bioc.cam.ac.uk/~rapper/rampage.php) to find the residues present in the allowed region and outlier region. The residues in the outlier region were refined by an iterative process available in ModLoop (Fiser *et al.* 2000; https:// modbase.compbio.ucsf.edu/modloop/). ModLoop generated pdb files were again analysed using Ramachandran Plot. This procedure was repeated till the structures appear to be containing more than 98% residues in favoured region and 0% in the outlier region. For stabilization of the refined protein structures, energy minimisation was done by YASARA Energy Minimization Server (http://www.yasara.org/minimizationserver.htm). The final PDB co-ordinates were collected and kept in suitable format for necessary import into database for further visualization of protein 3D structures.

Annotation and distribution of genomic elements on genome: For the purpose of annotating the available genomic regions onto the complete genome sequence, the off-line BLAST tool was downloaded from NCBI (http://www.ncbi.nlm.nih.gov/) whereas, the retrieved CDS sequences were used as query sequences in the off-line BLAST for annotation. Besides, the whole genome sequence retrieved from http://webapp.cabgrid.res.in/buffsatdb/chr/ was used as a background database in the local BLAST (Off-line BLAST tool). Three files, viz. NHR, NIN and NSQ were generated by BLAST while developing the local database on complete genome of *Bubalus bubalis.* The results obtained from BLAST homology search were stored and necessary PERL script was written to parse the relevant information, i.e. genomic co-ordinates of CDS on the genome.

Finally, the starting and end co-ordinates on genome for different CDS sequences along with the accession no, length of gene, chromosome number, E-value, % identity and strand information were collected and kept in a suitable format (tab-delimited ASCII) for importing the information into the database through standard commands of MySQL. A similar approach was followed for annotating promoter, UTR and mitochondrial DNA on genome. The distribution of CDS density was also analyzed from the annotations.

Information system on Buffalo Genome (ISBG)

A buffalo genome information system is proposed here to provide an efficient, informative and interactive interface for the user to explore and carry further research in buffalo species. The procedure is described here.

Database development: The information parsed in the CSV format from the available buffalo genome sequences were used to develop a database on buffalo genome. The Data Flow Diagram (DFD) of the developed database is shown in Fig.1 to project the data flow, data processes and data sources/destinations. Also, an Entity Relationship (ER) diagram is given in Fig.2 to show the relationships between Entities. Based on ER-diagram, the Buffalo genome databases with different tables were created. In addition, the attributes with different data types were defined in the phpMyAdmin. Using the principles of database design, the parsed data was imported into the database through "MySQL import" commands.

Software development: An Information System on Buffalo Genome (ISBG) with 3-tier architecture, was developed with the database at bottom layer (DBL), PHP as server side application-middle layer, HTML, CSS and Javascript at top layer as client side application layer. The ISBG using LAMP technology (Linux, Apache, MySQL, PERL/PHP) was developed.

