Intended for journal: G3 Genes|Genomes|Genetics – Genome Reports

Draft genome of the lowland anoa (*Bubalus depressicornis*) and comparison with buffalo genome assemblies (Bovidae, Bubalina)

Stefano Porrelli [1]; Michèle Gerbault-Seureau[2] ; Roberto Rozzi [3,4]; Rayan Chikhi [5]; Manon Curaudeau [2]; Anne Ropiquet [1]; Alexandre Hassanin [2]*

[1] Middlesex University, Department of Natural Sciences - Faculty of Science and Technology, The Burroughs, Hendon, London, NW4 4BT.

[2] Institut Systématique Evolution Biodiversité (ISYEB), Sorbonne Université, MNHN, CNRS, EPHE, UA, 57 rue Cuvier, CP 51, 75005 Paris, France.

[3] Museum für Naturkunde, Leibniz-Institut für Evolutions- und Biodiversitätsforschung, 10115 Berlin, Germany.

[4] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Synthesis Centre for Biodiversity Sciences (sDiv), Puschstr. 4, D-04103 Leipzig, Germany.

[5] Institut Pasteur, Université Paris Cité, Sequence Bioinformatics, 28 rue du Docteur Roux, 75015 Paris, France.

***Corresponding author**: Hassanin, Alexandre – alexandre.hassanin@mnhn.fr

Keywords: Bovidae, *Bubalus depressicornis*, lowland anoa, genome assembly, *de novo* assembly

Abstract

Genomic data for wild species of the genus *Bubalus* (Asian buffaloes) are still lacking while several whole genomes are currently available for domestic water buffaloes. To address this, we sequenced the genome of a wild endangered dwarf buffalo, the lowland anoa (*Bubalus depressicornis*), produced a draft genome assembly, and made comparison to published buffalo genomes.

The lowland anoa genome assembly was 2.56 Gbp long and contained 103,135 contigs, the longest contig being 337.39 kbp long. N50 and L50 values were 38.73 kbp and 19.83 kbp, respectively, mean coverage was 44x and GC content was 41.74%. Two strategies were adopted to evaluate genome completeness: (i) determination of genomic features with *de novo* and homology-based predictions using annotations of chromosome-level genome assembly of the river buffalo, and (ii) employment of benchmarking against universal single-copy orthologs (BUSCO). Homology-based predictions identified 94.51% complete and 3.65% partial genomic features. *De novo* gene predictions identified 32,393 genes, representing 97.14% of the reference's annotated genes, whilst BUSCO search against the mammalian orthologues database identified 71.1% complete, 11.7% fragmented and 17.2% missing orthologues, indicating a good level of completeness for downstream analyses. Repeat analyses indicated that the lowland anoa genome contains 42.12% of repetitive regions. The genome assembly of the lowland anoa is expected to contribute to comparative genome analyses among bovid species.

1. Introduction

The lowland anoa, *Bubalus depressicornis* (C. H. Smith, 1827), is a wild dwarf buffalo endemic to Sulawesi and Buton Islands, where it can be found in sympatry with the mountain anoa, *Bubalus quarlesi* (Ouwens, 1910). Both anoa species are currently classified as endangered with declining populations due to hunting and habitat loss (Burton et al. 2016). Because of their singular appearance, they were initially described in their own genus *Anoa* (Ouwens 1910). However, *Anoa* was not regarded as a valid genus in more recent classifications, in which both anoa species were ascribed to the genus *Bubalus*, together with the wild water buffalo – *Bubalus arnee* (Kerr, 1792) and the tamaraw - *Bubalus mindorensis* Heude, 1888 (Groves 1969; IUCN 2022). Molecular studies based on mitochondrial sequences have

supported a sister-group relationship between *Bubalus depressicornis* and *Bubalus quarlesi* (Schreiber et al., 1999; Priyono et al., 2020). In addition, the mitogenome of the lowland anoa was found to be equally distant from those of the two types of domestic water buffalo, the river buffalo from the Indian subcontinent and Mediterranean countries and the swamp buffalo from China and Southeast Asia (Hassanin et al., 2012). Since the same phylogenetic pattern was recovered from the analyses of two nuclear datasets, one based on 30 autosomal genes and the other based on two genes of the Y chromosome, Curaudeau et al. (2021) have concluded the existence of two species of domestic buffaloes: *Bubalus bubalis* (Linnaeus, 1758) for the river buffalo and *Bubalus kerabau* Fitzinger, 1860 for the swamp buffalo, which diverged during the Pleistocene at around 0.84 Mya. As discussed in Curaudeau et al. (2021), the two domestic species can easily be distinguished based on coat and horn characteristics (Castelló 2016), and they have different karyotypes: *Bubalis bubalis* has 2n = 50 chromosomes with a fundamental number (FN) equal to 58; whereas *Bubalus kerabau* has 2n = 48 chromosomes and FN = 56 (Nguyen et al., 2008).

With rapid progress and cost reduction in sequencing technologies, many whole genomes of domestic bovid species have been sequenced. Whole-genome sequencing has allowed the identification of variants involved in domestication and genetic improvement for several livestock species such as cattle and buffaloes (Zimin et al., 2009; Canavez et al., 2012; Li et al., 2020; Rosen et al., 2020). Chromosome-level genome assemblies include those of the domestic cow, *Bos taurus* (Zimin et al., 2009), the domestic river buffalo, *Bubalus bubalis* (Deng et al., 2016), the swamp buffalo, *Bubalus kerabau* (reported as *Bubalus carabanensis* in Luo et al. (2020) but see Curaudeau et al. (2021) for further taxonomic information), the domestic Yak, *Bos grunniens* (Zhang et al., 2021) and the zebu cattle, *Bos indicus* (Canavez et al. 2012). Whereas a total of eight chromosome- and scaffold-level genome assemblies are publicly available for domestic buffaloes, there is currently no genome data available for wild species of the genus *Bubalus*. To fill this gap, a biopsy of a living lowland anoa was used for next-generation sequencing, and a draft genome was assembled *de novo* for comparison to other buffalo genome assemblies available in international databases such as NCBI (National Center for Biotechnology Information) and BIG_GWH (Beijing Institute of Genomics Genome Warehouse database).

2. Material & Methods

2.1 DNA extraction, library preparation and genome sequencing

A living male adult of lowland anoa, named Yannick, was sampled at the *Ménagerie du Jardin des Plantes* of the Muséum national d'Histoire naturelle (MNHN, Paris, France) (Figure 1). A skin biopsy was performed in 2006 by a veterinary surgeon following protocols approved by the MNHN and in line with ethical guidelines. The same biopsy was previously used to determine its karyotype (2n = 48; FN = 58; Nguyen et al., 2008). DNA was extracted using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) following the manufacturer's protocol. DNA quantification was performed with a Qubit® 2.0 Fluorometer with Qubit dsDNA HS Assay Kit (Thermo Fischer Scientific, Walthan, MA, USA). Library preparation and sequencing were conducted at the *Institut du Cerveau et de la Moelle épinière*. The sample was sequenced on a NextSeq® 500 Illumina system generating 2 X 151 bp reads using the NextSeq 500 High Output Kit v2 with 300 cycles and aiming for an insert size of 350 bp.

