Indian Journal of Animal Sciences **81** (5): 38–00, May 2011

Whole-genome sequence assembly of the water buffalo (*Bubalus bubalis*)

M S TANTIA[1], R K VIJH[1], V BHASIN[1*], POONAM SIKKA[1#], P K VIJ[1], R S KATARIA[2], B P MISHRA[3], S P YADAV[4#], A K PANDEY[1#]. R K SETHI[5#], B K JOSHI[5], S C GUPTA[6*] and K M L PATHAK[7*]

National Bureau of Animal Genetic Resources, Karnal, Haryana 132 001India
** Animal Science Division, ICAR, New Delhi110 014*
 # Central Institute for Research on Buffaloes, Hisar 125 001 Haryana

ABSTARCT

The buffalo is an integral part of agriculture, particularly within the continent of Asia, providing a source of milk, meat, skin, hides, fertilizer, fuel, and draft power. The efficiency of this animal, compared to that of cattle, is higher in this region, though little is known about genome sequence of buffalo. The first version of assembly of a single female Murrah buffalo was constructed with Illumina paired end and mate pair short read sequencing using the cattle genome (Btau 4.0 assembly) as a reference. The assembly has read depth of 17-19X. The buffalo assembly represents ~ 91%-95% coverage in comparison to the cattle assembly Btau 4.0. The assembly has 185,150 contigs with the median contig length of 2.3 Kb and the largest contig length of 663 Kb. The mitochondrial genome is fully covered by a single contig. Whole genome comparison between this assembly and of cattle revealed 52 million mismatches/indels. The present analysis also unveils about 300 structural variants in the buffalo genome. The buffalo assembly has been integrated into a publically available genome browser with tracks for read pair insert distances, read depth, nucleotide variations, coverage, and the availability of custom tracks for scientific community. This assembly of the Water Buffalo is the first deep sequencing project that provides the resources to better understand the genomic basis of adaptable traits and genetic variation that distinguishes buffalo from cattle.

Key words: Buffalo, Genome sequence, Assembly

Buffalo is an important bovine species inhabiting the least developed and developing countries. With the presence of 188.3 million heads (FAO, 2009) contribute 5% of total milk production in the World. Asia has nearly 97% of buffaloes and is an integral part of agriculture in India, China, Pakistan, Nepal, Bangladesh, Thailand, Myanmar and Malaysia. The productivity of buffaloes in these regions is higher as compared to cattle. Buffalo being a multi-utility animal contributes towards milk, meat, skin, hides and draft power. Water buffaloes are defined as Riverine and Swamp type. The distribution of Riverine buffalo is concentrated in Indochina, the Mediterranean, and parts of South and Central America. Swamp buffalo are more easterly in distribution and inhabit Indochina and Southeast Asia as well as Australia (Roth and Myers, 2004).

Buffaloes belong to family Bovidae with a diploid count of chromosomes being 48 (Swamp) and 50 (Riverine). The synteny in buffalo and cattle chromosomes is quite large (Amaral et al., 2008). The 5 biarmed chromosomes are result of centric fusion of five pairs of cattle chromosomes. Iannuzzi and Di Meo (2009) reported 309 mapped loci on all buffalo chromosome arms mostly assigned by FISH. The recent effort of radiation hybrid mapping of Amaral et al., 2009, have 2621 cattle derived loci covering all the buffalo chromosomes. The first generation whole genome RH map for river buffalo when compared to Btau_4.0 genome sequence assembly showed the marker order with in linkage groups was consistent with cow assembly (Michelizzi1 et al., 2010). Stafuzza et al., 2009 generated RH map for Y chromosome with 28 markers in a single linkage group. These studies encouraged different workers interested in buffalo genomics to undertake buffalo genome mapping initiatives using cow genome resources. Till date the water buffalo gene and genomic resources are meagre as compared to other members of Bovidae like cow and sheep. Genome sequencing in livestock is advancing at a great speed with whole genome sequences being available for cow, sheep, horse, chicken and dog. Buffalo has a typical position and its genome shall shed light on the evolutionary biology and shall reveal the genomic basis of adaptability traits for tropical conditions. Next generation sequencing have expedited the pace of whole genome sequencing efforts of large number of model organisms and various techniques of NGS have been recently

Present address: [1]Principal Scientist (tantiams@gmail.com), [2]Senior Scientist, [3]Head, Animal Biotechnology, IVRI, Izatnagar, [4]Scientist (SS), [5]Director (rksethi7@rediffmail.com), [6]ADG (AP&B), [7]DDG (AS).

reviewed (Jiang et al., 2009) In the present study we generated paired end data and mate pair data from a single Murrah female with recorded pedigree. The data generated was analysed and compared in relation to its closely related species *Bos taurus*.

MATERIALS AND METHODS

Library preparation and sequencing

Paired End Library: Venous blood of a farm bred female Murrah buffalo was used for the isolation of genomic DNA using standard phenol-chloroform method. The DNA was further purified using Qiagen spin columns. Short-insert DNA libraries were prepared using standard Illumina protocol. DNA fragments generated from 3 microgram of genomic DNA by nebulisation at 25psi for 6 minutes were end repaired, incorporating an 'A' base to the 3' end followed by ligation of Illumina sequencing adapters. QIAquick Gel Extraction kit (Qiagen) was used to purify ~200bp insert size from the 2% agarose TAE gel. To increase sequence diversity of the library adapter ligated fragments from three different sets were pooled prior to amplification. 12 PCR cycles were carried out and the products were quantitated using an Agilent Bioanalyser.

