Genomics 113 (2021) 3599–3609

Contents lists available at ScienceDirect

Genomics

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

Adaptive selection signatures in river buffalo with emphasis on immune and major histocompatibility complex genes

Yan Ren [a], Callum MacPhillamy [a], Thu-Hien To [b], Timothy P.L. Smith [c], John L. Williams [a,d], Wai Yee Low [a,*]

[a] The Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, SA 5371, Australia
[b] Norwegian University of Life Sciences: NMBU, Universitetstunet 3, 1430 Ås, Norway
[c] USDA-ARS U.S. Meat Animal Research Center, Clay Center, NE, USA
[d] Dipartimento di Scienze Animali, della Nutrizione e degli Alimenti, Università Cattolica del Sacro Cuore, Piacenza, Italy

ARTICLE INFO

Keywords:
River buffalo immune genes
Positive selection in river buffalo
River buffalo gene families
Livestock major histocompatibility complex
Babesia
L1 LINE retrotransposons

ABSTRACT

River buffalo is an agriculturally important species with many traits, such as disease tolerance, which promote its use worldwide. Highly contiguous genome assemblies of the river buffalo, goat, pig, human and two cattle subspecies were aligned to study gene gains and losses and signs of positive selection. The gene families that have changed significantly in river buffalo since divergence from cattle play important roles in protein degradation, the olfactory receptor system, detoxification and the immune system. We used the branch site model in PAML to analyse single-copy orthologs to identify positively selected genes that may be involved in skin differentiation, mammary development and bone formation in the river buffalo branch. The high contiguity of the genomes enabled evaluation of differences among species in the major histocompatibility complex. We identified a Babesia-like L1 LINE insertion in the *DRB1-like* gene in the river buffalo and discuss the implication of this finding.

1. Introduction

The water buffalo (*B. bubalis*) is a domesticated species that is highly valued as a draft animal and for its meat, milk and hide. There are an estimated 207 million buffaloes in the world, the majority of which are found in Asia where there are 201 million head [1]. There are two types of water buffalo, river- and swamp- buffalo, which differ in chromosome number, phenotypes and geographic distribution. Water buffaloes have adapted to survive in hot and humid climates by spending their time in tall grasses, hiding in the shade of trees, or wallowing in mud, rivers or streams [2].

The river buffalo has many desirable traits, including tolerance to many diseases [3] although the mechanisms of this tolerance are not well understood. The availability of a high-quality river buffalo genome assembly and annotation [4] provides an opportunity for molecular evolutionary analyses of this species in the context of other mammalian species to uncover genes that may be responsible for adaption traits. Differences in members of gene families among species are common, some of which arose from the adaptation of a species to its environments [5]. The river buffalo appears to have acquired new gene functions through gene duplications [6]. Following a duplication event, the gene sequence is more amenable to change, which may lead to neo-functionalization, subfunctionalization or pseudogenization of the gene. Another way a new gene function can emerge is through positive selection of non-synonymous mutations in existing genes. A variety of methods are now available to investigate gene gains and losses [7] and detect beneficial non-synonymous mutations or positively selected sites [8]. The accuracy of these analyses depends on the quality of the genome assemblies available. Ideally these should be highly contiguous to enable accurate evaluation of conservation of synteny, which is necessary to identify and interpret origins of genes that belong to gene families which tend to occur in tandem on particular chromosomes. Additionally, the probability of error in each base of the genome should be low so that accurate gene models can be developed to detect mutations that alter the function of protein-coding genes.

The major histocompatibility complex (MHC) plays a vital role in initiating immune responses against both intra- and extracellular pathogens [9,10]. The MHC is a gene-rich locus present in all vertebrates

* Corresponding author.
E-mail address: wai.low@adelaide.edu.au (W.Y. Low).

https://doi.org/10.1016/j.ygeno.2021.08.021
Received 25 June 2021; Received in revised form 11 August 2021; Accepted 23 August 2021
Available online 26 August 2021

[11] and contains genes that encode molecules playing crucial roles in antigen presentation, inflammation regulation, and the innate and adaptive immune responses [12,13]. The mammalian MHC locus is the most gene-dense region of the genome; in humans, there are estimated to be around 250 genes (including pseudogenes) with approximately one gene per 16 kb [14]. The MHC genes are grouped into those coding for classical MHC class I and MHC class II molecules, along with the loosely defined MHC class III genes, many of which have immune related function [15,16]. The classical MHC genes are characterized by an extraordinary degree of polymorphism [15]. MHC class I molecules are expressed on the surface of all nucleated cells, and they allow the immune system to recognise self from non-self [9]. If a cell is infected with a pathogen, peptides are presented via the MHC class I antigen on the cell surface to CD8+ cytotoxic T cells leading to a swift cellular immune response to control the infection. MHC class II molecules are mainly found on antigen-presenting cells (APCs) of the immune system, such as dendritic cells (DCs), macrophages and B cells [17–19]. These cells phagocytose exogenous pathogens and proteins that are degraded and the resulting peptides are presented via the MHC Class II molecules to CD4+ helper T cells that stimulate B cells to produce antigen specific antibodies.

The river buffalo MHC region is located on chromosome 2 and has similar organization to the homologous cattle MHC region [20,21]. In cattle, the MHC Class II *DRB3* gene is highly polymorphic [22] and is known to be associated with diseases such as dermatophilosis [23], mastitis [24,25] and persistent lymphocytosis caused by bovine leukemia virus [26]. The river buffalo *DRB3* is also known to be highly polymorphic [27–29] and it is likely to be involved in disease resistance [30].

Here we use comparative genome alignment of mammalian species to study genome wide gene gains and losses and identify signs of positive selection in the river buffalo. Our focus was on immune related genes and in particular, an in-depth investigation of the molecular evolutionary pattern of MHC class I and II regions. We used the highly contiguous and accurate genome assemblies of the river buffalo, goat, pig, human and two cattle subspecies, which were annotated using the same genome annotation pipeline by the NCBI. We identified a Babesia retrotransposon in the intron of a *DRB1-like* MHC class II gene in the river buffalo and characterized genome wide distribution of this retroelement in the species studied.

2. Materials and methods

2.1. Genomes and annotations

Highly contiguous genome assemblies annotated using a standard pipeline were crucial for the comparative genomics and conservation of synteny analyses performed. The genomes selected for the analysis have been annotated by NCBI and have a contig N50 of more than 20 Mb, and are: river buffalo (*B. bubalis*; GenBank accession no GCF_003121395.1) [4], goat (*Capra hircus*; GenBank accession no GCF_001704415.1) [31], pig (*Sus scrofa*; GenBank accession no GCF_000003025.6) [32], indicine cattle (*Bos indicus*; GenBank accession no GCF_003369695.1) [33], taurine cattle (*Bos taurus*; GenBank accession no GCF_002263795.1) [34] and human (*Homo sapiens*; GenBank accession no GCF_000001405.39) [35]. The complete list of coding sequences, protein sequences and annotation files of these six genomes are given in Table S1.

2.2. Identification of gene families

The salmonid_synteny pipeline [36] (https://gitlab.com/sandve-lab/salmonid_synteny) (commit: 2f9ef3af5293ba3095e9d8147162a02730e1e393) that linked together OrthoFinder v2.4.0 [37] for orthogroups detection and building species trees, MACSE v2.03 [38] for aligning coding sequences, and treebest v1.9.2 (https://github.com/Ensembl/treebest) for phylogenetic gene tree construction was used to find orthogroups, build species trees and study gene synteny. The pipeline only saves the longest coding sequence and protein isoform of each gene prior to any sequence alignment. The configuration file for the synteny pipeline to reproduce the work is given in https://github.com/DaviesCentreInformatics/waterbuffaloComparativeGenomics (commit 9a590ca118245c27e7c18ddcdcae9cc330d96a4f). To categorize orthogroups into gene families, the Entrez IDs of human, taurine cattle, goat or pig were used to find corresponding UniProtKB IDs. Then, these UniProtKB IDs were used to search against the PANTHER database v15.0 [39] to determine the gene families. The orthogroups that shared the same PANTHER family ID were assigned to the same gene family. In cases where an orthogroup contained sequences matching multiple PANTHER family IDs, we categorized the gene family based on the PANTHER family IDs with the highest proportion in the orthogroup.

2.3. Gene family expansion and contraction

The count of genes in gene families was obtained by combining the count of orthogroup(s) that belong to the gene family. Only gene families found in at least two species were retained. To identify expansion or contraction of gene families we used Computational Analysis of gene Family Evolution (CAFE) v4.2.1 [7], which uses a probabilistic model to infer the rate and direction of the changes in gene size over a given ultrametric tree. The rooted amino acid tree from OrthoFinder was converted into an ultrametric tree using r8s v1.70 [40]. The calibration point of 96 million years between human and pig obtained from the TimeTree [41] was used as the reference calibration point to build the ultrametric tree.

The CAFE, multiple birth-death parameter (λ) and global λ were tested to find the best λ estimation to calculate the gene evolution ratio. We tested a multiple λ model, in which all ruminants including river buffalo, goat, indicine and taurine cattle had one $\lambda_{[ruminants]}$. The other species (pig, human) shared a second $\lambda_{[common]}$. By setting '-t 1000', the function lhtest was run 1000 times to produce the distribution of likelihood ratio 2 $[\log L_{global} - \log L_{multiple}]$. The global λ with the error correction model was better than the multiple λ model, as per the Akaike Information Criterion (AIC) [42]. The visualization of CAFE output and downstream analysis were carried out using custom R scripts (https://github.com/DaviesCentreInformatics/waterbuffaloComparativeGenomics).