2.2 De novo assembly

Data quality was assessed with FastQC v.0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and results were collated with MultiQC v1.12 (Ewels et al., 2016). Raw reads were quality trimmed and adapter sequences and contaminants removed with Trimmomatic v.0.36 (Bolger et al., 2014) with the following parameters: "ILLUMINACLIP: TruSeq3 -PE.fa:2:30:10 LEADING:33 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36". Data quality of quality-trimmed reads was re-assessed with FastQC. A *de novo* assembly was performed with MaSuRCA v.3.3.1 (Zimin et al., 2013; Zimin et al., 2017) using recommended parameters for mammalian genomes and paired-end Illumina-only data, as indicated in Zimin et al. (2017). Mean and standard deviation for the Insert size were estimated with an "estimate-insert-size" script (https://gist.github.com/rchikhi/7281991). Paired-end reads were error corrected using QuorUM (Marçais et al., 2015) and assembled into super-reads using a k-mer size of 99, as selected by the MaSuRCA assembler. The super-reads were then assembled into contigs using the CABOG assembler, part of the MaSuRCA pipeline (Zimin et al., 2017), followed by gap closing with the paired-end information (Zimin et al., 2013).

2.3 Assembly quality assessment

Genome assemblies publicly available for *Bubalus* and *Syncerus* genera were retrieved from NCBI and BIG_GWH for quality comparison and assessment. The dataset included two assemblies at the chromosome level for the river buffalo (*Bubalus bubalis*) with a coverage of 100x and 572x, four scaffold-level draft assemblies of river buffalo with coverage ranging between 69x and 119x, one chromosome-level assembly of swamp buffalo (*Bubalus kerabau*) with a mean coverage of 65x, and one scaffold-level draft assembly of the African buffalo (*Syncerus caffer*) with 162x coverage. The eight retrieved assemblies were sequenced and assembled with different methods, summarised in Table 1.

The quality of the lowland anoa genome assembly was assessed with QUAST-LG v.5.0.1 (Mikheenko et al., 2018) using the river buffalo NDDB_SH_1 genome assembly (Deng et al., 2016) as a reference. The default parameters for mammalian genomes were used to compare all assemblies in QUAST-LG: "MODE: large, threads: 50, eukaryotic: true, minimum contig length: 3,000, minimum alignment length: 500, ambiguity: one, threshold for extensive misassembly size: 7,000". All analysed assemblies were aligned to the river buffalo NDDB_SH_1 assembly and results were plotted with Circos v. 0.69.8 (Krzywinski et al., 2009) and Jupiter consistency plots (Chu, 2018).

We adopted two different strategies to evaluate genome completeness. Firstly, genomic features were predicted with the homology-based method by aligning the lowland anoa genome to that of the annotated river buffalo reference genome (NDDB_SH_1 and relative annotations retrieved from NCBI). Secondly, we used a *de novo* gene prediction method with GlimmerHMM v3.0.4 (Majoros et al. 2004). Thirdly, we employed benchmarking against universal single-copy orthologs (BUSCO v5.2.2; Manni et al. 2021) using the mammalia_odb10 dataset (19/02/2021, number of genomes: 24, number of BUSCOs: 9226) from OrthoDB (Kriventseva et al. 2019) and compared to other buffalo genome assemblies already deposited on NCBI and BIG_GWH (Table 1).

2.4 Repeats and gene annotation

Repetitive regions in the lowland anoa genome were identified, annotated and masked with RepeatMasker v.4.1.2-p1 (Tarailo-Graovac and Chen, 2009). Firstly, a *de novo* repeat library was constructed from the genome assembly with RepeatModeler v.2.0.2a. RepeatMasker was used with default parameters to produce a homolog-based repeat library and mask the genome's repetitive regions. The scripts "*calcDivergenceFromAlign.pl*" and

159 "*createRepeatLandscape.pl*" were used to calculate the Kimura divergence values and to plot
160 the resulting repeat landscape. The repeat landscape of *Bos taurus* was retrieved from the
161 RepeatMasker database for visual comparison.
162
163 ## ***3. Results & Discussion***
164 ### *3.1 Whole-genome sequencing and data QC*
165 Whole-genome sequencing generated 991,437,058 paired-end reads with a length of 151 bp.
166 Quality trimming removed 46,616,722 low quality, adapter-contaminated and PCR-
167 duplicated reads, representing approximately 0.5% of the total reads. A total of 944,820,336
168 clean paired-end reads were generated, covering the lowland anoa genome with an
169 estimated 56x depth based on a genome size of 2.56 Gbp. Estimation of insert size using in-
170 house script returned a mean of 377 and a standard deviation of 83.
171
172 ### *3.2 De novo assembly quality metrics*
173 The final lowland anoa genome assembly generated here contained 103,135 contigs, the
174 largest being 337.39 kbp long, an N50 of 38.73 kbp and an L50 of 19.83 kbp (Table 2). Total
175 length was 2.56 Gbp with a mean coverage of 44x, and GC content was 41.74%, in agreement
176 with other published assemblies (between 41.60% and 41.92%, Table 3). When aligned to the
177 NDDB_SH_1 genome assembly, the fraction of the anoa genome assembly was 95.41%, a
178 value comparable to other buffalo genome assemblies (Figure 2), with a total alignment
179 length of 2,515,453,843 bp. A total of 886 contigs could not be aligned to the river buffalo
180 genome assembly, whilst 8,085 contigs were only partially aligned, resulting in a total
181 unaligned length of 45,224,171 bp, which reflects the discrepancy between the total length
182 of the lowland anoa genome and the total aligned length to the reference river buffalo
183 genome assembly. Partially aligned and unaligned contigs could have resulted from structural
184 variations between the lowland anoa and the reference river buffalo assembly, such as large
185 INDELS (insertion/deletions), as well as repetitive regions and/or alternative haplotypes
186 causing assembly errors. The nature of short-read technology causes difficulties in
187 characterising genomic regions such as telomeres, centromeres, repetitive and highly
188 heterochromatic regions (Johnson et al. 2005; Low et al. 2019; Weissensteiner and Suh 2019),
189 which are notoriously difficult to assemble and could be better resolved with long-read
190 sequencing.

The lowland anoa genome assembly has a modest N50 compared to other buffalo genome assemblies (Table 3), indicating lower levels of contiguity, which is expected due to the short-read output of Illumina sequencing technology (read length = 151 bp). Additionally, repeat analysis revealed that 42.12% of the lowland anoa genome is composed of repetitive regions. This, coupled with low sequence coverage, sequencing and assembly errors, causes breaks in the assembly contiguity (Gnerre et al., 2011; Low et al., 2019). This is apparent even in high-quality chromosome-level genome assemblies that use multiple sequencing libraries and multiple sequencing technologies, such as the previous human genome assembly GRCh38, which contained hundreds of gaps (International Human Genome Sequencing Consortium 2004). In addition, the chromosome-level genome assemblies retrieved from NCBI (NDDB_SH_1, UOA_WB_1) were sequenced using multiple insert size libraries and sequencing technologies and were intensively verified with multiple methods such as optical mapping, Hi-C and RH (Deng et al., 2016; Low et al., 2019).