Mate Paired Library: Twenty microgram of genomic DNA fragmented by nebulisation at 10psi for 25 seconds using compressed nitrogen gas was used to prepare 5 Kb mate-paired library with Mate paired library preparation kit (Illumina). DNA fragments were end repaired and end labelled using biotin-dNTPs followed by selection of 5 Kb fragments in a 0.8% agarose gel. Purified DNA fragments were circularized by self ligation and the linear DNA was removed using DNA Exonuclease. The circularized molecules were further fragmented by nebulisation and 'merged ends' enriched using streptavidin coated magnetic beads (Dynabeads, Invitrogen). These fragments were end repaired adding 'A' overhangs and sequencing adapters ligation followed by amplification by 18 cycles of PCR.

These libraries were sequenced using Illumina Genome Analyser IIx for Paired-End sequencing. Each library was sequenced on 2-3 lanes for estimating the quantity of library to be used for optimal data output. Image analysis and base calling was performed using Illumina Real Time Analysis software. The 200bp short insert library was sequenced to generate 76nt paired end reads. The 5 Kb mate-paired library was sequenced for only 36nt to avoid sequencing the junctions or generating 'sequence chimeras'. To assess the quality of the sequencing runs Phix control library (Illumina) was also sequenced in one lane of each flow cell. Data from runs with a mismatch rate >0.1 were discarded based on alignment of Phix control reads to the respective phage genome. Three flow cells of sequence data of 76 bp read pairs with a 200 bp insert (paired end) and one flow cell of mate pair read sequences of 36 bp with a 5 Kb insert, were produced .

Editing of reads

Raw read sequences were converted from QSEQ format to FASTQ files using custom scripts and trimmed using BRAT v1.1.12 (arguments used: m=2; q=51) (Harris et al., 2010). If a single mate from a pair was filtered out for low quality, both mates were removed. Both paired end and single mate read files were created as output from the trimming and filtering process. Approximately 25% of the low quality paired end reads were filtered out in this process.

Read mapping and scaffold assembly

To map the trimmed paired end reads to the Bovine genome (Baylor release Btau 4.0 Liu *et al*. 2009) software BWA (v0.5.8c) was used (Li and Durbin 2009). Samtools (v0.1.7) was applied to format, sort, and manipulate the mapped sequences (Li *et al*., 2009). Custom scripts were used to compile the BAM output file into contigs. To build a scaffold assembly and for construction of buffalo sequence pseudomolecules the buffalo RH map (Amaral et al., 2008) was used. The coverage, depth, and gap statistics were calculated with custom scripts.

Analysis of regions with increased depth

The R packages BSgenome, (BSgenome.Btaurus.UCSC.bosTau4), and Biostrings were used to identify regions along each cattle chromosome with at least 60X depth with buffalo reads. These regions were interrogated in 100 Kb intervals and compared to the soft masked regions of the Btau 4.0 genome version. Masked regions were identified (Smit et al., 1996) by the following four criteria: assembly gaps (AGAPS), intra-contig ambiguities (AMB), RepeatMasker (RM) and Tandem Repeats Finder (TRF).

Identification of nucleotide variations

Sequence variations between cattle and buffalo were identified with the function 'varfilter' of Samtools package using the default setting. Two additional empirical criteria: 1) the minimum read quality value 40, and 2) the minimum read depth 10X were also applied. The R packages (org.Bt.eg.db, KEGG.db, and GO.db) were used to map the cow Ensembl IDs for regions where cattle and buffalo sequences has no variations. The Ensembl IDs were matched to the corresponding Entrez IDs to identify the conserved genes based on their biological functions.

Detection of structural variants (SVs)

In order to detect SVs in buffalo as compared to cattle improper read pairs were identified having mapping quality value no less than 30 and the insert size >2 kb, either on the same chromosome or the paired mates located on two different chromosomes. Each of the candidates form a 'linking' signature (Medvedev et al., 2009), suggesting that two distant regions in the cattle genome are likely to be in close proximity in the buffalo genome. A linking signature

candidate was selected using a 300-bp window around each mate sequence supported by at least five mate pairs. Linking signature candidates located in regions of >60X read depth or large strings of repeats in the cattle genome were not considered. This SV information, along with repeated regions and read depth, was plotted using Circos (Krzywinski et al., 2009).

Data availability

The buffalo genome version Bbu_2.0-alpha is available at NCBI's Short Read Archive (SRA) under accession numbers SRX016621 and SRX015182. The AGP (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/agp/AGP_Specification.shtml) formatted assembly detail file is also available at SRA. The genome annotation and assembly browser is available at http://210.212.93.84/cgi-bin/gb2/gbrowse/bovine/ with the ability to visually explore the assembly with custom tracks via GBrowse 2.0 (Stein, et al., 2002). Each pseudomolecule was constructed from contigs assembled using the bovine genome (Baylor release Btau_4.0) as a reference according to the whole genome RH map of the river buffalo constructed by Amaral *et al*. 2008. The markers identified from the BBU RH map was sparse (approximately 2,700 markers), though these markers were used to guide the pseudomolecule assemblies. For each pseudomolecule, ambiguous nucleotides are represented with the IUPAC notation (http://www.dna.affrc.go.jp/misc/MPsrch/ InfoIUPAC.html), with large gaps represented by a block of 1000 "Ns" and small gaps filled by the reference genome sequences in lowercase bases.