The species tree used to map gene gains and losses was based on the OrthoFinder output that used the Species Tree Inference from All Genes method (STAG) [43].

2.4. Positive selection analysis

The nucleotide sequences of single copy orthologues (SCOs) identified from the OrthoFinder pipeline were translated using the Transeq function from EMBOSS v6.5.7 [44]. The '–auto' mode of MAFFT v7.305 [45] was used to align the amino acid sequences. The codons were mapped to the aligned amino acid sequences using PAL2NAL v13 [46]. The FASTA alignment was converted into PHYLIP format using ALTER v1.3.4 [47] and then used as input for maximum-likelihood tree construction using RAxML v8.2.10 [48]. Taking the codon alignment and the tree of each SCO as input, CODEML branch site model A i.e. the alternative hypothesis (model = 2, NSsites = 2, fix_omega = 0, omega = 1.5) and null hypothesis (model = 2, NSsites = 2, fix_omega = 1, omega = 1) in PAML v4.8 [8] were compared to detect positively selected sites.

The log-likelihoods values for each SCO PAML set were extracted from the alternative hypothesis model and null hypothesis model. The likelihood ratio test was calculated as $2[\log \text{likelihood}_{alternative} - \log \text{likelihood}_{null}]$, and the *p*-value of the test was evaluated based on the null distribution is the 50:50 mixture of point mass 0 and chi-squared values. To account for multiple testing, the *p*-values were adjusted by False Discovery Rate (FDR) [49] and the results that passed FDR < 0.05 were used.

2.5. Immune gene identification

To determine which gene families or orthogroups belong to the immune system, a referenced database was downloaded from InnateDB [50], which included immune genes from ImmPort [51], Immunogenetic Related Information Source (IRIS) [52], Septic Shock Group [53], MAPK/NFKB Network, and information from Calvano et al. [54]. The UniprotKB ID for gene family or orthogroup was used to find the corresponding Entrez gene ID in InnateDB, in order to determine whether it belonged to particular immune gene sets.

2.6. Biological pathways analysis

The GO and KEGG enrichment analysis for gene families in the river buffalo branch with significant gene gains and losses used cattle as the reference species. This is because the river buffalo's pathways are likely to be more similar to cattle than human. For enrichment analysis with Reactome terms, we chose human as the reference species as Reactome terms for cattle were not available. Immune related genes among gene families with significant gains and losses in the river buffalo branch were identified by matching them to InnateDB. Then, the human genes from each of these gene families were used for GO and Reactome term enrichment analysis. Human representative genes were used, rather than cattle genes, because the InnateDB immune database does not capture cattle genes. After retrieving representative genes from each gene family, we applied goana or kegga functions from the limma [55] R package to find enriched GO and KEGG terms, respectively. The Reactome pathway terms were clustered for representative human genes and their immune genes using the enrichPathway function from ReactomePA R package [56].

For positively selected genes uncovered in the CODEML analysis, representative human genes were used to perform GO and Reactome enrichment analysis using the limma R package as described before.

2.7. Identification of MHC I and II genes

MHC I and II genes were identified from a literature search with cattle and human MHC I and II genes being used as input for subsequent BLASTP searches [57–62]. BLASTP databases were built for each of the six species and two multi-FASTA files, each containing the amino acid sequences of all identified cattle and human MHC I and MHC II genes, which were used to identify possible orthologs and paralogs in each species. BLASTP results that had e-value $<$1e-5 and a percent identity of $\geq$75 were retained. This cut-off was selected in order to maximise the likelihood of capturing all possible orthologs and paralogs within a given MHC gene family, but avoiding genes from other MHC gene families. We selected the protein sequence linked to each gene ID for each species as input for phylogenetic tree building. Where isoforms were present, the one with the longest amino acid sequence was selected. We aligned the amino acids using MUSCLE v3.8.31 [63] with default parameters. The aligned amino acid sequences were then trimmed using trimAl v1.2 [64] with the '-automated1' option, to remove poorly aligned regions and improve phylogenetic reconstruction [65]. Trimmed and aligned amino acid sequences were then used as input into RAxML v8.2.10 [48] for phylogenetic analysis. RAxML was run for 1000 bootstraps with the options '-m PROTGAMMAAUTO -p 12345 -x 12345'. Trees were then visualised in iTOL v5.0 [66] to determine orthologs and paralogs. The gene order of MHC I and II genes was visualised in Artemis v18.0.0 [58] and the NCBI Genome Data Viewer v4.8.11 (GDV).

The overlaps between these manually curated MHC gene coordinates and gene families undergoing gains and losses was performed using a customized R script (https://github.com/DaviesCentreInformatics/waterbuffaloComparativeGenomics). The overlaps between MHC genes and positively selected genes was done the same way.

2.8. Babesia retrotransposon in MHC II gene and the genome

The river buffalo gene LOC102408590 corresponds to the taurine cattle gene, *BOLA-DRB2* (MHC class II DR beta-chain). The first intron of LOC102408590 contains two open reading frames, ORF1 (819 bp) and ORF2 (3819 bp). BLASTP of ORF2 showed 98% identity to *Babesia ovata* retrovirus-related Pol poly LINE-1 (NCBI Accession no: GBE63528.1) across the entire length of the ORF2 protein. Using the protein sequence of GBE63528.1 as input, TBLASTN searches were done against the six study species. The TBLASTN results were filtered for alignment length ($>$1009) and percent identity ($>$90%). The chosen alignment length filter requires 90% of the input gene to be covered i.e. searching for match for most of the entire gene. These TBLASTN hits were overlapped with gene annotation for the downloaded genome assembly for six species (NCBI). To identify repetitive sequences in LOC102408590 and the *Babesia ovata* contig (BDSA01000072) carrying the retrotransposon Pol poly LINE-1, dot plots were done using Gepard v1.4 [67].

3. Results

3.1. Genome assemblies of river buffalo

Literature surveys and genomic database searches identified eight river buffalo and one swamp buffalo genome assemblies (Table S2). The highest contiguity among the river buffalo assemblies was the UOA_WB_1 reference (contig N50 $>$ 22 Mb), which was generated from a female Mediterranean buffalo [4]. The assembly was chosen to represent river buffalo as it was annotated by the NCBI, which is consistent with other high-quality assemblies used in this study. We note that a swamp buffalo genome was recently assembled and annotated [68], but the contig N50 was below 10 Mb and it was not annotated by the NCBI pipeline, and hence was excluded from further analysis. The six species/subspecies genomes used for this analysis are listed in Table 1.

3.2. Gene family expansion and contraction

3.2.1. Significant gains/losses in the river buffalo branch

OrthoFinder identified 18,423 homologous protein orthogroups in the six genomes, of which 18,353 orthogroups could be assigned to UniProtKB IDs. There were 16,752 orthogroups that could be classified into PANTHER gene families, which were used in gene gains and losses analyses. UniProtKB IDs from each orthogroup were used to identify 7425 gene families that could be mapped by the Panther database and passed our criteria for running the CAFE software to detect gene gains and losses. Most species in the comparison have a larger number of gene family gains than losses, with the pig being the only species with more losses than gains. The indicine cattle had a similar number of gene gains as losses. In total, 209 gene family expansions and 154 gene family contractions were identified in the river buffalo since divergence from cattle (Fig. 1). Among these families, 172 had statistically significant changes in gene gains whereas 96 had significant losses in the river buffalo (branch p-values $<$ 0.05) (Table S3).

River buffalo gene families that changed by more than one gene copy were ranked by p-values and the top 10 gene families with the most significant gene gains were the following: Inner Membrane Transporter Ygji-Related, G-Protein Coupled Receptor, MHC Class II-Related, Beta-1,3-N-Acetylglucosaminyltransferase, Hras-Like Suppressor-Related, Olfactory Receptor 56b1-Related, Translation Factor, Diacylglycerol O-Acyltransferase, 60s/50s Ribosomal Protein L6/L9 and MHC Class I NK Cell Receptor (Table S4). The top 10 families with the most significant contractions were: Zinc Finger Protein, Beta-Defensin, Nap1l1 Protein, Olfactory Receptor, Zinc Finger and Scan Domain-Containing, Cytochrome P450 508a4-Related, Solute Carrier Family 22 Member, MHC Class I-Related, E3 Ubiquitin-Protein, Ligase Trim and G-Protein-Recep-F1-2 Domain (Table S5).

Pathway enrichment analysis of significantly expanded gene families

Table 1
Assembly statistics of the six study species.