Moreover, quality metrics of publicly available assemblies are usually limited to reporting N50 and L50 values, which represent the shortest contig length needed to cover 50% of the total assembly size, and the number of contigs whose cumulative length covers 50% of the total assembly size, respectively (Bradnam et al., 2013). Such metrics are often used to compare and evaluate performances of the ever-growing assembly and annotation methods and software (Manchanda et al., 2020). However, we hereby show that reporting N50 and L50 metrics exclusively can be misleading, as they only provide a standard measure of assembly contiguity whilst omitting information such as gene content and completeness, as well as assembly correctness. Furthermore, N50 values can be artificially raised by deliberately excluding short contigs from analyses and by the presence of undetermined nucleotides (Ns) linking the scaffolded contigs (Gurevich et al. 2013). Therefore, to assess the quality of the lowland anoa genome assembly, we generated conventional N50 and L50 metrics and also determined genome completeness in terms of gene content and genome correctness by comparing our assembly to a chromosome level genome assembly of the river buffalo (*Bubalus bubalis*). Additionally, a swamp buffalo (*Bubalus kerabau,* CUSA_SWP) and a more distantly related African buffalo species (*Syncerus caffer*, ABF221) were also included in our comparison.

Regardless of the modest N50 value, the lowland anoa genome assembly is in good agreement with the NDDB_SH_1 assembly, with 95.91% of contigs correctly mapped to the

25 reference chromosomes of the river buffalo and fewer misassembled blocks compared to other draft assemblies (Figure 3). The genome assembly of the Egyptian river buffalo (EGYBUF_1.0) had an abnormally high number of misassembled blocks with respect to the reference genome, followed by the genome assembly of a female Italian river buffalo (UOA_WB_1). To investigate this, misassemblies and structural variation metrics were computed in QUAST-LG (Table 4). The Egyptian river buffalo assembly (EGYBUF_1.0) showed the highest number of mismatches and the highest number of Ns, followed by the Jaffrabadi river buffalo (AAUIN_1). The genome assembly of the African buffalo (*S. caffer*, ABF221) showed a larger number of mismatches (Table 4), but this can be explained by the higher sequence divergence between *Syncerus* and *Bubalus*, as the two genera have separated in the Late Miocene (Hassanin et al., 2012). Misassemblies and structural variation metrics could not explain the misassembled blocks of the UOA_WB_1 assembly observed in the Circos plot of Figure 3. However, some of these misassembled blocks could be due to unplaced contigs. To investigate this, the UOA_WB_1 assembly was aligned to the NDDB_SH_1 reference to generate Jupiter consistency plots. When using the largest 26 contigs of the UOA_WB_1 assembly to cover 100% of the reference river buffalo genome, an almost perfect level of synteny was observed (Figure 4a). Although this result was expected for genomes of the same species, it also indicates a good level of assembly quality in terms of correctness. However, when including all 509 contigs of the UOA_WB_1 assembly, several misassembled regions were observed (Figure 4b). Three non-exclusive hypotheses can be advanced to interpret this result: possible genomic rearrangements, genome assembly errors, and repetitive regions. Whether the results of the consistency plots are due to the factors mentioned above or other factors, such as contamination, remains speculative. Nevertheless, the results of the quality metric comparison conducted here further indicate the unreliability of using exclusively N50 and L50 metrics when assessing assembly quality. Instead, contiguity metrics should be supplemented with genome completeness and correctness metrics.

3.3 Genomic features, gene prediction and annotation

Homology and *de novo* gene predictions performed on the lowland anoa genome assembly were in agreement with each other and indicated a good level of genome completeness. Results were comparable to other published genome assemblies (Tables 5 and 6), and an

improvement over the Bangladeshi river buffalo (Bubbub_1.0), the Egyptian river buffalo (EGYBUF_1.0) and Mediterranean river buffalo (UMD_CASPUR_WB_2.0) assemblies. Interestingly, these three assemblies showed higher contiguity (N50) than the draft assembly of the lowland anoa, further indicating the unreliability of using exclusively N50 and L50 metrics when assessing genome assembly quality.

Out of the 1,921,249 genomic features annotations of the reference assembly NDDB_SH_1, homology prediction identified 1,815,794 (94.51%) complete and 69,929 (3.63%) partial features in the lowland anoa genome assembly, which is comparable to other published assemblies (Figure 5), indicating a good level of genome completeness. GlimmerHMM *de novo* predicted 1,027,469 unique genomic features (mRNA and coding sequences, CDS), which is an improvement over some of the water buffalo assemblies used for quality comparison (Table 5). Homology-based gene prediction identified 32,393 genes in the lowland anoa genome assembly, representing 97.14% of the genes annotated in NDDB_SH_1 (n= 33,348). Of these, 59.11% (19,148) were complete and 40.88% (13,245) were partial, probably reflecting the level of fragmentation of the lowland anoa genome assembly. Nevertheless, the total number of genes predicted still represents an improvement over some of the compared assemblies (Table 6).

When predicting mammalian orthologs with BUSCO, the lowland anoa genome assembly contained 6,556 (71.1%) complete BUSCOs, of which 6,412 (69.5%) were single-copy and 144 (1.6%) were duplicated. The number of fragmented BUSCOs was 1,076 (11.7%), whilst 1,594 (17.2%) were missing. The BUSCO results indicate an acceptable level of genome completeness (<70%, Simão et al., 2015) for downstream analyses for the anoa genome assembly, and a slight improvement over the Egyptian river buffalo assembly (EGYBUF_1.0, Figure 6).

Mammalian genomes contain large families of repeats (Goodier and Kazazian, 2008), such as long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and long-terminal repeats (LTRs). RepeatMasker revealed that 42.12% of the lowland anoa genome is composed of repetitive regions (Table 7), which is comparable to data previously published for genome assemblies of river buffalo and other bovids (Deng et al., 2016; Low et al., 2019; Mintoo et al,. 2019; El-Khishin et al., 2020). Results also agree with the repetitive content in the cattle genome (Figure 7b). Both lowland anoa and cattle genomes showed two waves of repeat expansion in their repeat landscape (Figure 7a and 7b), suggesting a shared

inheritance of such repeats. In the lowland anoa, the LINEs were more abundant, representing 30.04% of the repeats, followed by LTRs representing 3.10% and SINEs representing 1.03% (Table 7).

4. Conclusion

To date, whole-genome sequencing has allowed identification of variants involved in domestication and genetic improvement for several livestock species (Zimin et al., 2009; Canavez et al., 2012; Li et al., 2020; Rosen et al., 2020). However, the lack of wild buffalo genomes hinders further analyses addressing functional and evolutionary aspects of this group, as well as possible conservation efforts. The draft genome assembly of the lowland anoa reported here is expected to contribute to this gap in data availability, as this is the first draft genome assembly for wild Asian buffaloes. Furthermore, we showed that short-read Illumina sequencing data can still provide a cost-effective way of sequencing mammalian genomes to an adequate level of completeness for downstream comparative analyses.