RESULTS

Assembly summary

A total of over 1.5 billion reads (1.05x1011 total bp) were generated using Illumina platform from a single Murrah female buffalo. The paired end reads were 76 bp in length and were generated using a standard paired end library of 200 bp. Data was also generated for 5 kb insert library for mate pair reads of 36 bp in length with insert sizes of 5 Kb. After quality control and filtering, about 75% of reads were retained. The reference guided assembly mapped 62% of quality filtered reads (Btau 4.0 version from Baylor; Table 1).

Table 1. Buffalo read statistics

Total reads	1,561,456,346 (780,728,173 pairs)
Read lengths	3 FC 76mers (200 bp insert) 1 FC 36mer (5Kb insert)
Trimmed reads	
Total reads	1,176,593,782 (588,296,891 pairs)
Singleton reads	384,862,564
Both read mates mapped	821,464,222
Singleton mapped reads	59,428,437 (5.05%)
Total mapped reads	880,892,659 (74.89%)
Properly mapped read pairs	733,764,562 (62.36%)

The coverage buffalo genome contigs across equivalent cattle chromosomes 1-29 and X ranged from 91%-95% with N50 values ranging from 931 to 3,370 bp (Table 2). The mean depth ranged from 17X to 19X across these 30 cattle chromosomes. The maximum contig length was 448 kb. There were several regions on all the chromosomes where the depth was very high. The depth threshold of 60X was determined empirically for a minimum depth value across all chromosomes and was outside of the general distribution. These regions with a coverage depth >60X were evaluated at an interval of 100 Kb for identification of the possible reasons for such increase in depth. The plausible reasons for such increase in depth could be low complexity or highly repetitive elements. These regions might align at multiple locations throughout the genome leading substantial increase in depth of coverage. The analysis revealed that except a small region on chromosome 6 (chr6:6,300,000-6,400,000) and 7 (chr7: 53,500,000-53,600,000), the regions of depth greater than 60X were attributed to soft masked regions in the cattle genome attributed to assembly gaps or intra-contig ambiguities. The two 100 Kb regions identified on chromosomes 6 and 7 demonstrated high sequence similarity for roughly one-third of their sequence length, sharing 99% identity (29,796/30,293), and both having a top BLAST (Altschul SF et al. 1990) hit to 'Bovine genomic fragment for #1.709 satellite DNA'. Chromosome depth plots were constructed with the soft masked regions represented by blue lines (at least 60X depth and at least 30% of the bases in a 100 Kb interval are masked) and non-masked regions represented by green lines (at least 60X depth and less than 30% of the bases in a 100 Kb interval are masked) along the chromosome at y=100.

When examining the gap regions, we found that 7% of the cattle genome is not mapped by any buffalo read of good quality. The average gap lengths across the cattle chromosomes 1-29 and X ranged from 731 bp to 1,382 bp with large gaps lengths ranging from 124,586 bp to 459,542 bp (Table 3). For all cattle chromosomes (1-29, X), these regions of large gap lengths for buffalo reads mapping to the cattle genome are attributed to assembly gaps in the cattle genome assembly. Among these maximum gap lengths in Table 3 for each cattle chromosome, at least 97% of the bases in the Btau 4.0 assembly are represented by arbitrary nucleotides ("N"), indicating that the coverage of buffalo reads mapped to the cattle genome was underestimated in the present study.

Identificationof variations (buffalo vis-à-vis cattle)

The nucleotide variation between buffalo and cattle were identified from the buffalo reads that were mapped to the Btau 4.0 genome. In total, we identified 52,001,319 variations with 40,547,546 (78%) of them being high quality. This

Table 2. Contig coverage summary statistics for buffalo read mapping assembled across the Btau_4.0 chromosomes

Chromosome		Total Length	Contig length	Coverage	Count	Minimum	Maximum	Mean	Median	N90
Cattle	Buffalo									
1.	1q	161,112,571	148,689,783	92%	12,605	35	475,672	11,796	1,800	35,387
2.	2q	140,809,139	132,108,058	94%	9,644	41	365,632	13,698	1,975	40,874
3.	6	127,931,374	118,378,207	93%	8,838	45	626,611	13,394	1,655	40,464
4.	8	124,461,602	117,031,333	94%	8,591	49	385,693	13,623	1,993	41,453
5.	4q	125,851,629	116,531,865	93%	9,655	31	479,417	12,070	1,505	35,596
6.	7	122,567,560	112,098,606	91%	10,269	36	401,009	10,916	1,840	32,480
7.	9	112,086,926	104,509,580	93%	8,433	37	607,838	12,393	1,569	35,460
8.	3q	116,952,631	107,607,137	92%	7,894	32	555,273	13,632	1,957	41,069
9.	10	108,154,237	100,651,363	93%	8,587	42	662,504	11,721	1,666	34,822
10.	11	106,392,721	100,016,561	94%	7,239	45	455,078	13,816	1,665	40,875
11.	12	110,177,331	103,360,885	94%	6,415	48	426,998	16,112	2,062	46,543
12.	13	85,365,658	79,834,680	94%	7,563	28	443,615	10,556	1,444	30,720
13.	14	84,426,694	79,428,116	94%	3,843	45	451,441	20,668	2,250	66,048
14.	15	81,352,385	76,761,875	94%	4,986	44	459,952	15,395	1,808	46,897
15.	16	84,636,695	77,910,307	92%	7,991	50	534,606	9,750	1,150	25,979
16.	5q	77,911,411	71,906,287	92%	5,300	43	538,518	13,567	1,619	40,986
17.	17	76,512,898	70,559,927	92%	5,218	32	411,739	13,522	2,160	38,550
18.	18	66,145,125	60,816,937	92%	3,857	34	547,169	15,768	1,641	47,523
19.	3p	65,321,398	60,605,098	93%	3,460	43	573,912	17,516	1,786	52,660
20.	19	75,802,968	70,350,645	93%	5,444	45	393,446	12,923	1,735	39,413
21.	20	69,177,455	64,581,383	93%	4,858	40	406,188	13,294	1,732	40,903
22.	21	61,853,906	58,455,445	95%	2,672	47	498,578	21,877	3,370	66,821
23.	2p	53,383,219	49,900,510	93%	3,447	41	362,934	14,477	1,933	42,632
24.	22	65,027,238	60,974,074	94%	3,896	41	510,230	15,650	2,195	46,124
25.	24	44,066,150	41,280,505	94%	2,145	53	501,231	19,245	2,526	58,084
26.	23	51,757,727	47,863,689	92%	3,776	43	383,352	12,676	1,484	36,991
27.	1p	48,755,914	44,741,898	92%	4,277	56	299,495	10,461	931	33,321
28.	4p	46,088,657	42,810,178	93%	3,013	35	407,874	14,208	2,165	42,521
29.	5p	52,001,983	47,325,309	91%	4,401	49	274,712	10,753	1,305	31,455
X	X	88,519,689	81,348,221	92%	6,832	41	430,183	11,907	1,261	35,322
Mitochondrion					16,367	16,367	100%	1		