Assembly name	Genome size (Gb)	Contig N50 (kb)	Scaffold N50 (Mb)	Number of contigs	Breed Or Origin	Type	Species	Reference
UOA_WB_1	2.66	22,441.5	117.2	919	Mediterranean	River	*Bubalus bubalis*	[4]
GRCh38.p13	3.10	57,879.4	67.8	998	NA	NA	*Homo sapiens*	[35]
Sscrofa11.1	2.50	48,231.2	88.2	1118	Duroc	NA	*Sus scrofa*	[32]
ARS-UCD1.2	2.72	25,896.1	103.3	2597	Hereford	NA	*Bos taurus*	[34]
UOA_Brahman_1	2.68	26,764.3	104.5	1552	Brahman	NA	*Bos indicus*	[33]
ARS1	2.92	26,244.6	87.3	30,399	San Clemente	NA	*Capra hircus*	[31]

All genomes are annotated by the NCBI (Accessed on 24/08/2020). NA denotes not available.

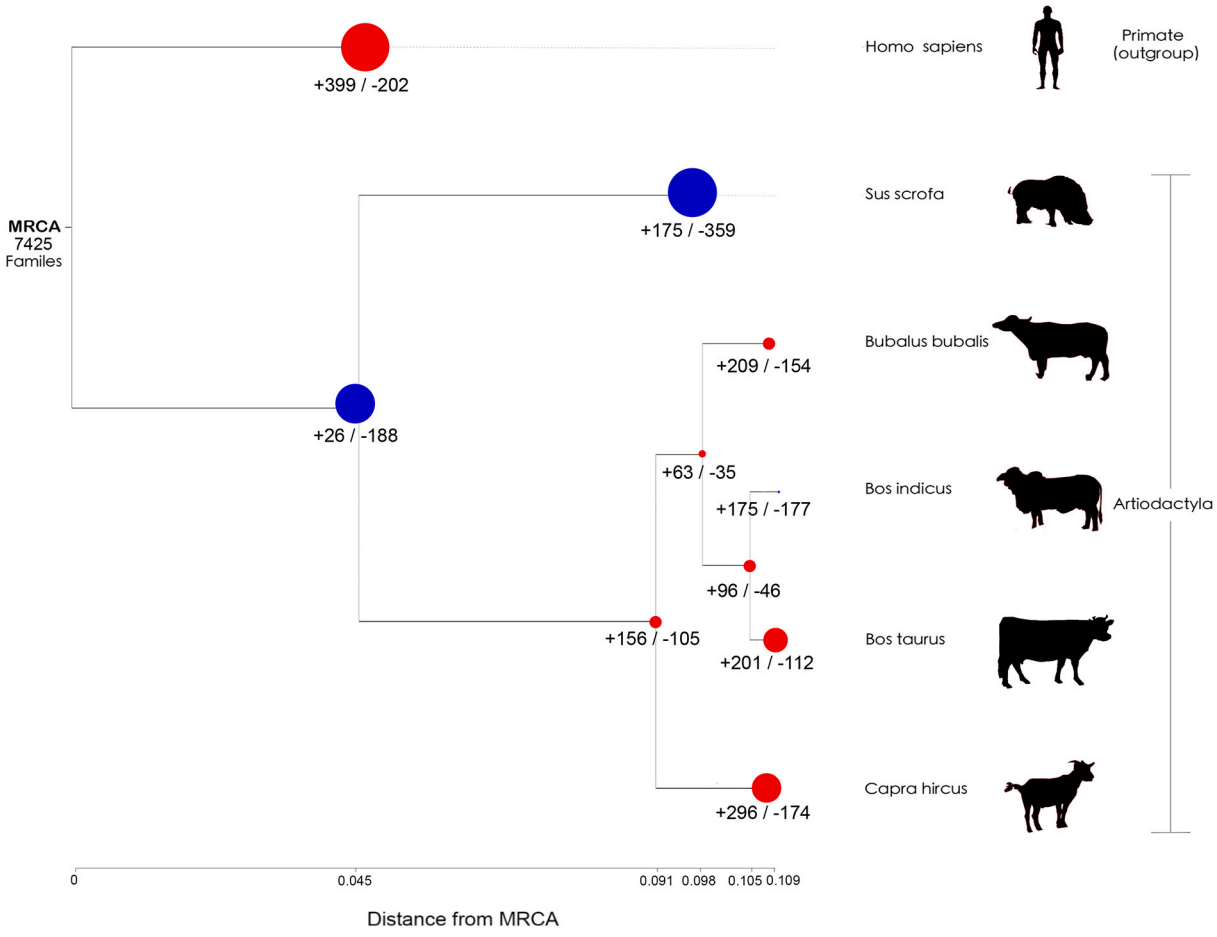

Fig. 1. Gene family expansion and contraction across the six study species.
The species tree was based on the Species Tree Inference from All Genes (STAG) method. The number of gene families involved in expansions and contractions are shown in red and blue, respectively, with the size of each node representing the significance of the expansion or contraction. Human was the outgroup used for this analysis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

on the river buffalo branch was performed using cattle as the reference species. The analysis revealed calcium ion binding as the only enriched GO term, and 33 enriched KEGG terms (FDR < 0.05) (Table S6). The top 5 ranked KEGG terms based on FDR adjusted *p*-values were Olfactory transduction, Antigen processing and presentation, Graft-versus-host disease, Asthma and Autoimmune thyroid disease. The GO ancestor chart could not be used for expanded gene families as there was only one GO term found. When significant contracted gene families on the river buffalo branch were used for pathway enrichment analysis, 15 GO terms and 19 KEGG terms were identified (FDR < 0.05) (Table S7, Table S8). Clustering of the 15 GO terms to create a GO ancestor chart [69] showed the interconnectedness of child GO terms, and identified the following biological processes as enriched: Protein K63-linked ubiquitination, Protein K48-linked ubiquitination and Protein K11-linked ubiquitination (Fig. S1). These GO terms also clearly revealed the molecular function of ubiquitin-protein transferase activity, which suggests gene losses in protein degradation pathways. The second most significant KEGG term was ubiquitin mediated proteolysis (Table S8), which agreed with the GO term result.

Reactome pathway analysis used human representative genes because cattle gene Reactome data were not available. There were 271 and 67 Reactome pathway terms identified as significant for gene family expansions and contractions, respectively (FDR < 0.05). The top 5 Reactome terms for expanded gene families were as follows:

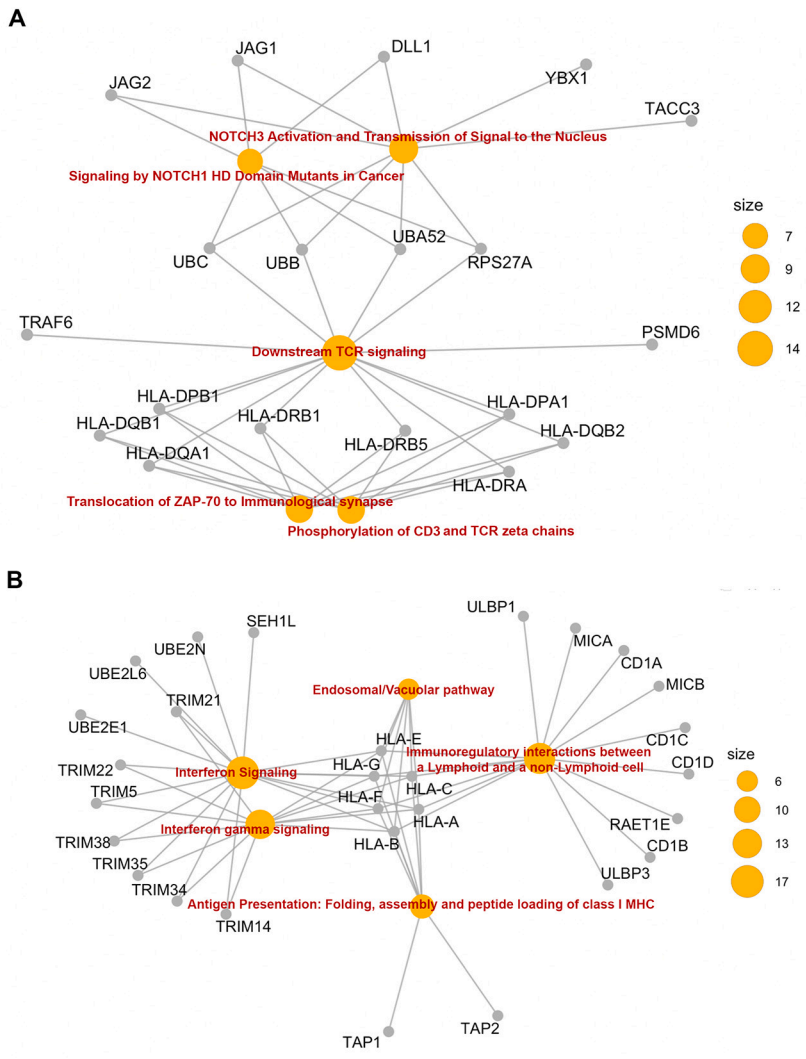

Fig. 2. The clustering of immune related genes to Reactome pathways using genes with significant (A) gene gains and (B) losses in the river buffalo.
The significantly enriched Reactome terms (FDR < 0.01) are shown as orange nodes, and the gene symbols that contributed to the term are shown in black. The size of the orange nodes is determined by the number of genes that connect to it. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Translocation of ZAP-70 to Immunological synapse, Phosphorylation of CD3 and TCR zeta chains, PD-1 signaling, DNA Double Strand Break Response and Recruitment and ATM-mediated phosphorylation of repair and signaling proteins at DNA double strand breaks (Fig. S2). The top 5 Reactome terms for contracted gene families were as follows: Organic cation/anion/zwitterion transport, Synthesis of active ubiquitin (roles of E1 and E2 enzymes), Protein ubiquitination, CYP2E1 reactions, and G beta (gamma signaling through BTK) (Fig. S3).