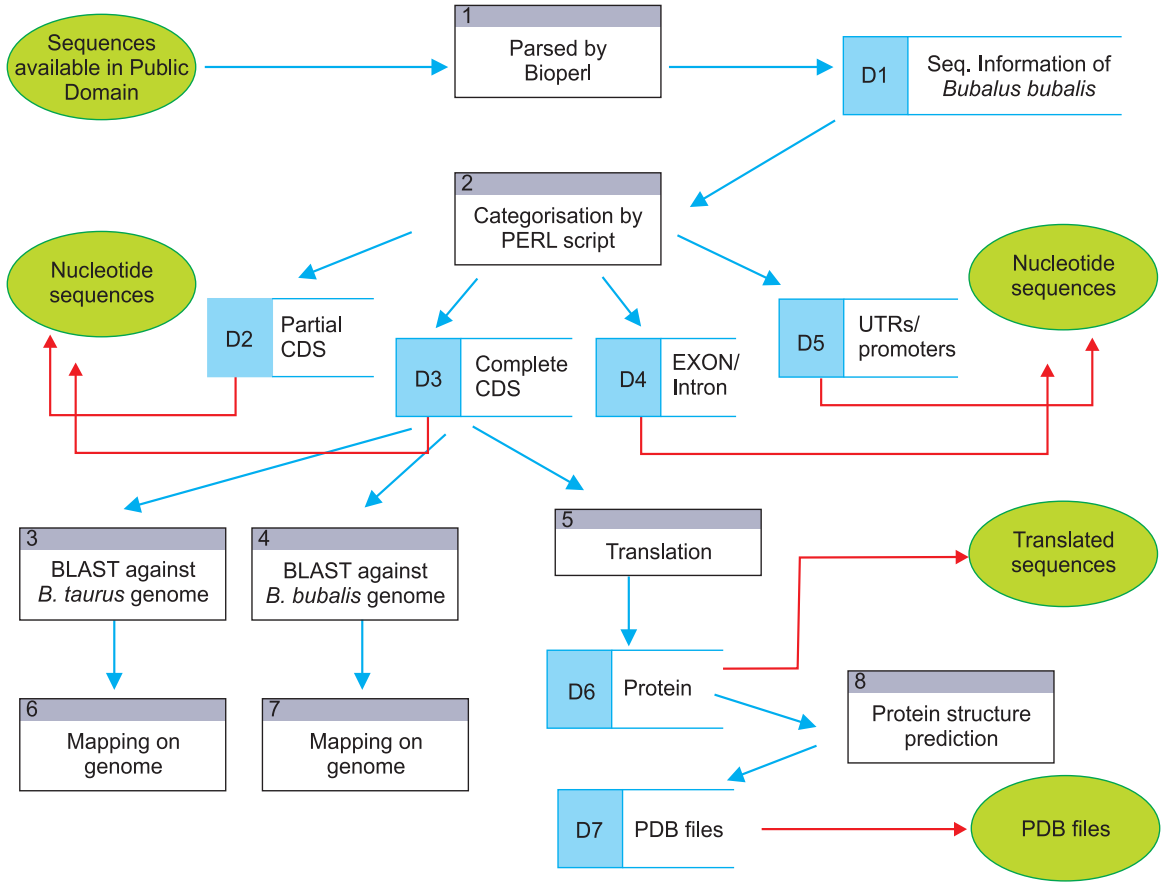

Fig. 1. Dataflow diagram (level-0) of Information System on Buffalo genome.

Buffalo genome browser: Genome Browsers are used to visualize genomic elements on genome through visual graphic display. Here, a genome browser was developed to represent genomic elements on genome. The back-end database was used as input for the development of genome browser. The proposed Buffalo genome browser is based on LWGV (Light Weight Genome Viewer) tool and represents buffalo genome features as colour-coded tracks on rectangular bars of chromosomes. The detailed information about each feature was shown by "mousing-over" the track. These features ware described in a text-file, written in a specific format called annotation (.ann) file. Commonly used feature sets and configuration parameters were stored in separate files and included in the annotation file with "#include" statement to prevent regenerating the same features in contexts where only part of the data analysis is dynamic. The genome browser also contains onMouseOver facility with the details of the annotated sequences like, E-value, GenBank link, Accession no and Identity (%). In addition to that, Zoom-in and Zoom-out facilities were given to visualize further the compressed regions represented in blue colour.

Web-interface: A Web interface with ISBG and Buffalo genome browser was developed by keeping the interface simple, consistent with proper page layout, colour and texture and made available to users in a more user-friendly manner. Necessary PHP scripts were written for the purpose. The Web-interface was provided at http:/cabgrid.res.in:8080/bgis/. Facilities like, search, filter, view, print, download, browse, links to public domain databases together with the information on gene, chromosome, accession number, etc. were given in the web-interface.

RESULTS AND DISCUSSION

Based on the methods described under methodology, the retrieved sequences from NCBI were classified into complete CDS, Exon, Intron, Promoter, Partial CDS, Mitochondrial DNA, Promoter sequences and UTRs. The number of sequences falling under different categories is given in Table 1. For the purpose of classification, BioPERL script is written and provided in the supplementary materials. The PERL scripts written to parse the sequences

Table 1. Categorization of sequence information of buffalo species retrieved from NCBI

Functional element	Number of sequences
Complete CDS	930
Partial CDS	1154
Exon	656
Intron	237
Mitochondrial DNA	1709
Promoter	73
UTR	67
Whole Genome Chromosomes	24