analysis suggests that the minimum difference between the cattle and buffalo genome is estimated to be 1.4%. The variations occurring in coding regions were first investigated. Of the total 40 million variations, there are only about 1.0% (409,751) in the protein coding regions. As the total protein coding regions span about 1.3% of the cattle genome (based on the Ensembl cattle genome annotation), the frequency of variations in protein coding regions, as commonly expected, is lower than those non-protein coding regions ($p<0.001$). The regions with both the highest and lowest variation density were identified across the protein coding regions of the cattle genome. A density distribution was computed and regions with a variation density greater than 0.04 were classified as high variation density regions while regions of variation density equal to 0 were classified as low variation density regions (density peak at 0.01). From the high variation density regions, 49 unique Ensemble parent IDs were identified, of which 21 had annotated gene information. It is assumed that these regions of large genetic variation between cattle and buffalo indicate the most divergent areas between the two species. Among the proteins that these genes code for, olfactory receptor (OR) family proteins were represented most (6 occurrences). This particular gene family is of interest as it been reported that there are over 1,800 genes in this family in the cattle genome, making up approximately 5% of the total annotated genes (total of ~33,000 annotated genes). The next most represented protein is trophoblast Kunitz domain protein 4, which is identified twice, than other proteins such as S100 calcium binding protein A8, secretoglobin, secretory leukocyte peptidase inhibitor, cathelicidin antimicrobial peptide, and matrix metallopeptidase 3 are represented once.

The genome regions of low variation density suggest the more conserved regions between the cattle and buffalo, since genetic variation and species divergence is assumed to be very low. As expected from the well known similarity between these two species, there are many more genomics regions with low variation density than high variation density. In fact, 1,144 Ensembl parent IDs map to the regions of low variation density, of which 530 had annotated gene

Table 3. Gap summary statistics for buffalo read mapping across the Btau_4.0 chromosomes

Chromosome		Count	Total Length	Max	Mean	Median	N90
Cattle	Buffalo						
1.	1q	12,604	12,422, 788	154,597	986	63	435
2.	2q	9,643	8,701,081	459,542	902	64	440
3.	6	8,837	9,553,167	158,059	1,081	70	510
4.	8	8,590	7,430,269	173,814	865	64	422
5.	4q	9,654	9,319,764	188,898	965	74	526
6.	7	10,268	10,468,954	195,893	1,020	66	430
7.	9	8,432	7,577,346	186,082	899	68	490
8.	3q	7,893	9,345,494	220,962	1,184	68	498
9.	10	8,586	7,502,874	224,289	874	67	457
10.	11	7,238	6,376,160	162,856	881	70	405
11.	12	6,414	6,816,446	180,510	1,063	66	446
12.	13	7,562	5,530,978	136,214	731	72	435
13.	14	3,842	4,998,578	136,330	1,301	67	590
14.	15	4,985	4,590,510	132,669	921	67	452
15.	16	7,990	6,726,388	206,601	842	78	476
16.	5q	5,299	6,005,124	154,566	1,133	71	502
17.	17	5,217	5,952,971	187,766	1,141	65	486
18.	18	3,856	5,328,188	166,761	1,382	77	577
19.	3p	3,459	4,716,300	156,849	1,363	76	636
20.	19	5,443	5,452,323	228,494	1,002	65	431
21.	20	4,857	4,596,072	205,014	946	66	417
22.	21	2,671	3,398,461	124,586	1,272	66	546
23.	2p	3,446	3,482,709	137,017	1,011	69	473
24.	22	3,895	4,053,164	144, 114	1,041	66	467
25.	24	2,144	2,785,645	171,253	1,299	69	545
26.	23	3,775	3,894,038	146,070	1,032	73	469
27.	1p	4,276	4,014,016	181,093	939	90	554
28.	4p	3,012	3,278,479	216,333	1,088	67	489
29.	5p	4,400	4,676,674	151,937	1,063	77	476
X	X	6,831	7,171,468	154,075	1,050	86	825

information. To best summarize the large list of genes, the information was mapped to Gene Ontology (GO) terms (biological process) and the top 10 terms were plotted in Fig, 1. G-protein coupled receptor protein signaling was the most enriched biological process among the low variation density regions, of which OR genes are part of this family. The next two most enriched biological processes are regulation of transcription and translation. Then GTPase signal transduction, multicellular organism development, defense response, and other more general biological processes.