3.2.2. *Pathway enrichment of immune genes in the river buffalo branch*

More than two thirds (67%; 116 of 172) of the significantly expanded families are immune-related according to the InnateDB database. Immune-related genes were enriched for 565 GO terms and 271 Reactome terms (FDR < 0.05) using human representative genes for the enrichment analysis. GO ancestor chart analysis of the top 10% GO terms revealed enrichment of the GO terms Innate immune response, Cytokine-mediated signaling pathway and Regulation of apoptotic process (Fig. S4). The top 5 enriched Reactome terms were NOTCH3 Activation and Transmission of Signal to the Nucleus, Translocation of ZAP-70 to Immunological synapse, Downstream TCR signaling, Phosphorylation of CD3 and TCR zeta chains and Signaling by NOTCH1 HD Domain Mutants in Cancer (Fig. 2A).

Forty one of the 96 (43%) significantly contracted families are immune-related when matched InnateDB database. These genes were enriched for 214 GO terms and 31 Reactome terms (FDR < 0.05). The top 10% GO terms indicated the biological process term of positive regulation of natural killer cells on the GO ancestor chart (Fig. S5). The top 5 enriched Reactome terms were Endosomal/Vacuolar pathway, Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell, Interferon signaling, Interferon gamma signaling and Antigen Presentation (Folding, assembly and peptide loading of class I MHC) (Fig. 2B). All the top 5 Reactome terms, of both expanded and contracted gene families, were connected by a network of genes.

3.2.3. *MHC I and II gene gains and losses*

Significant gene losses in MHC class I and significant gene gains in MHC class II were observed among the immune gene families in river buffalo (Fig. S6). However, the grouping of orthologs into gene families for the MHC class I genes was inconsistent between the bioinformatics pipeline and manual grouping. Manual curation identified a total of 47 class I genes (Fig. 3A, Table S9) but our automated pipeline found 180 genes as a result of the Panther database grouping “MHC class I related” and classical “MHC class I" genes into the same gene family. The bioinformatics pipeline grouping and manual curation were more consistent for the MHC class II genes, supporting the general validity of the pipeline. The river buffalo had 14 MHC class II genes from manual curation and 16 were identified by the pipeline. The difference between the two methods was because the Panther database includes two copies

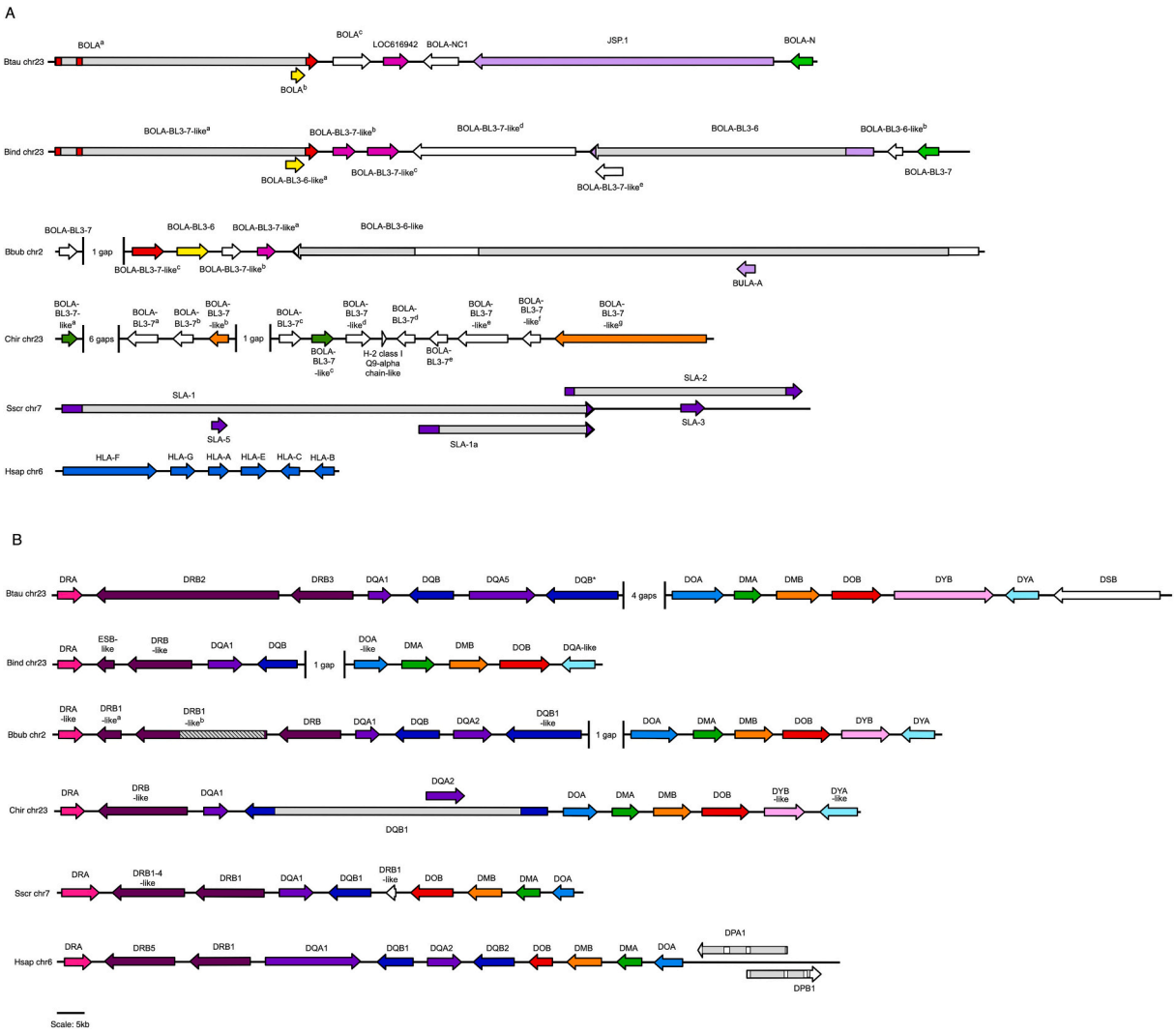

Fig. 3. An illustration of gene order in MHC class I and II genes.
Full-length genes are shown as arrowed boxes and the direction of the arrows represents the direction of transcription. The boxes are coloured to indicate orthologs and paralogs with bootstrap support more than or equal to 70. Where bootstraps were below 70, gene order was used in conjunction with bootstrap values to assign orthologs. The white boxes represent genes that could not be assigned as an ortholog based on bootstrap values and gene order. The "||" symbol represents a sequence gap in the contig with the number of gaps between regions in between. The meaning of gap here refers to genome assembly gaps, which are ambiguous Ns inserted during the scaffolding of contigs. Overlapping genes have their introns shown in grey. Gene names are shown and the corresponding NCBI protein IDs can be found in Table S9 (A) A scaled illustration of MHC class I in the six study species. There is one goat *BOLA-BL3-7-like* gene on an unplaced scaffold that is not shown. (B) A scaled illustration of MHC class II. The stripped rectangle on the *DRB1-like*^{b} gene on river buffalo (Bbub chr2) represents the *B. ovata* retrotransposon. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of *B2M* genes that encodes beta-2-microglobulin as MHC class II, which is incorrect as *B2M* is a subunit of MHC Class I antigens (Fig. 3B).. Therefore, although significant gene gains were predicted in MHC class II for the river buffalo branch from our automated pipeline, careful manual curation of these genes did not support the same conclusion.

3.3. Genes displaying positive selection

We found 602 SCO gene sets under positive selection with FDR < 0.05 (Table S10). *TRPM4* was the gene with the highest statistical significance from the FDR *p*-value ranking and encodes a protein that may play a role in signaling in skin tissues [70]. Our list of genes under selection included five growth factors (*TGFB1*, *TGFBR3*, *TBRG1*, *IGFBP4*, *VEGFD*), 14 genes involved in protein degradation (*NEURL4*, *UBE3D*, *MMP23B*, *RFFL*, *NEDD4L*, *USP3*, *OBI1*, *OTUD3*, *NEURL2*, *DTX3L*, *UBFD1*, *UBE2J1*, *USP40*, *DTX4*) and three heat shock related proteins (*HSF5*, *HSF1*, *DNAJC2*).

Pathway enrichment analysis of all positively selected genes revealed 125 enriched GO terms at FDR < 0.05. As the top 10% GO terms had too few entries in the GO ancestor chart, we used the top 50% enriched GO terms for clustering. We found the following GO terms as significant, microtubule cytoskeleton organization, glycerolipid biosynthetic process, cellular protein modification process and microtubule (Fig. S7). No significant Reactome pathway were identified for the positively selected genes.

Among the 602 SCOs that showed signs of positive selection, 205 (34%) were immune related genes. There were 290 enriched GO terms for these immune genes (FDR < 0.05). In the GO ancestor chart of the top 10% enriched terms at the biological processes level were animal organ development and cell surface receptor signaling pathway (Fig. S8). Two significant Reactome terms found were NOD1/2 Signaling Pathway and R-HSA-168643 Nucleotide-binding domain, leucine rich repeat containing receptor (NLR) signaling pathways (Fig. S9).