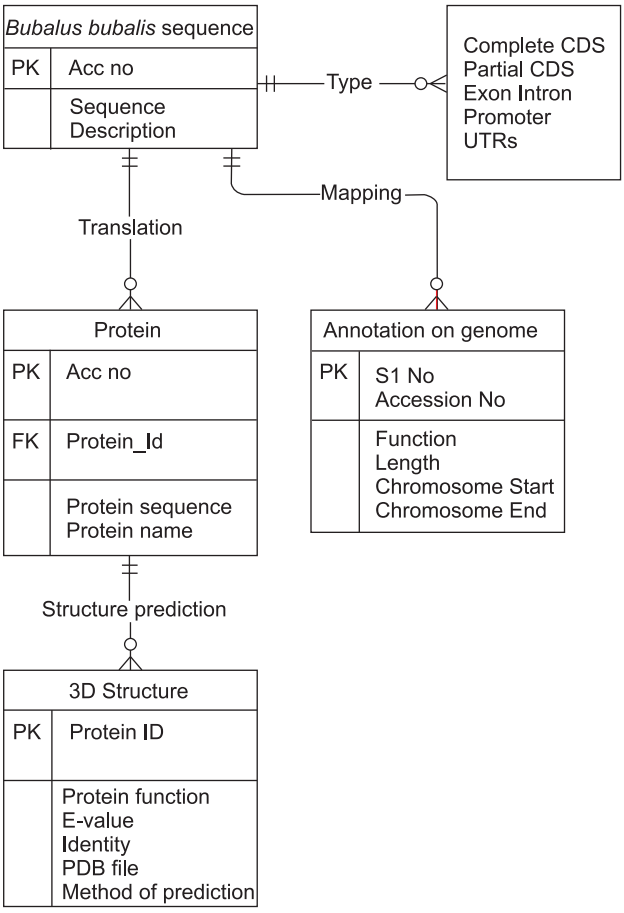

Fig. 2. ER diagram of Buffalo genome database.

of *Bubalus bubalis* are also given in supplementary materials. As described in the methodology, the accession number, description of the sequences and the sequences of genomic regions arranged in CSV format were imported into MySQL database.

A total of 930 protein primary sequences were subjected to structure prediction by using Phyre[2]. The protein structure (.pdb files) generated from Phyre[2] were further refined by Modloop and then every refined structure was rigorously validated by Ramachandran's Plot through an iterative process till all the residues fell in the allowed region. The final refined structures were again subjected to energy minimisation by using YASARA energy minimization server as outlined in the methodology. For example, three energy minimised protein 3D structures are given in Fig.3. The assembled whole genome sequence of female Murrah breed, obtained from http://webapp.cabgrid.res.in/buffsatdb/chr/ was taken as "subject-sequence" in NCBI offline BLAST tool and the complete CDS sequences were taken as "query" sequences. The BLASTn results were parsed and mapped onto genome by collecting the start and end co-ordinates of CDS sequences on chromosomes using BioPERL script. The mapped information, i.e. the length of CDS, query-start, query-end, subject-start (start co-ordinate on chromosome), subject-end (end co-ordinate on chromosome), chromosome number, E-value, Identity and Strand information for a sample of CDS sequences is given in Fig. 4.

The parsed information of different functional elements and mapped information of CDS onto genome were stored in the database. Also, a snapshot of the PHP script written for the development of software is given in supplementary Fig. S2.

The mapping information present in the database was used as an input for developing the annotation file (.ann) required for developing genome browser. The .ann file also has the "onMouseOver" facility with details on E-value, GenBank link, Accession No., and % identity. The homepage of ISBG is given in Fig.5.

As mentioned earlier in materials and methods, the Web-interface was developed by writing PHP scripts. This interface contains the database, information system, genome browser and user-choice. The Web-interface is made available in public domain at http://cabgrid.res.in:8080/bgis/ . The snapshots of different pages are given in Fig 6a and 6b.

The chromosomal distribution of buffalo genome along with the visual display of buffalo genes on genome can be seen from Fig. 6(b). The regions on the chromosomes are compressed and shown with blue colour. The decompression of blue regions facilitates the visual display of genes present in the compressed region. The distribution of genes on chromosome is given in Table 2.

No gene is found to be present on chromosome number 8, 16, 19 and 23. This may be due to non-availability of gene information in public domain or due to low-coverage of whole genome assembly available at http://webapp.cabgrid.res.in/buffsatdb/chr/. Out of 930 genes of buffalo, available at NCBI, a total 837 genes are mapped onto *Bubalus bubalis* genome.

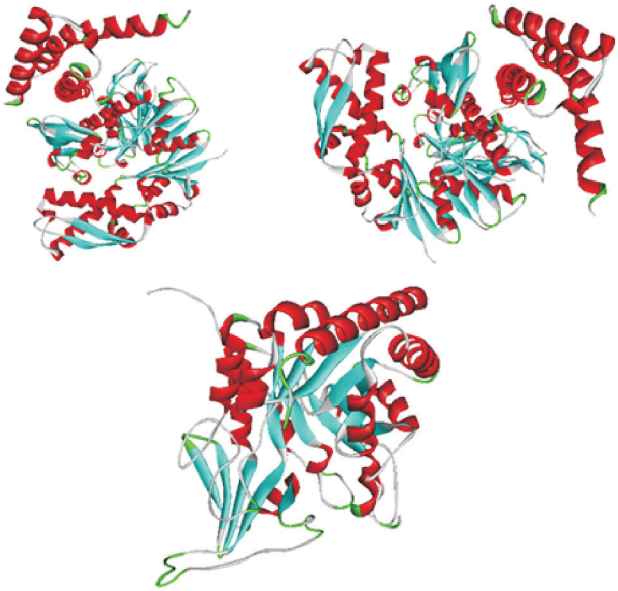

Fig. 3. Three energy minimized structures of Buffalo genes (GU183099, GU183099, JQ859818).