Data availability

The genome assembly of the lowland anoa is available on NCBI under accession XXXXXXXXXXXXX. The raw data is available on the Sequence Read Archive (SRA) on NCBI under accession XXXXXXXXXX (under embargo until review).

Acknowledgements

We thank the people of the *Ménagerie du Jardin des Plantes* who helped to collect the biopsy of the lowland anoa used in this study: Norin Chai, Gerard Dousseau, Christelle Hano, Abderrahmane Latreche, Claire Rejaud, Roland Simon, and Rudy Wedlarski. The authors would like to thank Huw Jones for the proofreading of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

Funding

R.R. was supported by sDiv, Synthesis Centre of the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, funded by the German Research Foundation

(DFG– FZT 118, 202548816) and by the German Research Foundation (DFG Research grant RO 5835/2-1).

Literature cited:

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 30(15):2114–2120. doi:10.1093/bioinformatics/btu170.

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, et al. 2013. Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species. Gigascience. 2(1):1–31. doi:10.1186/2047-217X-2-10.

Burton J, Wheeler P, Mustari A. 2016. *Bubalus depressicornis*. IUCN Red List Threat Species 2016. e.T3126A46. doi:10.2305/IUCN.UK.2016-2.RLTS.T3126A46364222.

Canavez FC, Luche DD, Stothard P, Leite KRM, Sousa-Canavez JM, Plastow G, Meidanis J, Souza MA, Feijao P, Moore SS, et al. 2012. Genome sequence and assembly of *Bos indicus*. J Hered. 103(3):342–348. doi:10.1093/jhered/esr153.

Castelló JR. 2016. Bovids of the World: Antelopes, Gazelles, Cattle, Goats, Sheep, and Relatives. Princeton, New Jersey, USA: Princeton University Press.

Chu J. 2018. Jupiter plot: a Circos-Based tool to Visualize Genome Assembly Consistency (Version 1.0). Github. [accessed 2022 Feb 22]. https://github.com/JustinChu/JupiterPlot.

Curaudeau M, Rozzi R, Hassanin A. 2021. The genome of the lowland anoa (*Bubalus depressicornis*) illuminates the origin of river and swamp buffalo. Mol Phylogenet Evol. 161(March):107170. doi:10.1016/j.ympev.2021.107170. https://doi.org/10.1016/j.ympev.2021.107170.

Deng T, Pang C, Lu X, Zhu P, Duan A, Tan Z, Huang J, Li H, Chen M, Liang X. 2016. De Novo transcriptome assembly of the Chinese swamp buffalo by RNA sequencing and SSR marker discovery. PLoS One. 11(1):1–20. doi:10.1371/journal.pone.0147132.

El-Khishin DA, Ageez A, Saad ME, Ibrahim A, Shokrof M, Hassan LR, Abouelhoda MI. 2020. Sequencing and assembly of the Egyptian buffalo genome. PLoS One. 15(8 August):1–14. doi:10.1371/journal.pone.0237087. http://dx.doi.org/10.1371/journal.pone.0237087.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: Summarize analysis results for

multiple tools and samples in a single report. Bioinformatics. 32(19):3047–3048. doi:10.1093/bioinformatics/btw354.

Fitzinger LJ. 1860. Der Sunda-Büffel (*Bubalus kerabau*). In: Wissenschaftlich-populäre Naturgeschichte der Säugethiere in ihren sämmtlichen Hauptformen, V. Kaiserlich-Königlichen Hof- und Staatsdruckerei. Wien. p. 329.

Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A. 108(4):1513–1518. doi:10.1073/pnas.1017351108.

Goodier JL, Kazazian HH. 2008. Retrotransposons Revisited: The Restraint and Rehabilitation of Parasites. Cell. 135(1):23–35. doi:10.1016/j.cell.2008.09.022.

Groves CP. 1969. Systematics of the anoa (Mammalia, Bovidae). Beaufortia. 17:1–12.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: Quality assessment tool for genome assemblies. Bioinformatics. 29(8):1072–1075. doi:10.1093/bioinformatics/btt086.

Hassanin A, Delsuc F, Ropiquet A, Hammer C, Jansen Van Vuuren B, Matthee C, Ruiz-Garcia M, Catzeflis F, Areskoug V, Nguyen TT, et al. 2012. Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. Comptes Rendus - Biol. 335(1):32–50. doi:10.1016/j.crvi.2011.11.002.

Heude PM. 1888. Note sur le petit buffle sauvage de l'ile de Mindoro (Philippines). Mémoires Concern l'histoire Nat l'Empire chinois. 2(4):50.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature. 431(7011):931–945.

IUCN. 2022. *Bubalus depressicornis.* IUCN Red List Threat Species. [accessed 2022 Feb 15]. The IUCN Red List of Threatened Species.

Johnson JM, Edwards S, Shoemaker D, Schadt EE. 2005. Dark matter in the genome: Evidence of widespread transcription detected by microarray tiling experiments. Trends Genet. 21(2):93–102. doi:10.1016/j.tig.2004.12.009.

Kerr R. 1792. Arnee Bos arnee. In: Strahan A, Cadell T, editors. The Animal Kingdom or zoological system of the celebrated Sir Charles Linnaeus. Class I. Mammalia. Edinburgh & London. p. 336.

382 Kriventseva E V., Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM. 2019.
383 OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and
384 viral genomes for evolutionary and functional annotations of orthologs. Nucleic
385 Acids Res. 47(D1):D807–D811. doi:10.1093/nar/gky1053.
386 Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009.
387 Circos: An information aesthetic for comparative genomics. Genome Res.
388 19(9):1639–1645. doi:10.1101/gr.092759.109.
389 Li X, Yang J, Shen M, Xie XL, Liu GJ, Xu YX, Lv FH, Yang H, Yang YL, Liu C Bin, et al. 2020.
390 Whole-genome resequencing of wild and domestic sheep identifies genes associated
391 with morphological and agronomic traits. Nat Commun. 11(1):1–16.
392 doi:10.1038/s41467-020-16485-1.
393 Linnaeus. 1758. *Bubalus bubalis.* GBIF Secr. [accessed 2022 Mar 14].
394 https://www.gbif.org/species/7422937.
395 Low WY, Tearle R, Bickhart DM, Rosen BD, Kingan SB, Swale T, Thibaud-Nissen F, Murphy
396 TD, Young R, Lefevre L, et al. 2019. Chromosome-level assembly of the water buffalo
397 genome surpasses human and goat genomes in sequence contiguity. Nat Commun.
398 10(1):1–11. doi:10.1038/s41467-018-08260-0. http://dx.doi.org/10.1038/s41467-
399 018-08260-0.
400 Luo X, Zhou Y, Zhang B, Zhang Y, Wang X, Feng T, Li Z, Cui K, Wang Z, Luo C, et al. 2020.
401 Understanding divergent domestication traits from the whole-genome sequencing of
402 swamp- and river-buffalo populations. Natl Sci Rev. 7(3):686–701.
403 doi:10.1093/nsr/nwaa024.
404 Majoros WH, Pertea M, Salzberg SL. 2004. TigrScan and GlimmerHMM: Two open source ab
405 initio eukaryotic gene-finders. Bioinformatics. 20(16):2878–2879.
406 doi:10.1093/bioinformatics/bth315.
407 Manchanda N, Portwood JL, Woodhouse MR, Seetharam AS, Lawrence-Dill CJ, Andorf CM,
408 Hufford MB. 2020. GenomeQC: A quality assessment tool for genome assemblies
409 and gene structure annotations. BMC Genomics. 21(1):1–9. doi:10.1186/s12864-
410 020-6568-2.
411 Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: Novel and
412 Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for
413 Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol. 38(10):4647–

4654. doi:10.1093/molbev/msab199. https://doi.org/10.1093/molbev/msab199.