3.4. Major histocompatibility complex I and II gene synteny

We used alignment to cattle and human MHC I genes (Table S11) to identify 47 MHC I genes among the six genomes included in the study (Table S9) with the gene number ranging from five in pig to 13 in goat (Fig. 3A). Four orthologous sets of genes were identified between taurine cattle, indicine cattle and the river buffalo, but no clear orthologues of these genes could be established in goat, pig and human. In the river buffalo, these genes were *BoLA-BL3–7-like*^{c}, *BoLA-BL3–6*, *BoLA-BL3–7-like*^{a} and *BuLa-A*. All human MHC I genes occurred in a monophyletic group and this pattern was also observed in the pig MHC I genes. Interestingly, the pig MHC I genes displayed a noticeable degree of overlap where *SLA-1* spanned the full length of both *SLA-5* and *SLA-1a*; *SLA-2* spanned the length of *SLA-3*.

Cattle and human MHC II genes were used as reference sequences to identify 72 MHC II genes in the six genomes, which were grouped into orthologs based on chromosomal locations, gene order and phylogenetic relationship (Table S11, Fig. S10). Orthologs for the *DMB*, *DOA* and *DRA* genes as well as the DY gene family had good bootstrap scores. The remaining genes had lower bootstrap support, suggesting a high diversity of these sequences. The number of MHC class II genes was lowest for indicine cattle and pigs (10 genes) and highest for river buffalo and taurine cattle (14 genes) (Fig. 3B). Only three genes, *DRB*, *DQA*, and *DQB* were duplicated in any of the six genomes. The goat genome lacks a *DRB* gene duplication (Fig. 3B). *DQA* was duplicated in human, goat, river buffalo and taurine cattle but not in indicine cattle or pig. *DQB* was duplicated in human, taurine cattle and river buffalo, but not in indicine cattle, goat and pig. Finally, the overall gene order and number of genes in the MHC I and II regions were less similar for the MHC I genes than for the MHC II genes among the six genomes included in the study.

3.5. Babesia retrotransposon in river buffalo MHC II gene

Two large open reading frames (ORFs) in the intron of the river buffalo LOC102408590 gene were identified. This gene is designated as *DRB1-like* by the NCBI. A BLASTP of LOC102408590 to cattle refseq protein database showed 96.5% identity to the DRB5 protein (XP_024839853.1). ORF1 is 819 bp and ORF2 is 3819 bp. ORF1 showed 98.5% identity to *Babesia ovata* LINE-1 type transposase. A BLASTP analysis of the ORF1 sequence against Swissprot non-redundant database showed it has Transposase_22, spc7 and Tnp_22_dsRBD protein domains. ORF2 showed 98% identity to *Babesia ovata* retrovirus-related Pol poly LINE-1. ORF2 has the RVT_1 and Exo_endo_phos protein domains, which are known to be found in species of mammalian and non-mammalian origin [71]. Both of these ORFs are missing in the same gene in the short read-based river buffalo genome assembly (UMD_CASPUR_WB_2.0) from the same animal but their presence in intron 1 of the LOC102408590 gene is supported by whole genome shotgun short read data (Fig. S11). The retrotransposon is not found in gene GWHGAAJZ000112 of the swamp buffalo, the gene orthologous to water buffalo LOC102408590.

The *Babesia ovata* genome was recently sequenced [72] and the transposase and Pol genes were found in a single contig (NCBI accession: BDSA01000072). A dotplot of the Babesia contig and LOC102408590 showed the entire contig sequence of 8691 bp is a close match to the first intron of LOC102408590 (Fig. 4A). The first intron of the gene contains inverted repeat sequences (Fig. 4B), which were not identified in the Babesia contig. This may be a Babesia genome assembly error, where the presence of repeats triggered a break in genome assembly continuity. The predicted retrotransposon structure is given in Fig. 4C. Using the *Babesia ovata* retrovirus-related *Pol* poly LINE-1 as search input, we found that this gene is distributed widely across all chromosomes in ruminants (Fig. 5), but is not found in pig or human genomes with filter criteria set to retrieve approximately the full length of *Babesia* sequence at >90% sequence identity (Fig. S12). However, lowering thresholds to allow shorter match length and lower sequence identity did detect LINE-1 *Pol*-like sequences in these species (Fig. S11). Using the *Pol* gene as an indicator of Babesia-like retrotransposon insertion, a total of 2540 were found across the six genomes. Only about 1% of the human genome is protein-coding [73] and the other mammalian genomes studied here are likely to have a similar proportion. However, an average of 24.80% of the Babesia-like retrotransposons were found within protein-coding genes in the studied species suggesting that these elements preferentially insert into genes.

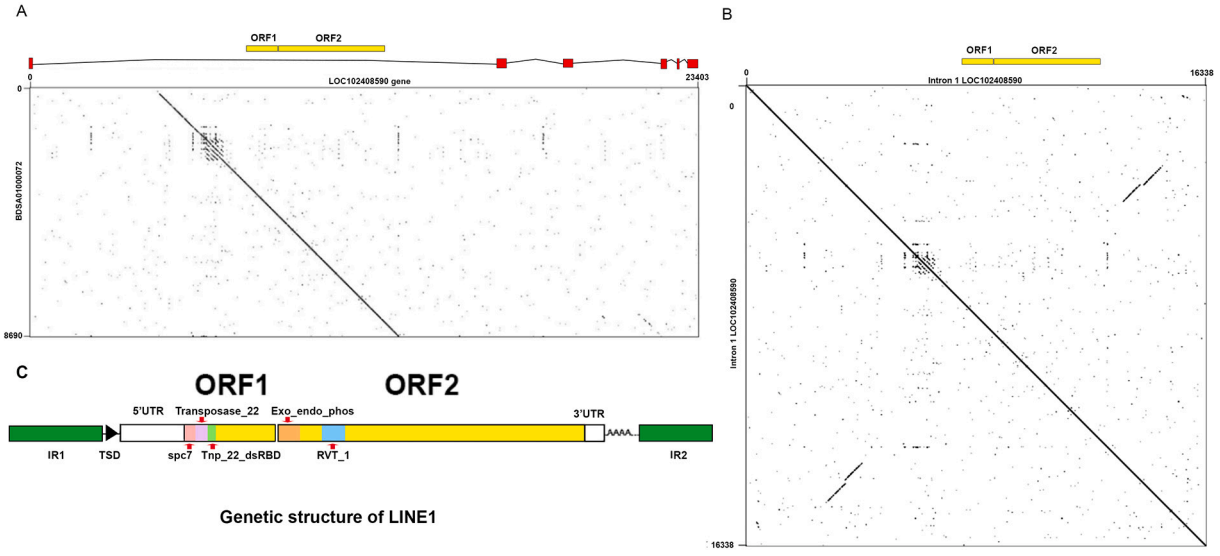

Fig. 4. The dotplots of LOC102408590 and *Babesia ovata* contig BDSA01000072.
(A) The dotplot between LOC102408590 and *B. ovata*. The predicted exon regions are shown in light blue. The open reading frame 1 and 2 (ORF1 and ORF2) are highlighted in yellow. (B) The dotplot between the first intron of LOC102408590 and itself. The dotplot shows the inverted repeat sequences in the intron 1 region of LOC102408590. (C) The predicted retrotransposon structure in the first intron of LOC102408590. The three domains in ORF1 are Transposase_22, spc7 and Tnp_22_dsRBD. The two domains in ORF2 are RVT_1 and Exo_endo_phos. IR represents inverted repeats and TSD represents tandem sequence duplication. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

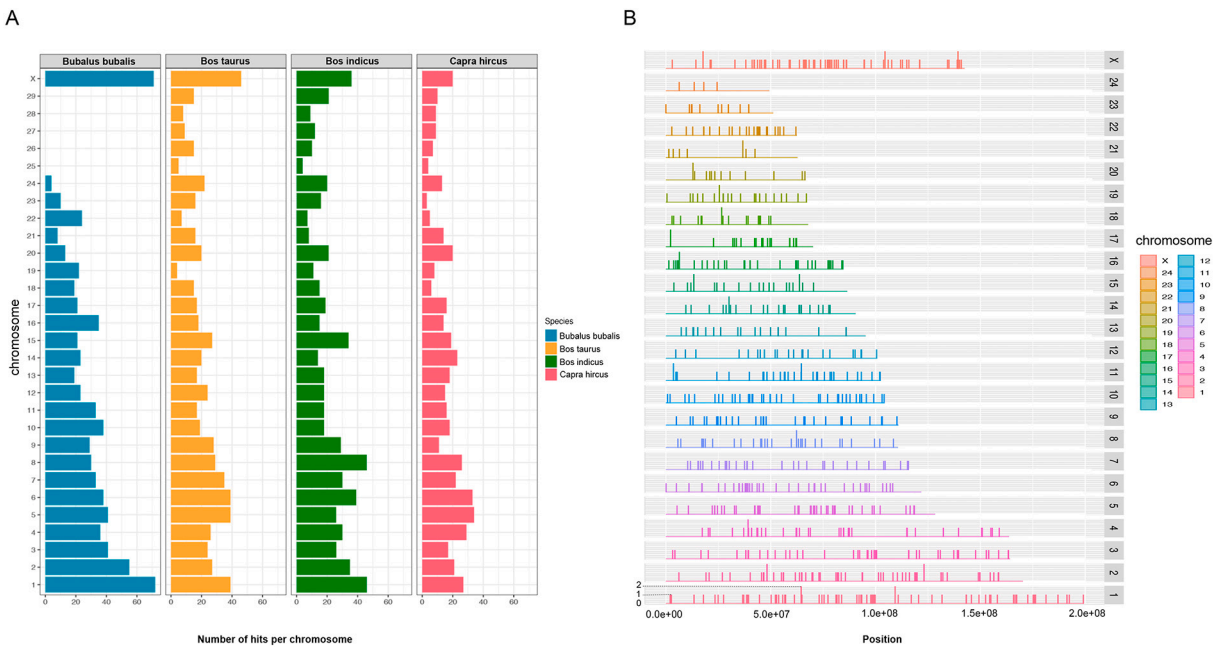

Fig. 5. The TBLASTN hits of *Babesia ovata* retrovirus-related Pol poly LINE-1 on each chromosome.
The barplot shows the number of TBLASTN hits on each chromosome for the four ruminant species. *B. bubalis* has no chromosome 25 to 29.