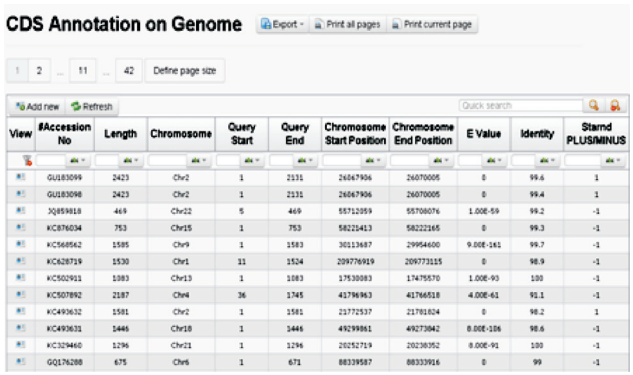

Fig. 4. Chromosomal position of few CDS sequences and their other sequence information on genome.

Fig. 5. Homepage of ISBG.

Fig. 6 (a-b). **(a)** CDS sequence annotation on genome, **(b)** Buffalo genome browser.

Table 2. Distribution of *Bubalus bubalis* genes on its chromosome

Chromosome number	Number of genes
Chromosome 1	20
Chromosome 2	43
Chromosome 3	66
Chromosome 4	324
Chromosome 5	33
Chromosome 6	33
Chromosome 7	22
Chromosome 9	5
Chromosome 10	9
Chromosome 11	8
Chromosome 12	2
Chromosome 13	10
Chromosome 14	4
Chromosome 15	15
Chromosome 17	12
Chromosome 18	155
Chromosome 20	7
Chromosome 21	4
Chromosome 22	17
Chromosome 24	1

Recently Williams *et al.* (2017) assembled 2 species of domestic water buffalo, the river (2n=50) and the swamp (2n=48) buffalo. They described a draft quality reference sequence for the river buffalo created from Illumina GA and Roche 454 short read sequences using the MaSuRCA assembler. Annotation of the genome was supported by transcriptome data from 30 tissues and identified 21,711 predicted protein coding genes, however, in their study they did not categorize all the other non coding sequences and there is no browser to visualize all the coding sequences over the genome. In addition to that, predicted protein structures for coding sequences were not available. However, Low *et al*. (2019) and Young *et al*. (2019) have given chromosome level assembly and gene expression atlas of water buffalo respectively.

Arora *et al.* (2013) developed a Buffalo microsatellite database BuffsatDb: http://webapp.cabgrid.res.in/buffsatdb/ that is mainly customized of retrieval of chromosome-wise microsatellite markers alone. An extensive comparison between river buffalo and domestic cattle was done by Amaral *et al.* (2008) through RH map. Here radiation hybrid chromosome maps for the entire river buffalo genome based on cattle-derived markers were obtained. Possible rearrangements between the chromosomes of river buffalo and cattle were detected from the alignments of RH maps.

Another comparative sequence alignment was done by Li *et al.* (2018). This study characterizes differences in gene content, regulation and structure between taurine cattle and river buffalo. 127 deletion CNV regions in river buffalo were identified representing 5 annotated cattle genes. Transcriptome analysis in various tissue types on river buffalo confirmed the absence of four cattle genes. But protein structure prediction of the relevant genes has not been done. The visualization of genomic regions has also

been ignored. Thus, the present ISBG is going to be an asset for the animal breeders and biotechnologists.

The buffalo species (*Bubalus bubalis*) is an important livestock species and contribution of this species towards milk, meat, draught is enormous in Indian rural economy. With the recent developments in genome sequencing technologies Indian scientist made a breakthrough by generating whole genome sequence assembly of Murrah Breed of the said species. Also, a large amount of sequence information of *Bubalus bubalis* is available in public domain. However, to the best of our knowledge, it seems difficult to find the processed and parsed information at one place. Moreover, the predicted protein 3D structures of buffalo are likely to be not available in most used protein databases. Besides, the sequence information has not been fully mapped yet and annotated on buffalo genome. However, the proposed Information System on Buffalo Genome is having a user interactive search facility of functional elements, protein structures, mapping information of complete CDS, and genome browser. It is worth noting that, out of 930 genes available in public domain a total 836 were found to be mapped onto *Bubalus bubalis* genome. The number of genes mapped onto chromosome number 4, is highest followed by chromosome number 18. It is also observed that no gene got mapped on chromosomes 8, 16, 19 and 23. This may be due to non-availability of gene information or coverage of genome assembly with lesser depth.