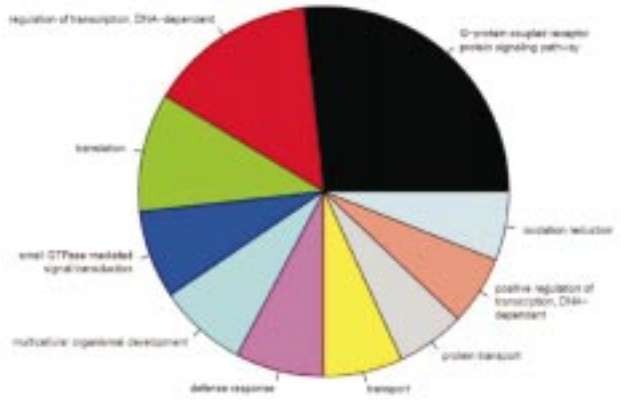

Fig. 1. Pie chart representing the distribution of enriched GO terms (biological response) from gene coding regions containing no variation between cattle and buffalo

Structural variant (SV) detection

In a reference-based assembly, not only can genetic variation at nucleotide level but also structural variants (SVs) between the buffalo and cattle genomes. One approach for identifying such instances is with paired-end mapping (PEM) combined with the depth of read coverage (reviewed by Medvedev et al., 2009). An example of this is provided in Fig. 2, where a deletion of about 400 bp occurs in the buffalo genome (corresponding to the cattle chromosome 2), supported by a dozen paired end mates and a dramatic decrease in read depth. This type of structural difference between the buffalo and cattle can be detected using PEM. To uncover these instances of genomic rearrangements, indicating the divergence of these two genomes, we employed

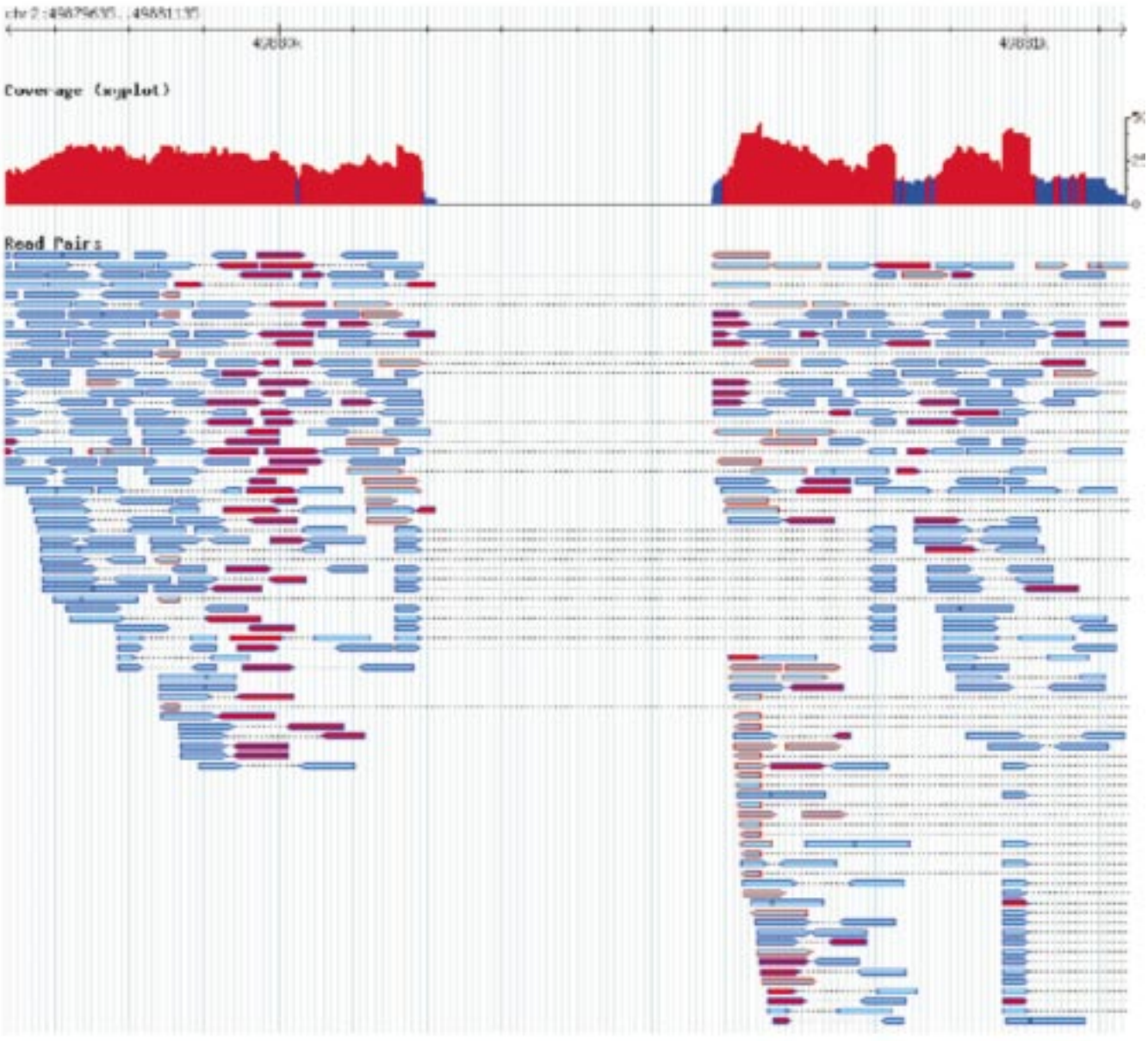

Fig. 2. An example of structural variant between buffalo and cattle. The track of coverage gives an xy plot of the depth of the reads within the selected region, that is, between 49,879,635 bp and 49,881,135 bp in the chromosome 2 of the cattle genome. Buffalo sequence reads are represented by bars in the track of "Read Pairs" in one of the two colors: cyan, if the mapping of the read is unique; red, otherwise. The orientation of the reads are marked by the arrow of the bar and the two mates from the same pair are connected by the horizontal dashed line, where "proper" pairs are outlined in blue and "improper" pairs in red. The gap in the coverage track suggests an insertion of about 400 bp in the central region of the cattle genome, which is absent in the buffalo genome.