Marçais G, Yorke JA, Zimin A. 2015. QuorUM: An error corrector for Illumina reads. PLoS One. 10(6):1–13. doi:10.1371/journal.pone.0130821.

Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 34(13):i142–i150. doi:10.1093/bioinformatics/bty266.

Mintoo AA, Zhang H, Chen C, Moniruzzaman M, Deng T, Anam M, Emdadul Huque QM, Guang X, Wang P, Zhong Z, et al. 2019. Draft genome of the river water buffalo. Ecol Evol. 9(6):3378–3388. doi:10.1002/ece3.4965.

Nguyen TT, Aniskin VM, Gerbault-Seureau M, Planton H, Renard JP, Nguyen BX, Hassanin A, Volobouev VT. 2008. Phylogenetic position of the saola (Pseudoryx nghetinhensis) inferred from cytogenetic analysis of eleven species of Bovidae. Cytogenet Genome Res. 122(1):41–54. doi:10.1159/000151315.

Ouwens PA. 1910. Contribution a la connaissance des mammiféres de Célébeès. Bull Dépt Agric indes Néerl. 38(Zool., 6):1–7.

Priyono DS, Solihin DD, Farajallah A, Purwantara B. 2020. The first complete mitochondrial genome sequence of the endangered mountain anoa (Bubalus quarlesi) (Artiodactyla: Bovidae) and phylogenetic analysis. J Asia-Pacific Biodivers. 13(2):123–133. doi:10.1016/j.japb.2020.01.006. https://doi.org/10.1016/j.japb.2020.01.006.

Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, Rowan TN, Low WY, Zimin A, Couldrey C, et al. 2020. De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience. 9(3):1–9. doi:10.1093/gigascience/giaa021.

Schreiber A, Seibold I, Nötzold G, Wink M. 1999. Cytochrome b gene haplotypes characterize chromosomal lineages of anoa, the Sulawesi dwarf buffalo (Bovidae: *Bubalus* sp.). J Hered. 90(1):165–176. doi:10.1093/jhered/90.1.165.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31(19):3210–3212. doi:10.1093/bioinformatics/btv351. https://doi.org/10.1093/bioinformatics/btv351.

Smith CH. 1827. The seventh order of the Mammalia. The Ruminantia. In: Griffith E, Smith CH, Pidgeon E, editors. The animal kingdom arranged in conformity with its organization, by the Baron Cuvier, member of the Institute of France, with additional

descriptions of all the species hitherto named, and of many not before noticed. Whittaker G.B., London. p. 293.

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinforma.(SUPPL. 25):1–14. doi:10.1002/0471250953.bi0410s25.

Weissensteiner MH, Suh A. 2019. Repetitive DNA: The Dark Matter of Avian Genomics. In: Kraus, R. (eds) Avian genomics in ecology and evolution. Springer, Cham. https://doi.org/10.1007/978-3-030-16477-5_5.

Zhang S, Liu W, Liu X, Du X, Zhang K, Zhang Y, Song Y, Zi Y, Qiu Q, Lenstra JA, et al. 2021. Structural Variants Selected during Yak Domestication Inferred from Long-Read Whole-Genome Sequencing. Mol Biol Evol. 38(9):3676–3680. doi:10.1093/molbev/msab134.

Zimin A V., Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS, et al. 2009. A whole-genome assembly of the domestic cow, *Bos taurus*. Genome Biol. 10(4). doi:10.1186/gb-2009-10-4-r42.

Zimin A V., Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. Bioinformatics. 29(21):2669–2677. doi:10.1093/bioinformatics/btt476.

Zimin A V., Puiu D, Luo MC, Zhu T, Koren S, Marçais G, Yorke JA, Dvořák J, Salzberg SL. 2017. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res. 27(5):787–792. doi:10.1101/gr.213405.116.

Tables and figures:

Table 1: Information regarding genome assemblies available for buffalo species.

Species / Assembly name	Breed	Geographic location	ID	Assembly accession no	Sequencing technology	Assembly method	Coverage	Assembly level
Bubalus bubalis NDDB_SH_1_ (**RefSeq**)	Murrah	India	NDDB_SH_1	GCF_019923935.1	PacBio Sequel; 10X and BioNano Optical Map	Falcon+Scaff10X+BioNano v. 2019-02-25	572x	Chromosome
Bubalus bubalis Jaffrabadi_v3.0	Jaffrabadi	India	AAUIN_1	GCA_000180995.3	454; Illumina NextSeq 500	MaSuRCA v. 2.3.2b	100x	Scaffold
Bubalus bubalis UOA_WB_1	Mediterranean	Italy	UOA_WB_1	GCA_003121395.1	PacBio	Falcon-Unzip v. 1.8.7	69x	Chromosome
Bubalus bubalis Bubbub1.0	Bangladesh	Bangladesh	Bubbub1.0	GCA_004794615.1	Illumina HiSeq 2000	Soapdenovo v. 2.04	119x	Scaffold
Bubalus bubalis ASM299383v1	Egyptian	Egypt	EGYBUF_1.0	GCA_002993835.1	SOLiD	Velvet v. 1.1; Bowtie2 v. 2.1.0; SHRiMP v. 2.2.3	70x	Scaffold
Bubalus bubalis UMD_CASPUR_WB_2.0	Mediterranean	USA	UMD_CASPUR_WB_2.0	GCA_000471725.1	Illumina GAIIx; Illumina HiSeq; 454	MaSuRCA v. 1.8.3	70x	Scaffold
*Bubalus depressicornis** MNHNYannick_LA_1	-	Indonesia	MNHNYannick_LA_1	Assembled MaSuRCA	Illumina NextSeq 500	MaSuRCA v. 3.3.1	44x	Scaffold
Bubalus kerabau CUSA_SWP	Fuzhong	China	CUSA_SWP	GWHAAJZ00000000	PacBio 57.8	Wtdbg 1.2.8	65x	Chromosome
Syncerus caffer ASM640878v2	African Buffalo	South Africa	ABF221	GCA_006408785.2	Illumina HiSeq	Platanus v. 1.2.4	162x	Scaffold

*= this study

Table 2: Draft assembly statistics of the lowland anoa genome

Contig statistics	value
Total length	2,565,510,706
Number of contigs	103,135
Largest contig	337,395
GC (%)	41.74
N50	38,737
L50	19,832

Table 3: Comparison of assembly quality metrics of the lowland anoa (*Bubalus depressicornis*) and other buffalo assemblies.