4. Discussion

4.1. Gene family expansion and contraction in the river buffalo

Gene duplication provides genetic material for natural selection to work on [74], whereas gene loss could be the result of diverged duplicated genes that are no longer under functional constraint. Together, the dynamics of gene gains and losses reflect the interplay of genomic changes and selection. In the context of an agricultural animal species, such as the river buffalo, gene gains and losses may be linked to artificial selection for meat yield, milk production, heat tolerance and disease resistance traits. To accurately identify genome wide gene gains and losses, high quality genome assemblies and annotations of several species are required. The river buffalo reference genome was assembled to chromosome-level and annotated with about 15 billion RNA-Seq reads from more than 50 tissues [4]. This latest river buffalo genome has led to the successful identification of loci linked to convergent signatures of domestication [75]. Here we detected 268 gene families with significant gene gains or losses in the river buffalo since divergence from cattle. By analysing these gene families individually and clustering them into significant biological pathways using GO, KEGG and Reactome terms, the four major biological functions that emerged were protein degradation, olfactory receptor sensing system, detoxification and immune system.

4.2. Protein degradation

We identified significant loss of genes encoding proteins in the protein degradation pathway of buffalo, in particular those that belong to ubiquitinating enzymes. Protein degradation enables cells to dynamically adjust their proteome to respond to internal and external signals. The balance between protein synthesis and degradation is important to ensure efficient utilization of energy. *Bos indicus* cattle may be better adapted to harsher and hotter environments by having lower rates of protein turnover [33,76]. Like *Bos indicus* cattle, river buffalo is adapted to survive in harsh environments. Finding that there are significant gains and losses of genes in biological pathways involved in protein degradation may be a sign that the river buffalo has adapted to its environment by regulating protein turnover for a lower metabolic rate.

4.3. Olfactory receptor

Adaptation to a specific ecological niche is accompanied by changes in sensory perception, which at the molecular level is seen as changes in the olfactory receptors (ORs) [77]. The OR gene families consist of G-protein coupled receptor genes that function to perceive chemosensory signals [78]. These genes have undergone extensive gene duplications and contractions in several taxa [79–81]. The dynamic gains and losses of the river buffalo OR gene family suggests that it has specific genes which are used for sensory perception in its niche environment. The closest species to the river buffalo is cattle, which has 1071 OR-related sequences in the UMD3.1 genome assembly, but only 881 are considered as functional genes [82]. About 6% of cattle functional ORs genes are cattle-specific and hence, it is expected that the river buffalo would also have specific OR genes. Indeed, one of the largest OR studies in mammals [77] has shown that many species have specific ORs, which are likely to be important for adaption.

4.4. Detoxification

Genes that play a role in detoxification are known to evolve rapidly to deal with environmental toxins [83–85]. Gene losses in the river buffalo gene families associated with the CYP2E1 reaction Reactome pathway suggest potential impact on xenobiotic and endogenous toxin metabolism, in particular those that belong to the cytochrome P450 monooxygenases (P450s), which generally perform oxidation reactions and work in concert with other enzymes such as the glutathione S-transferases, to turn toxins into a more water-soluble and less toxic chemicals [86]. CYP2E1 is known to be highly expressed in the liver of humans [87] and cattle [88], with known metabolic activities to protect against a range of xenobiotics [89].

4.5. Major histocompatibility complex

We found that the *DOA*, *DMA*, *DMB* and *DOB* genes were conserved as SCOs in all species, even though the order of these genes was reversed in the ruminant species in comparison with the monogastric species (pig and human). In contrast, the DR and DQ gene families had varying degrees of gene duplications among the six genomes. The goat genome was

the only one lacking a duplication of the *DRB* gene, while *DQA* was duplicated in river buffalo, goat, and taurine cattle but not indicine cattle, humans, or pigs. We found duplications of *DQB* in taurine cattle and river buffalo. This relatively high level of gene duplication in the MHC class IIA region (DR, DQ gene families) is congruent with the literature [90,91], with the DR and DQ gene families also known to be highly polymorphic [12]. We observed a ruminant specific class IIA gene, *DY* (i.e., *DYA*; *DYB*, alias *DIB*) which is most closely related to the *DQ* gene (i.e. *DQA*; *DQB*) by phylogenetic analysis. This suggests that a duplication of the *DQ* gene led to the evolution of the ruminant *DY* gene [92,93], which is supported by the likely chromosomal inversion in the ruminant MHC region [94–96]. *DY* may be active in dendritic cells to mediate the interactions between the microbe rich rumen and the immune system [92]. Further evidence for a ruminant specific function of *DY* can be seen in the high levels of conservation in promoter (97%), coding region (94%) and intronic regions (91%) between sheep and cattle [57].

4.6. Positively selected genes in the river buffalo

One of the signatures of adaptive evolution is a higher rate of non-synonymous than synonymous substitution. Our work takes advantage of maximum likelihood model evaluation framework [97] to detect codons that showed a higher rate of non-synonymous substitution. The most significant positively selected gene we identified was *TRPM4,* which is a member of the transient receptor potential (TRP) superfamily of ion channels that plays a role in keratinocyte differentiation [70]. In mouse, *Trpm4* is known to play a role in transduction of taste stimuli [98]. Interestingly, in a study on the association between the abomasum transcriptome and microbiota in cattle, the expression of *TRPM4* was found to be positively correlated with bacteria of the *Desulfovibrio* genus [99]. This suggests the gene may play a role in the immune system by regulating Ca^{2+} influx in response to bacterial infection. We also found growth factors genes to be positively selected, notably *TGFB1* and *IGFBP4*. In the bovine mammary gland, the expression of transforming growth factor-β1 (*TGFB1*) is increased during mammogenesis and lactogenesis [100], and the gene is known to play a role in mammary development [101]. The sequenced river buffalo in our study was from a breed selected for milk production, therefore the positively selected sites may be linked to this trait. Insulin-like growth factor (IGF)-binding protein 4 (*IGFBP4*) is highly expressed in osteoblasts and inhibits IGF-1. It is also known to be involved in bone formation [102] and in human, it is linked to fetal growth restriction in gestation [103]. *TRPM4* and *TGFB1* have not been studied in the river buffalo. *IGFBP4* is upregulated in early pregnancy in the caruncle of the water buffalo and it has been suggested that it inhibits fetally derived IGF2 [104]. The functional significance of these positively selected genes remains to be validated.

4.7. A retrotransposon in the intron of the river buffalo MHC II gene

Babesiosis is a zoonotic disease of global importance and is caused by tick-borne protozoan of the genus *Babesia* [105]. To complete its life-cycle, this parasite needs to be transmitted by ixodid ticks to a vertebrate host to replicate in the red blood cells (RBCs). The water buffalo is a known host for the tick *Rhipicephalus haemaphysaloides,* which can transmit *B. orientalis* [106]. Only a few relatively high quality genome assemblies of the parasites such as *B. ovata* and *B. divergens* [105] are available. We identified a retrotransposon in the river buffalo that has 98% identity to the *B. ovata Pol* gene at the protein level. The retrotransposon, found in the first intron of a *DRB1-like* gene, is made up of the *Pol* gene, a reverse transcriptase, and the long inverted retrotransposon terminal repeats, but does not include *env* gene that encodes the envelope protein.

The *Babesia*-related retrotransposon was not found in any other buffalo MHC genes or in the MHC of other species using our alignment criteria. Indeed it was only found in the long read based genome assembly (UOA_WB_1), but not in the short read based assembly (UMD_CASPUR_WB_2.0) of the same animal. This is likely to be due to the repetitive sequences in the retrotransposon interrupting the short read assembly of the region. A study of the human MHC II has shown that various types of repeats including retrotransposons make up more than 20% of the entire region [107]. Therefore, it is not surprising that this region is hard to assemble, which has prevented the analysis of conservation of synteny up to now. The large number of repetitive sequences in MHC II region is perhaps due to the high density of genes that are actively transcribed [108].