ACKNOWLEDGEMENTS

Authors acknowledge the facilities provided at Centre for Agricultural Bioinformatics (CABin), ICAR-IASRI and financial support from ICAR, New Delhi.

REFERENCES

Amaral M E, Grant J R, Riggs P K, Stafuzza N B, Rodrigues F E A, Goldammer T, Weikard R, Brunner R M, Kochan K J, Greco A J and Jeong J. 2008. A first generation whole genome RH map of the river buffalo with comparison to domestic cattle. *BMC genomics* **9**(1): 631.

Arora V, Iquebal M A, Rai A and Kumar D. 2013. In silico mining of putative microsatellite markers from whole genome sequence of water buffalo (*Bubalus bubalis*) and development of first BuffSatDB. *BMC genomics* **14**(1): 43.

Bakker P I W, DePristo M A, Burke D F and Blundell T L. 2002. *Ab initio* construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model. *Proteins: Structures, Functions, and Genetics* **51**: 21–40.

Bernnett-Lovsey R M, Herbert A D, Sternberg M J E and Kelley L A. 2008. Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. *Proteins* **70**(3): 611–25.

Cole C, Barber J D and Barton G J. 2008. The Jpred 3 secondary structure prediction server. *Nucleic Acids Research* **36**(Web server issue): W197–W201.

Faith J J, Olson A J, Gardner T S and Sachidanandam R. 2007. Lightweight genome viewer: portable software for browsing genomics data in its chromosomal context. *BMC Bioinformatics* **8**: 344.

Fiser A, Do A K and Sali A. 2000. Modelling of loops in protein structures. *Protein Science* **9**(9): 1753–73.

Jiang H, Wang F, Dyer N P and Wong W H. 2010. CisGenome Browser: A flexible tool for genomic data visualization. *Bioinformatics* **26**(14): 1781–82.

Kelley L A, Mezulis S, Yates C M, Wass M N and Sternberg M J. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols* **10**(6): 845.

Li W, Bickhart D M, Ramunno L, Iamartino D, Williams J L and Liu G E. 2018. Comparative sequence alignment reveals River Buffalo genomic structural differences compared with cattle. *Genomics* **111**(3): 418–25

Low W Y, Tearle R, Bickhart D M, Rosen B D, Kingan S B, Swale T, Thibaud-Nissen F, Murphy T D, Young R, Lefevre L and Hume D A. 2019. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nature Communications* **10**: 260.

McGuffin L J, Bryson K and Jones D T. 2000. The PSIPRED protein structure prediction server. *Bioinformatics* **16**: 404–05.

Mitchell E S, Uzilov A V, Lincoln D S, Christopher J M and Holmes I H. 2009. Jbrowse: A next-generation genome browser. *Genome Research* **19**(9): 1630–38.

Nanda A S and Nakao T. 2003. Role of buffalo in the socioeconomic development of rural Asia: Current status and future prospectus. *Animal Science Journal* **74**(6): 443–55.

Pan X, Stein D L and Brendel V. 2005. SynBrowse: A synteny browser for comparative sequence analysis. *Bioinformatics* **21**: 3461–68.

Pollastri G, Przybylski D, Rost B and Baldi P. 2002. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* **47**: 228–35.

Stein L D, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich J E, Harris T W, Arva A and Lewis S. 2002. The generic genome browser: A building block for a model organism system database. *Genome Research* **12**(10): 1599–1610.

Tantia M S, Vijh R K, Bhasin V and Sikka P. 2011. Whole-genome sequence assembly of water buffalo (*Bubalus bubalis*). *Indian Journal of Animal Sciences* **81**: 38–46.

Ward J J, McGuffin L J, Bryson K, Buxton B F and Jones D T. 2004. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**: 2138–39.

Williams J L, Iamartino D, Pruitt K D, Sonstegard T, Smith T P, Low W Y and Coletta A. 2017. Genome assembly and transcriptome resource for river buffalo, *Bubalus bubalis* (2 n= 50). *GigaScience* **6**(10): gix088.

Young R, Lefevre L, Bush S J, Joshi A, Singh S H, Jadhav S K, Lisowski Z M, Iamartino D, Summers K M, Williams J L and Archibald A L. 2019. A gene expression atlas of the domestic water buffalo (*Bubalus bubalis*). *Frontiers in Genetics* **10**: 668.