Name/assembly name (NCBI)	ID	Genome fraction %	Total aligned length	Largest alignment	Scaffolds count	N50	L50	GC%
Bubalus bubalis NDDB_SH1 (**RefSeq**)	NDDB_SH_1	-	-	-	26	116,997,125	9	41.75
Bubalus bubalis Jaffrabadi_v3.0	AAUIN_1	83.189	2,299,810,356	834,863	75,621	104,127	9,942	41.78
Bubalus bubalis UOA_WB_1	UOA_WB_1	98.851	2,605,694,501	34,949,624	509	117,219,835	9	41.81
Bubalus bubalis Bubbub1.0	Bubbub1.0	86.537	2,309,804,413	9,328,338	14,905	7,025,746	116	41.6
Bubalus bubalis ASM299383v1	EGYBUF_1.0	36.01	974,053,149	2,013,276	6,313	3,666,815	234	41.92
Bubalus bubalis UMD_CASPUR_WB_2.0	UMD_CASPUR_WB_2.0	93.634	2,473,056,510	7,952,377	5,714	1,545,294	508	41.73
Bubalus depressicornis MNHNYannick_LA_1	MNHNYannick_LA_1	95.415	2,515,453,834	337,395	103,135	38,737	19,832	41.74
Bubalus kerabau CUSA_SWP	CUSA_SWP	97.086	2,557,653,758	23,566,932	1,534	117,253,548	8	41.83
Syncerus caffer ASM640878v2	ABF221	73.046	1,942,672,810	4,692,267	13,167	2,448,414	351	41.72

Table 4: QUAST-LG statistics of all buffalo assemblies with respect to the river buffalo NDDB_SH_1 reference.

	B. depressicornis MNHNYannick_LA_1	*B. bubalis* AAUIN_1	*B. bubalis* Bubbub1.0	*B. bubalis* EGYBUF_1.0	*B. bubalis* UMD_CASPUR_WB_2.0	*B. bubalis* UOA_WB_1	*B. kerabau* CUSA_SWP	*S. caffer* ABF221
Misassemblies	4,949	19,238	3,561	131	4,040	1,724	2,111	6,565
Relocations	1,447	13,540	2,761	85	1,434	1,051	1,199	3,397
Translocations	3,203	4,714	757	10	2,569	647	896	3,032
Inversions	299	984	43	36	37	26	16	136
Misassembled contigs	4,550	15,988	1,049	45	1,943	255	533	1,727
Misassembled contigs length	159,179,266	1,334,096,556	2,506,642,146	55,459,162	1,891,377,139	2,639,940,877	2,594,120,526	2,486,555,687
Local misassemblies	7,014	73,267	241,261	6,933	7,100	4,870	9,940	435,454
Possible TEs	164	874	886	10	544	136	158	654
Unaligned mis. contigs	287	2,378	548	2,522	63	104	381	1,324
Unaligned contigs	886 + 8,085 partial	2,555 + 57,865 partial	297 + 7,280 partial	2,806 + 3,472 partial	182 + 3,290 partial	1 + 416 partial	140 + 1110 partial	900 + 7,314 partial
Unaligned length	45,224,171	596,227,806	299,544,303	1,673,093,194	82,826,374	49,291,638	51,316,520	779,611,955
Genome fraction (%)	95.415	83.189	86.537	36.01	93.634	98.851	97.086	73.046
Duplication ratio	1.007	1.425	1.076	1.36	1.034	1.005	1.013	1.045
Mismatches	16,233,421	19,654,061	23,375,163	17,890,296	10,863,130	10,118,782	15,844,866	114,608,168
Indels	1,578,224	746,243	705,955	6,440,610	1,136,878	1,400,310	1,534,735	2,128,964
Indels length	12,654,316	56,163,406	24,209,936	35,356,432	24,745,254	23,411,739	33,123,824	18,236,722
Mismatches per 100 kbp	649	901	1,030	1,895	442	390	622	5,983
Indels per 100 kbp	63	34	31	682	46	54	60	111
indels (<= 5 bp)	1,297,998	598,354	515,830	5,758,980	893,802	1,227,309	1,269,689	1,641,754
indels (> 5 bp)	280,226	147,889	190,125	681,630	243,076	173,001	265,046	487,210
N's	493,027	850,098,824	138,209,713	328,128,682	73,946,361	373,500	22,116,406	59,283,755
N's per 100 kbp	19.22	22,942	5,040.03	11,097	2,820.18	14.06	840.50	2,131.26

499 **Table 5**: Gene features (CDS and mRNA) predicted with GlimmerHMM

Name/assembly name (NCBI)	ID	predicted gene features (unique)	predicted gene features (>= 0 bp)	predicted gene features (>= 300 bp)	predicted gene features (>= 1500 bp)	predicted gene features (>= 3,000 bp)
Bubalus bubalis Jaffrabadi_v3.0	AAUIN_1	1,065,654	1,087,174 + 1,214 part	719,235 + 911 part	129,801 + 19 part	24,579 + 7 part
Bubalus bubalis UOA_WB_1	UOA_WB_1	1,055,791	1,059,972 + 21 part	762,464 + 17 part	154,594 + 0 part	29,659 + 0 part
Bubalus bubalis Bubbub1.0	Bubbub1.0	948,732	958,663 + 101 part	655,839 + 73 part	136,045 + 4 part	27,867 + 1 part
Bubalus bubalis ASM299383v1	EGYBUF_1.0	826,048	826,155 + 69 part	530,835 + 37 part	96,365 + 0 part	16,243 + 0 part
Bubalus bubalis UMD_CASPUR_WB_2.0	UMD_CASPUR_WB_2.0	963,177	964,473 + 138 part	669,508 + 117 part	134,780 + 5 part	26,448 + 2 part
Bubalus depressicornis MNHNYannick_LA_1	MNHNYannick_LA_1	1,027,469	1,023,163 + 5,278 part	702,282 + 4,582 part	131,966 + 204 part	24,994 + 37 part
Bubalus kerabau CUSA_SWP	CUSA_SWP	1,042,862	1,046,662 + 87 part	752,170 +70 part	151,809 + 10 part	29,488 + 6 part
Syncerus caffer ASM640878v2	ABF221	1,061,091	1,064,542 + 229 part	750,719 + 171 part	150,033 + 10 part	29,460 + 1 part

500

501 **Table 6**: Genes predicted with homology-based prediction method.