The *Babesia* retrotransposon was found dispersed throughout the genomes of the three ruminant species studied here, but was not observed in the monogastric species i.e. pig and human. This suggests that the retrotransposon could have inserted into the genome of the ruminant ancestor after diverging from pig. This may be because ruminants and non-ruminants are exposed to different *Babesia* parasites, ruminants being predominantly exposed to *B. orientalis* whereas *Babesia trautmanni* is more common in pigs [106,109]. The retrotransposon seems to be under purifying selection as it maintains a high level of sequence identity between *Babesia* and the ruminant species studied. Similar retroelement sequences are found abundantly in the ruminant genomes and tend to overlap genes, suggesting that these elements could still be transposing and insert into actively transcribed regions. Whether there is an advantage for host or pathogen in the integration of retroviral elements into the genome is unclear, however, there is the suggestion that presence of retroviral sequences stimulate the host immune memory and hence protection against infection [110]. Buffalo do show a higher level of tolerance to babesia than other species often with no clinical signs of disease [111]. We note possible alternative explanations for the presence of apparent *Babesia* sequences in ruminant genomes include horizontal transfer of bovid retrotransposon into the parasite, or artifactual contamination of the *B. ovata* genome assembly with cattle DNA. To assemble the genome of *B. ovata,* in vitro growth of the parasite used purified bovine RBCs that may have carried cattle sequences [72]. Future work should examine Babesia genomes using long read technologies to confirm whether the retrotransposon is indeed integrated into them. Furthermore, the functional significance of the retrotransposon insertion in the river buffalo *DRB1-like* gene should be investigated as it may impact on immune gene expression.

5. Conclusion

In summary, gene gains and losses analyses of the river buffalo genome have revealed significant changes in genes families involved in protein degradation, olfactory receptor sensing system, detoxification and immune system. Searches for positive selection have uncovered river buffalo genes that play a role in skin differentiation, mammary development and bone formation, among the hundreds of genes under selection. The finding of a *Babesia* retrotransposon in the intron of the river buffalo MHC II *DRB1-like* gene requires further investigation as its functional relevance is not known.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2021.08.021.

Acknowledgements

We would like to thank David L. Adelson and Atma Ivancevic for their assistance on retrotransposon analysis. Mention of trade names or commercial products in this publication is solely for information and does not imply recommendation or endorsement by USDA. USDA is an equal opportunity provider and employer.

Authors contributions

Conceived and designed the experiments: YR, CM, WYL, JLW, TPLS, THT; Performed analyses and interpreted results: YR, CM, THT, WYL;

Secured funding for project: WYL, JLW, TPLS; Wrote the paper: YR, CM, WYL.

References

[1] The Food and Agriculture Organization of the United Nations, About Live Animals, Data on Buffaloes [Internet], FAOSTAT, 2020 [cited 2020 Oct 15] Available from: http://www.fao.org/faostat/en/#data/TA.

[2] J. Roth, Bubalus bubalis [Internet], in: Animal Diversity Web, 2004 [cited 2020 Oct 16] Available from: https://animaldiversity.org/accounts/Bubalus_bubalis/.

[3] M.A. Villanueva, C.N. Mingala, G.A.S. Tubalinal, et al., Emerging infectious diseases in water buffalo: an economic and public health concern, in: M. A. Villanueva, C.N. Mingala, Tubalinal GAS (Eds.), Emerging Infectious Diseases in Water Buffalo - An Economic and Public Health Concern, InTech, 2018.

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

[5] A.L. Hughes, M. Nei, Evolution of the major histocompatibility complex: independent origin of nonclassical class I genes in different groups of mammals, Mol. Biol. Evol. 6 (1989) 559–579.

[6] S. Ohno, Evolution by Gene Duplication, Springer Berlin Heidelberg, Berlin, Heidelberg, 1970.

[7] M.V. Han, G.W.C. Thomas, J. Lugo-Martinez, et al., Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3, Mol. Biol. Evol. 30 (2013) 1987–1997.

[8] Z. Yang, PAML: a program package for phylogenetic analysis by maximum likelihood, Comput. Appl. Biosci. 13 (1997) 555–556.

[9] E.W. Hewitt, The MHC class I antigen presentation pathway: strategies for viral immune evasion, Immunology 110 (2003) 163–169.

[10] C. Barra, B. Alvarez, S. Paul, et al., Footprints of antigen processing boost MHC class II natural ligand predictions, Genome Med. 10 (2018) 84.

[11] S. Abduriyim, Y. Nishita, P.A. Kosintsev, et al., Evolution of MHC class I genes in Eurasian badgers, genus Meles (Carnivora, Mustelidae), Heredity 122 (2019) 205–218.

[12] T. Shiina, K. Hosomichi, H. Inoko, et al., The HLA genomic loci map: expression, interaction, diversity and disease, J. Hum. Genet. 54 (2009) 15–39.

[13] B. Lukasch, H. Westerdahl, M. Strandh, et al., Major histocompatibility complex genes partly explain early survival in house sparrows, Sci. Rep. 7 (2017) 6571.

[14] C. Vandiedonck, J.C. Knight, The human major histocompatibility complex as a paradigm in genomics research, Brief Funct. Genomic Proteomic 8 (2009) 379–394.

[15] C. Alfonso, L. Karlsson, Nonclassical MHC class II molecules, Annu. Rev. Immunol. 18 (2000) 113–142.

[16] A. Halenius, C. Gerke, H. Hengel, Classical and non-classical MHC I molecule manipulation by human cytomegalovirus: so many targets—but how many arrows in the quiver? Cell Mol. Immunol. 12 (2015) 139–153.

[17] C.V. Harding, H.J. Geuze, Class II MHC molecules are present in macrophage lysosomes and phagolysosomes that function in the phagocytic processing of Listeria monocytogenes for presentation to T cells, J. Cell Biol. 119 (1992) 531–542.

[18] F.D. Batista, N.E. Harwood, The who, how and where of antigen presentation to B cells, Nat. Rev. Immunol. 9 (2009) 15–27.

[19] J. Banchereau, R.M. Steinman, Dendritic cells and the control of immunity, Nature 392 (1998) 245–252.

[20] L. Iannuzzi, D.S. Gallagher, G.P. Di Meo, et al., Chromosomal localization of the lysozyme gene cluster in river buffalo (Bubalus bubalis L.), Chromosom. Res. 1 (1993) 253–255.

[21] N.B. Stafuzza, A.J. Greco, J.R. Grant, et al., A high-resolution radiation hybrid map of the river buffalo major histocompatibility complex and comparison with BoLA, Anim. Genet. 44 (2013) 369–376.

[22] S. Takeshima, Y. Nakai, M. Ohta, et al., Short communication: characterization of DRB3 alleles in the MHC of Japanese shorthorn cattle by polymerase chain reaction-sequence-based typing, J. Dairy Sci. 85 (2002) 1630–1632.

[23] J.C. Maillard, D. Martinez, A. Bensaid, An amino acid sequence coded by the exon 2 of the BoLA DRB3 gene associated with a BoLA class I specificity constitutes a likely genetic marker of resistance to dermatophilosis in Brahman zebu cattle of Martinique (FWI), Ann. N. Y. Acad. Sci. 791 (1996) 185–197.

[24] T. Yoshida, H. Mukoyama, H. Furuta, et al., Association of BoLA-DRB3 alleles identified by a sequence-based typing method with mastitis pathogens in Japanese Holstein cows, Anim. Sci. J. 80 (2009) 498–509.

[25] T. Yoshida, H. Mukoyama, H. Furuta, et al., Association of the amino acid motifs of BoLA-DRB3 alleles with mastitis pathogens in Japanese Holstein cows, Anim. Sci. J. 80 (2009) 510–519.

[26] M.A. Juliarena, M. Poli, L. Sala, et al., Association of BLV infection profiles with alleles of the BoLA-DRB3.2 gene, Anim. Genet. 39 (2008) 432–438.

[27] J. Mosafer, M. Heydarpour, E. Manshad, Distribution of BoLA-DRB3 allelic frequencies and identification of two new alleles in Iranian buffalo breed, Sci. World 2012 (2012), 863024, https://doi.org/10.1100/2012/863024. Epub 2012 Feb 14. PMID: 22454612; PMCID: PMC3289872.

[28] S. De, R.K. Singh, G. Butchaiah, MHC-DRB exon 2 allele polymorphism in Indian river buffalo (Bubalus bubalis), Anim. Genet. 33 (2002) 215–219.

[29] N.B. Stafuzza, L.M. Olivatto, B.C.M. Naressi, Analysis of DRB3 gene polymorphisms in Jafarabadi, Mediterranean, and Murrah buffaloes from Brazil, Genet. Mol. Res. 15 (1) (2016) 1–8, https://doi.org/10.4238/gmr.15016368. PMID: 27051015.

[30] O.E. Othman, M.G. Khodary, A.H. El-Deeb, et al., Five BoLA-DRB3 genotypes detected in Egyptian buffalo infected with Foot and Mouth disease virus serotype O, J. Genet. Eng. Biotechnol. 16 (2018) 513–518.

[31] D.M. Bickhart, B.D. Rosen, S. Koren, et al., Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome, Nat. Genet. 49 (2017) 643–650.

[32] A. Warr, N. Affara, B. Aken, et al., An improved pig reference genome sequence to enable pig genetics and genomics research, BioRxiv 9 (6) (2020).