a simple approach and identified 287 SV candidates in this study. Many (66%) of the linking signatures span within the same chromosome in a "proper" orientation, with a median insert size of 4.2 Kb, suggesting insertion signatures are common in these candidates. Further manual inspection using BLAST searches of the sequence regions revealed that the top hits from most, if not all, of these SV are related to transposable elements.

Comparison of the buffalo pseudomolecule assembly to the Btau 4.0 cattle assembly using additional publicly available Buffalo reads

To provide a source of quality control for the contigs and buffalo pseudomolecules constructed, we obtained all available 563 nucleotide sequences from the Murrah Buffalo available in GenBank (including 275 cDNA, 216 EST, and 72 genome survey sequences). A BLAST search was then

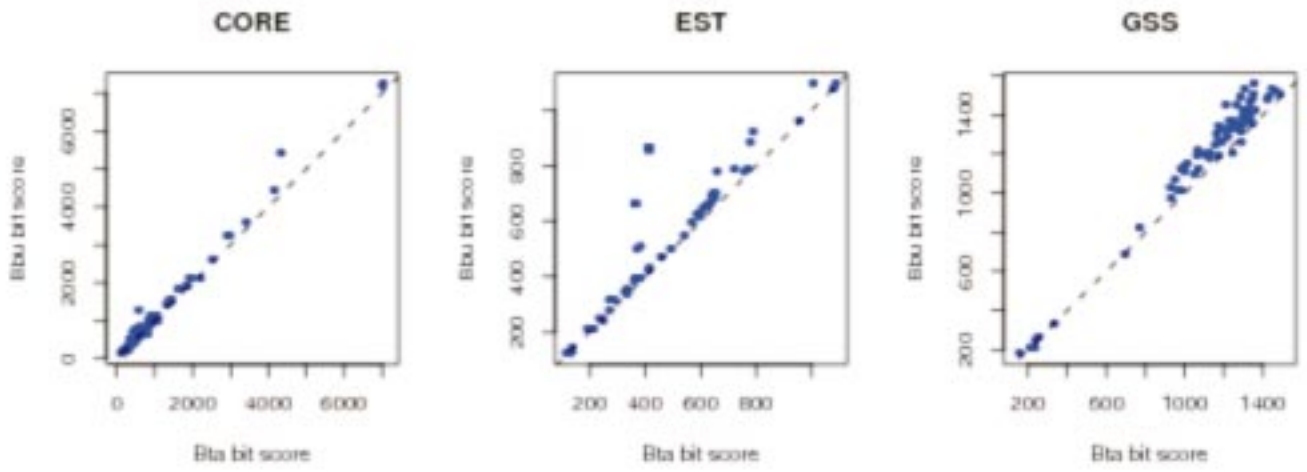

Fig. 3. Comparison of sequence similarity for publicly available buffalo reads to the buffalo pseudomolecule assembly versus the Btau 4.0 cattle genome assembly. The dotted lines represent the equation y=x. Thus, points above the dotted line indicate that the sequences have higher similarity to the buffalo assembly than the cattle genome.

performed against both our buffalo pseudomolecule assembly and the Btau 4.0 cattle genome assembly. The bit scores for the top hit corresponding to each of the query sequences was obtained when searching against both assemblies and compared. For all three sequence classifications (cDNA, EST, and GSS), the bit scores of the BLAST search were always higher or similar when aligning to the buffalo pseudomolecules, as compared to the Btau 4.0 genome assembly (Fig. 3).

DISCUSSION

This work illustrates the utility in using a closely related mammalian genome to construct a meaningful reference-guided assembly, in the absence of adequate sequence content. Illumina short read sequences generated from a river buffalo were mapped to the Btau 4.0 release cattle genome. The assembly of both paired end and mate pair buffalo reads into pseudomolecules has an average read depth of 17-19X and 91%-95% coverage to that of the cattle genome. Variation that distinguish buffalo from cattle were identified throughout the genome and the protein-coding regions were associated with biological processes that are most enriched to better understand genetic diversity between these two species. OR family proteins were most prevalent among those protein coding regions with the highest variation density. This gene family has been reported as an example of the birth-and-death evolutionary model, which proposes that within a gene family, gene duplication events occur, creating new genes that persist in the genome for a long period of time, while other genes are inactivated, or obliterated from the genome over time (Nei and Roopney 2005). Such a model differs from concerted evolution, which supports members of a gene family evolving together as a group (Nei and Roopney 2005). The enrichment of genetic variation within the OR gene family between the buffalo and cattle supports the secondary point of this evolutionary theory, where certain genes are inactivated or omitted over time. Similar to the divergence in this OR gene family observed between fish and mammals, where different species adapt to acquire different odorants, cattle and buffalo show a similar pattern for enriched genetic variation in this gene family.