Name/assembly name (NCBI)	ID	Genes	Partial genes	Total	% Of reference's annotated genes (n= 33,348)
Bubalus bubalis Jaffrabadi_v3.0	AAUIN_1	10,804	20,895	31,699	95.05
Bubalus bubalis UOA_WB_1	UOA_WB_1	30,810	1,955	32,765	98.25
Bubalus bubalis Bubbub1.0	Bubbub1.0	11,039	20,983	32,022	96.02
Bubalus bubalis ASM299383v1	EGYBUF_1.0	1,345	23,770	25,115	75.31
Bubalus bubalis UMD_CASPUR_WB_2.0	UMD_CASPUR_WB_2.0	18,656	13,271	31,927	95.74
Bubalus depressicornis MNHNYannick_LA_1	MNHNYannick_LA_1	19,148	13,245	32,393	97.14
Bubalus kerabau CUSA_SWP	CUSA_SWP	28,349	3,419	31,768	95.26
Syncerus caffer ASM640878v2	ABF221	8,763	21,575	30,338	90.97

502

503

504

505

506

507

508

509 **Table 7**: Repeat sequence composition of the lowland anoa genome.

Family	Copy number of elements	Length occupied (bp)	% Genome
SINEs	296,064	26,945,915	1.03%
LINEs	2,864,468	786,815,034	30.04%
LINE1	1,203,360	282,366,346	10.78%
LINE2	101,415	13,911,301	0.53%
RTE/Bov-B	1,461,651	481,114,012	18.37%
LTR elements	362,123	81,208,077	3.10%
DNA transposon	255,003	38,433,935	1.47%
Small RNA	139,586	14,174,190	0.54%
Satellites	269	52,169	0.00%
Simple repeats	500,363	20,187,327	0.77%
Low complexity	81,685	3,956,146	0.15%
Unclassified	611,789	100,086,577	3.82%
Total			42.12%

510

511

512

513

514 **Figure 1**: Lowland anoa (*Bubalus depressicornis*) housed at the Ménagerie du Jardin des

515 Plantes (© Alexandre Hassanin - MNHN).

516

517

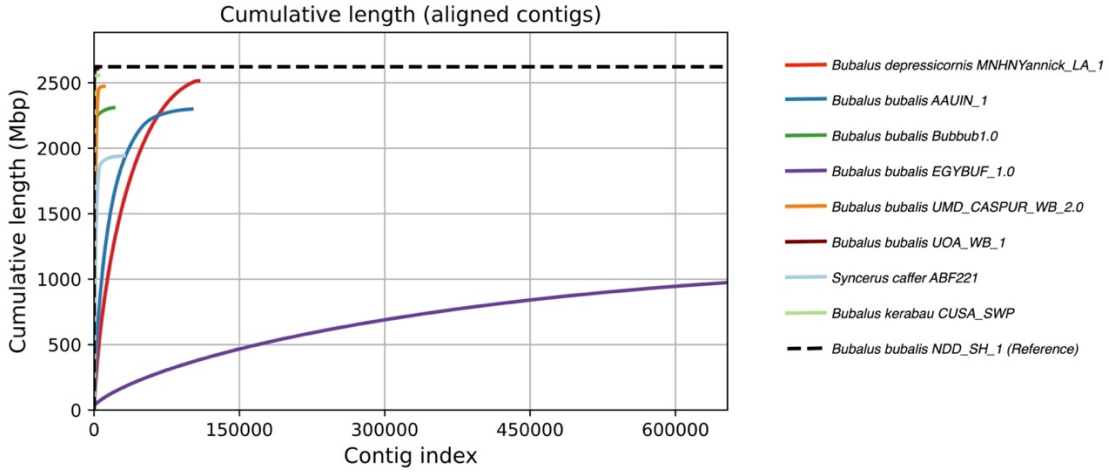
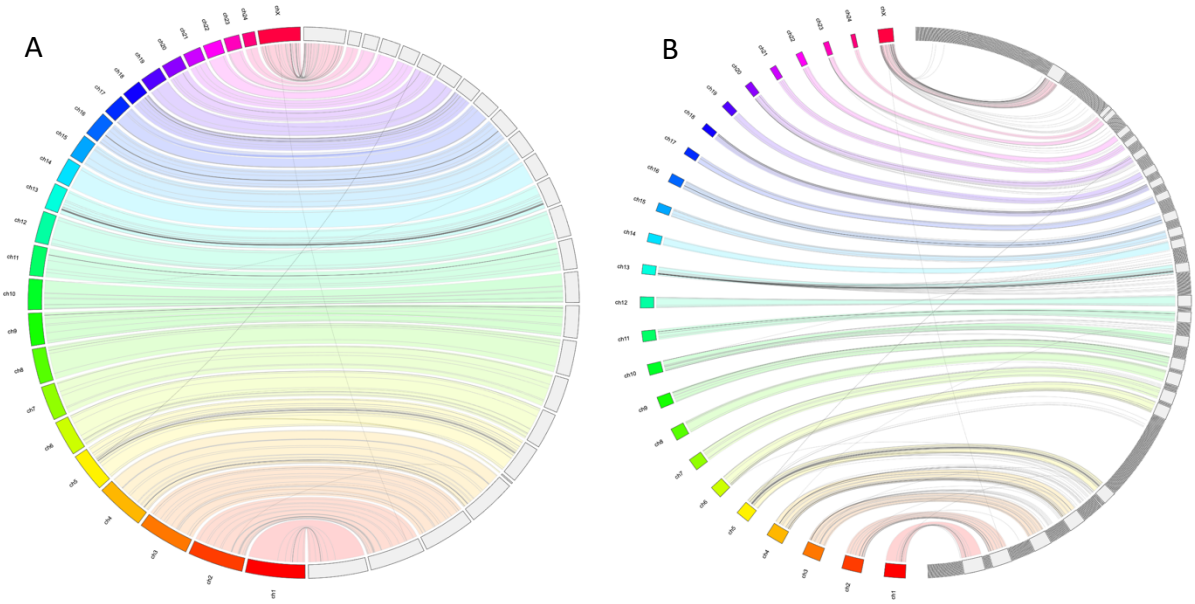


Figure 2: Cumulative length of aligned contigs of the lowland anoa (red line) against the river buffalo NDDB_SH_1 reference genome assembly (dashed line) and compared to other buffalo genome assemblies available on NCBI.