Among those regions with no genetic variation between cattle and buffalo, G-protein coupled receptor protein signaling pathways were most prevalent, of which OR family proteins are associated. In contrast to the variation -enriched OR region, the minimization of genetic variation between cattle and buffalo within this gene family supports the arm of the birth-and-death evolutionary model that explains gene duplication events that persist over a long period of time (Nei and Roopney 2005).

A limitation of this study lies in the reference-based methodology of using a different species for the reference genome from the species that was sequenced. With no known buffalo reference genome, the cattle was selected - the evolutionarily closest sequenced species. Though the buffalo and cattle have high homology, differences in both structural rearrangements and repetitive elements between the two species are not completely resolved in this assembly. Additionally, the sparsity of the RH map that was used to guide the buffalo pseudomolecule assembly is a limitation. With only ~2,700 markers to guide the buffalo assembly, regions where few markers were identified, or large distances between markers, can harbor structural rearrangements in the buffalo genome, compared to the cattle genome, that were not completely accounted for in this buffalo sequence assembly. We do, however, demonstrate that the paired end mapping approach (Medvedev et al., 2009) can identify regions of SVs between these two genomes. We also found that the top SV candidates, when the sequence regions were BLAST searched, were associated with transposable

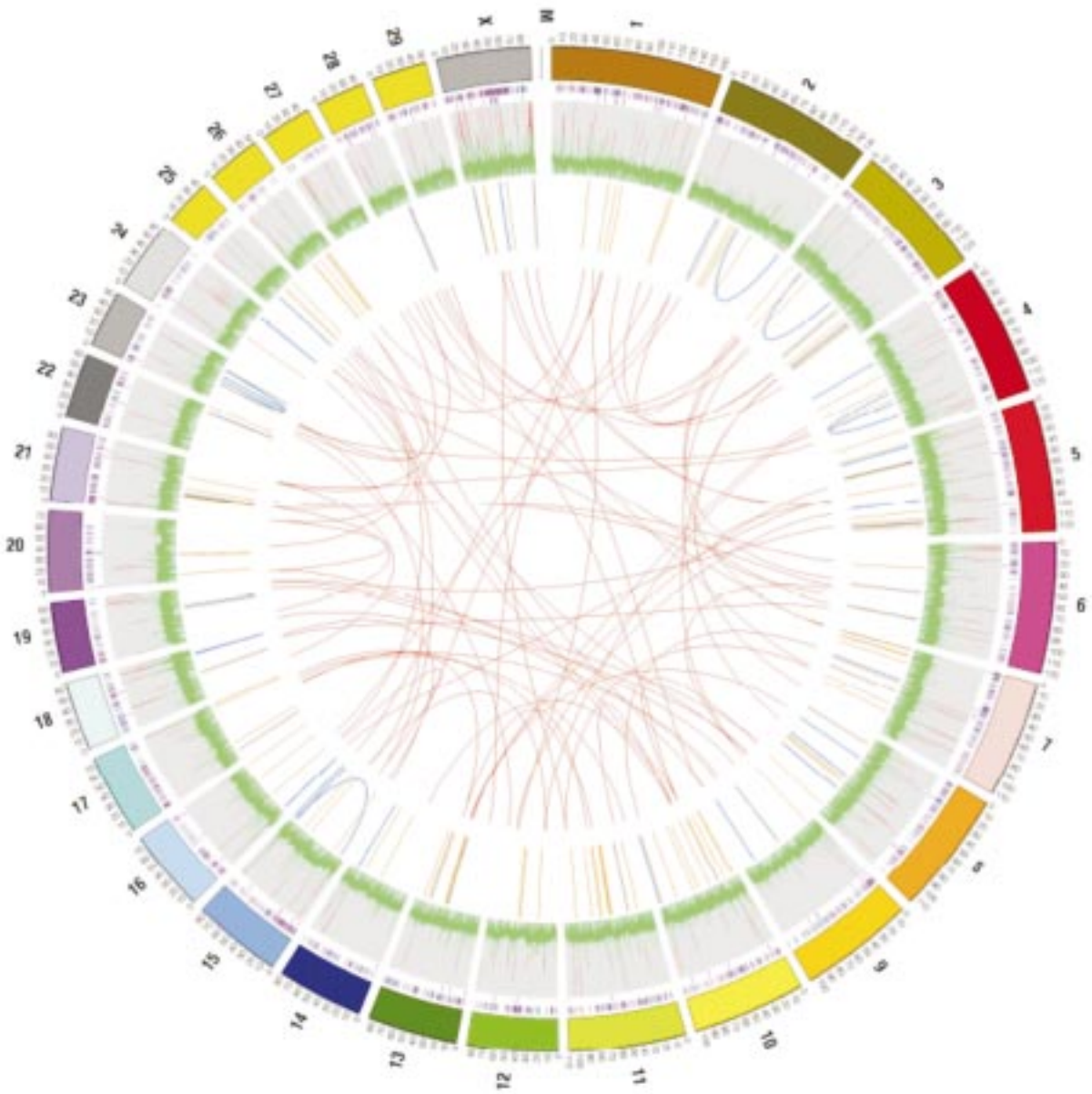

Fig. 4. Summary plot of buffalo reads mapped to the cattle chromosomes. Information included is as follows: (outside track) regions of interrupted repeats in 10 Kb steps with long interspersed elements (LINES) as purple tiles, short interspersed elements (SINES) as red tiles, and all other repeat classes (long terminal repeats (LTRs), simple repeats, etc.) as blue tiles; (second track) read depth (green = depth _ 60; red = depth > 60); (third track) putative regions of insertions, inversions, or duplications in the buffalo genome relative to the cattle genome within a 10 Kb window (orange lines) and greater than a 10 Kb window (blue lines) on the same cattle chromosome; (fourth track) the same mapping patterns as track three between read pairs of 200 bp insert size, but between different cattle chromosomes.

elements. A summary overview for the assembly by chromosome, inclusive of repeat regions, read depth, and SVs are represented in Fig. 4.