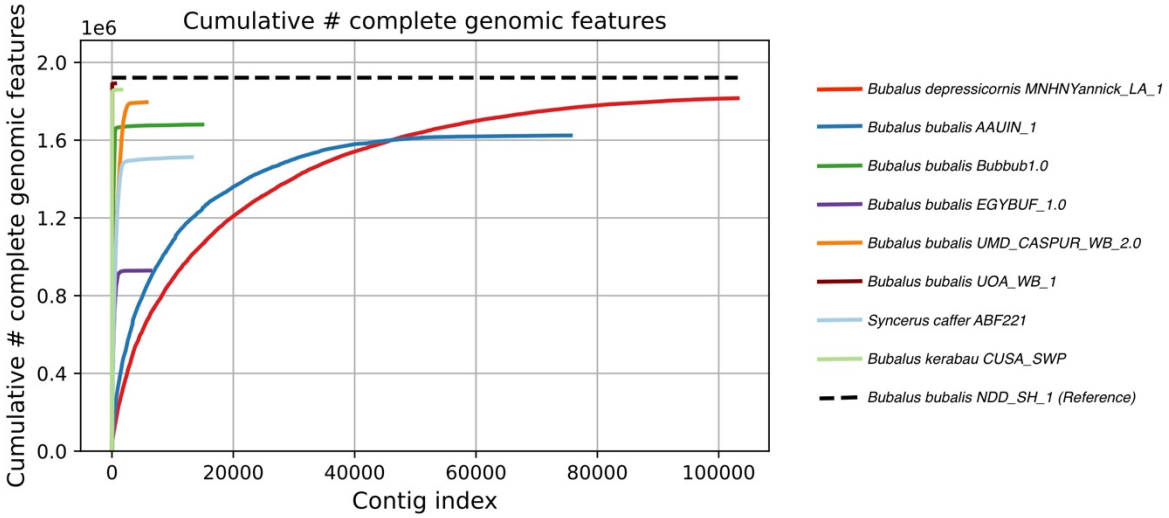
Figure 3: Circos plot of scaffolds mapped to NDD_SH_1 reference genome assembly (*Bubalus bubalis*). Outer circle represents reference sequence with GC% heatmap (0% = white, 69% = black). Inner circles represent assembly tracks, with heatmap representing correct contigs (green) and misassembled blocks (red).



527

528 **Figure 4:** Jupiter consistency plot showing alignment between the river buffalo genome
529 assemblies UO_AWB_1 and NDDB_SH_1. The left of the plots shows the numbered
530 NDDB_SH_1 chromosomes. The right of the plots shows (A) the 26 longest contigs of the
531 UOA_WB_1 assembly needed to cover 100% of the reference genome, and (B) all the 509
532 contigs of the UO_AWB_1 assembly. Coloured bands represent synteny between the
533 genomes. Lines represent genomic rearrangements, break points in the scaffolds or
534 assembly errors. Absence of lines connecting the UO_AWB_1 blocks to the NDDB_SH_1
535 chromosomes indicates contigs that could not be aligned to the reference.

536



537

538 **Figure 5**: Complete genomic features identified in the lowland anoa assembly and compared
539 to other assemblies using the river buffalo (*Bubalus bubalis*) NDD_SH1 reference sequence
540 and annotations.

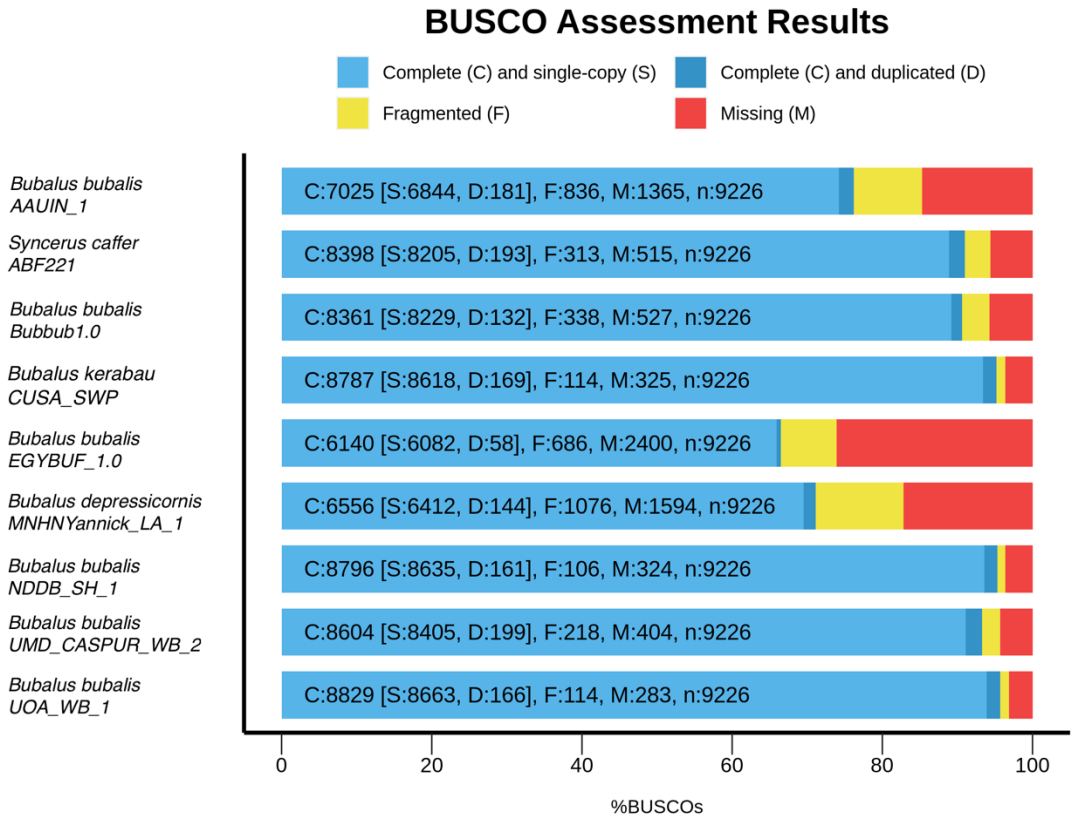


Figure 6: BUSCO results of the genome assembly of the lowland anoa (*Bubalus depressicornis*) compared to other publicly available buffalo genome assemblies.

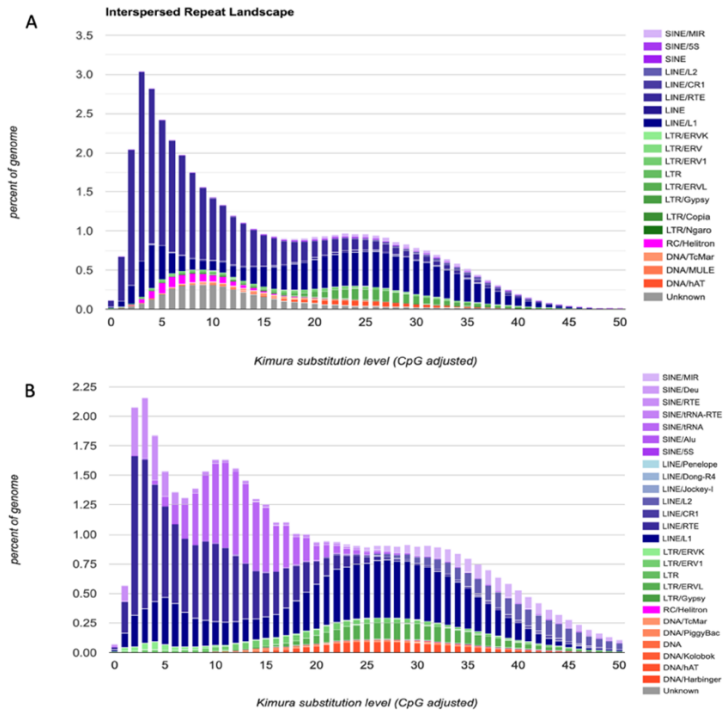


Figure 7: Interspersed Repeat Landscape of (**A**) the lowland anoa genome assembled in this study and (**B**) *Bos taurus*.