Though our buffalo pseudomolecules are still a preliminary draft, our buffalo contigs may serve as stable units for the reconstruction of the new buffalo assembly (both contigs and pseudomolecules are available at http://210.212.93.84/data/buffalo_v2.0). A more comprehensive RH map would benefit this assembly and help identify those major structural differences between the two genomes. We implemented a quality control approach to demonstrate the specificity of the pseudomolecule assembly to the buffalo genome from that of the cattle reference genome. An independent set of publicly available Murrah Buffalo core, EST, and genome survey sequences (GSS) were aligned against both animal assemblies and all query sequences

showed higher bit scores for alignment to the buffalo pseudomolecule assembly than the Btau 4.0 genome assembly. This suggests that the pseudomolecule buffalo assembly is more specific to buffalo than cattle, which is of particular importance for exhibiting specificity, since the cattle genome was used as a reference in the buffalo assembly.

Future work will focus on the refinement of this genome and incorporation of *de novo* approaches to better identify large structural rearrangements in the buffalo genome. As more sequence data becomes available for the Water Buffalo, this assembly will continue to be refined and provide a great resource for continued genome sequencing work in buffaloes.

ACKNOWLEDGEMENT

The whole genome sequencing and bioinformatic work was outsourced to M/s Sandor Proteomics Pvt Ltd, Hyderabad. The help rendered in assembly analysis and for generating buffalo genome browser resources is duly acknowledged.

REFERENCES

Amaral M E, Grant J R, Riggs P K, Stafuzza N B, Filho E A, Goldammer T, Weikard R, Brunner R M, Kochan K J, Greco A J, Jeong J, Cai Z, Lin G, Prasad A, Kumar S, Saradhi G P, Mathew B, Kumar M A, Miziara M N, Mariani P, Caetano A R, Galvao S R, Tantia M S, Vijh R K, Mishra B, Kumar S T, Pelai VA, Santana A M, Fornitano L C, Jones B C, Tonhati H, Moore S, Stothard P and Womack J E. 2008. A first generation whole genome RH map of the river buffalo with comparison to domestic cattle. *BMC Genomics* **9**: 631. PubMed PMID: 19108729; PubMed Central PMCID: PMC2625372.

Altschul S F, Gish W, Miller W, Myers E W and Lipman D J.1990. Basic local alignment search tool". *Journal of Molecular Biology* **215**: 403–10. doi:10.1006/jmbi.1990.9999. PMID 2231712

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27:** 573-80.

FAO, 2009. http://faostat.fao.org

Harris E Y, Ponts N, Levchuk A, Roch K L and Lonardi S. 2010. BRAT: bisulfite-treated reads analysis tool. *Bioinformatics* **26**: 572-73 Erratum in: *Bioinformatics* **26**:2499. PubMed PMID: 20031974.

Iannuzzi L and Di Meo G. 2009. Water Buffalo. In: Cockett NE and Kole C, Editors. Genome Mapping and Genomics in Domestic Animals. Berlin Herdelberg, Germany: Springer-Verlag.

Jiang Z, Rokhsar D S, and Harland R.M. 2009. Old can be new again: HAPPY whole genome sequencing, mapping and assembly. *International Journal of Biological Sciences* **5**: 298-303.

Krzywinski M, Schein J, Birol I, Cannors J, Gascoyne R, Horsman D, Jones S J and Marra M A. 2009. Circos: an Information Aesthetic for Comparative Genomics. *Genome Research* **19**: 1639-45

Li H and Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. **25**: 1754-60. PubMed PMID:19451168; PubMed Central PMCID: PMC2705234.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R. 2009. 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**: 2078-79. [PMID: 19505943]

Liu Y, Qin X, Song X Z, Jiang H, Shen Y, Durbin K J, Lien S, Kent M P, Sodeland M, Ren Y, Zhang L, Sodergren E, Havlak P, Worley K C, Weinstock G M and Gibbs R A. 2009. Bos taurus genome assembly. *BMC Genomics* **10**: 180. PubMed PMID: 19393050; PubMed Central PMCID: PMC2686734.

Medvedev P, Stanciu M and Brudno M. 2009. Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods* **6**: S13-20. PubMed PMID: 19844226.

Michelizzi V N, Dodson M V, Pan Z, Amaral M E, Michal J J, McLean D J, Womack J E and Jiang Z. 2010. Water Buffalo Genome Science Comes of Age. *International Journal of Biological Sciences* **6**: 333-49

Nei M and Roopney A P. 2005. Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics* **39**: 121-52.

R Development Core Team. 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Roth J and Myers P. 2004. Bubalus bubalis. http://animaldiversity.ummz.umich.edu/site/ accounts/information/Bubalus_bubalis.html.

Smit A F A, Hubley R and Green P. 1996. *RepeatMasker Open-3.0*. www.repeatmasker. org

Stafuzza N B, Abbassi H, Grant J R, Rodrigues-Filho E A, Ianella P, Kadri S M, Amarante M V, Stohard P, Womack J E, de León F A, Amaral M E. 2009. Comparative RH maps of the river buffalo and bovine Y chromosomes. *Cytogenetics and Genome Research* **126**: 132-38.

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