[33] W.Y. Low, R. Tearle, R. Liu, et al., Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle, Nat. Commun. 11 (2020) 2071.

[34] B.D. Rosen, D.M. Bickhart, R.D. Schnabel, et al., De novo assembly of the cattle reference genome with single-molecule sequencing, Gigascience 9 (2020).

[35] V.A. Schneider, T. Graves-Lindsay, K. Howe, et al., Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly, Genome Res. 27 (2017) 849–864.

[36] A.C. Bertolotti, R.M. Layer, M.K. Gundappa, et al., The structural variation landscape in 492 Atlantic salmon genomes, Nat. Commun. 11 (2020) 5176.

[37] D.M. Emms, S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics, Genome Biol. 20 (2019) 238.

[38] V. Ranwez, E.J.P. Douzery, C. Cambon, et al., MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons, Mol. Biol. Evol. 35 (2018) 2582–2584.

[39] P.D. Thomas, M.J. Campbell, A. Kejariwal, et al., PANTHER: a library of protein families and subfamilies indexed by function, Genome Res. 13 (2003) 2129–2141.

[40] M.J. Sanderson, r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock, Bioinformatics 19 (2003) 301–302.

[41] S.B. Hedges, J. Dudley, S. Kumar, TimeTree: a public knowledge-base of divergence times among organisms, Bioinformatics 22 (2006) 2971–2972.

[42] G. Tsagkogeorga, S. Müller, C. Dessimoz, et al., Comparative genomics reveals contraction in olfactory receptor genes in bats, Sci. Rep. 7 (2017) 259.

[43] D. Emms, S. Kelly, *STAG: species tree inference from all genes*, BioRxiv 2018 (2018) 267914.

[44] P. Rice, I. Longden, A. Bleasby, EMBOSS: the european molecular biology open software suite, Trends Genet. 16 (2000) 276–277.

[45] K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, Mol. Biol. Evol. 30 (2013) 772–780.

[46] M. Suyama, D. Torrents, P. Bork, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments, Nucleic Acids Res. 34 (2006) W609–W612.

[47] D. Glez-Peña, D. Gómez-Blanco, M. Reboiro-Jato, et al., ALTER: program-oriented conversion of DNA and protein alignments, Nucleic Acids Res. 38 (2010) W14–W18.

[48] A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics 30 (2014) 1312–1313.

[49] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B Methodol. 57 (1995) 289–300.

[50] K. Breuer, A.K. Foroushani, M.R. Laird, et al., InnateDB: systems biology of innate immunity and beyond–recent updates and continuing curation, Nucleic Acids Res. 41 (2013) D1228–D1233.

[51] S. Bhattacharya, P. Dunn, C.G. Thomas, et al., ImmPort, toward repurposing of open access immunological assay data for translational and clinical research, Sci. Data 5 (2018) 180015.

[52] J. Kelley, B. de Bono, J. Trowsdale, IRIS: a database surveying known human immune system genes, Genomics 85 (2005) 503–511.

[53] M. Singer, C.S. Deutschman, C.W. Seymour, et al., The third international consensus definitions for Sepsis and Septic Shock (Sepsis-3), JAMA 315 (2016) 801–810.

[54] S.E. Calvano, W. Xiao, D.R. Richards, et al., A network-based analysis of systemic inflammation in humans, Nature 437 (2005) 1032–1037.

[55] G.K. Smyth, Limma: Linear models for microarray data, in: R. Gentleman, V. J. Carey, W. Huber (Eds.), Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Springer, New York, 2005, pp. 397–420.

[56] G. Yu, Q.-Y. He, ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization, Mol. BioSyst. 12 (2016) 477–479.

[57] J.D. Behl, N.K. Verma, N. Tyagi, et al., The major histocompatibility complex in bovines: a review, ISRN Vet. Sci. 2012 (2012) 872710.

[58] T. Carver, S.R. Harris, M. Berriman, et al., Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data, Bioinformatics 28 (2012) 464–469.

[59] M. Wieczorek, E.T. Abualrous, J. Sticht, et al., Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation, Front. Immunol. 8 (2017) 292.

[60] J.E. Wosen, D. Mukhopadhyay, C. Macaubas, et al., Epithelial MHC class II expression and its role in antigen presentation in the gastrointestinal and respiratory tracts, Front. Immunol. 9 (2018) 2144.

[61] S.Y. Choo, The HLA system: genetics, immunology, clinical testing, and clinical implications, Yonsei Med. J. 48 (2007) 11–23.

[62] The Comparative MHC Nomenclature Committee, MHC Bovins species [Internet], IPD-MHC Database [cited 2020 Oct 1]. Available from: https://www.ebi.ac.uk/ipd/mhc/group/BoLA/species/.

[63] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (2004) 1792–1797.

[64] S. Capella-Gutiérrez, J.M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses, Bioinformatics 25 (2009) 1972–1973.
[65] G. Talavera, J. Castresana, Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments, Syst. Biol. 56 (2007) 564–577.
[66] I. Letunic, P. Bork, Interactive tree of life (iTOL) v4: recent updates and new developments, Nucleic Acids Res. 47 (2019) W256–W259.
[67] J. Krumsiek, R. Arnold, T. Rattei, Gepard: a rapid and sensitive tool for creating dotplots on genome scale, Bioinformatics 23 (2007) 1026–1028.
[68] X. Luo, Y. Zhou, B. Zhang, et al., Understanding divergent domestication traits from the whole-genome sequencing of swamp- and river-buffalo populations, Natl. Sci. Rev. 7 (2020) 686–701.
[69] R.P. Huntley, D. Binns, E. Dimmer, et al., QuickGO: a user tutorial for the web-based Gene Ontology browser, Database (Oxford) 2009 (2009) bap010.
[70] H. Wang, Z. Xu, B.H. Lee, et al., Gain-of-function mutations in TRPM4 activation gate cause progressive symmetric erythrokeratodermia, J. Invest. Dermatol. 139 (2019) 1089–1097.
[71] A.M. Ivancevic, R.D. Kortschak, T. Bertozzi, et al., LINEs between species: evolutionary dynamics of LINE-1 retrotransposons across the eukaryotic tree of life, Genome Biol. Evol. 8 (2016) 3301–3322.
[72] J. Yamagishi, M. Asada, H. Hakimi, et al., Whole-genome assembly of Babesia ovata and comparative genomics between closely related pathogens, BMC Genomics 18 (2017) 832.
[73] R.F. Zhao, ENCODE: Deciphering Function in the Human Genome, 2012.
[74] J. Zhang, Evolution by gene duplication: an update, Trends Ecol. Evol. (Amst) 18 (2003) 292–298.
[75] P. Dutta, A. Talenti, R. Young, et al., Whole genome analysis of water buffalo and global cattle breeds highlights convergent signatures of domestication, Nat. Commun. 11 (2020) 4739.
[76] R.D. Sainz, L.G. Barioni, P.V. Paulino, Growth patterns of Nellore vs British beef cattle breeds assessed using a dynamic, mechanistic model of cattle growth and composition, in: Nutrient Digestion and Utilization in Farm Animals: Modelling Approaches, 2006, p. 160, books.google.com.
[77] G.M. Hughes, E.S.M. Boston, J.A. Finarelli, et al., The birth and death of olfactory receptor gene families in mammalian niche adaptation, Mol. Biol. Evol. 35 (2018) 1390–1406.
[78] L. Buck, R. Axel, A novel multigene family may encode odorant receptors: a molecular basis for odor recognition, Cell 65 (1991) 175–187.
[79] S. Hayden, M. Bekaert, T.A. Crider, et al., Ecological adaptation determines functional mammalian olfactory subgenomes, Genome Res. 20 (2010) 1–9.
[80] S. Hayden, M. Bekaert, A. Goodbla, et al., A cluster of olfactory receptor genes linked to frugivory in bats, Mol. Biol. Evol. 31 (2014) 917–927.
[81] I. Khan, Z. Yang, E. Maldonado, et al., Olfactory receptor subgenomes linked with broad ecological adaptations in Sauropsida, Mol. Biol. Evol. 32 (2015) 2832–2843.
[82] K. Lee, D.T. Nguyen, M. Choi, et al., Analysis of cattle olfactory subgenome: the first detail study on the characteristics of the complete olfactory receptor repertoire of a ruminant, BMC Genomics 14 (2013) 596.
[83] H.M. Tan, W.Y. Low, Rapid birth-death evolution and positive selection in detoxification-type glutathione S-transferases in mammals, PLoS One 13 (2018), e0209336.
[84] W.Y. Low, H.L. Ng, C.J. Morton, et al., Molecular evolution of glutathione S-transferases in the genus Drosophila, Genetics 177 (2007) 1363–1375.
[85] J.H. Thomas, Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates, PLoS Genet. 3 (2007), e67.
[86] D. Sheehan, G. Meade, V.M. Foley, et al., Structure, function and evolution of glutathione transferases: implications for classification of non-mammalian members of an ancient enzyme superfamily, Biochem. J. 360 (2001) 1–16.
[87] I. Bièche, C. Narjoz, T. Asselah, et al., Reverse transcriptase-PCR quantification of mRNA levels from cytochrome (CYP)1, CYP2 and CYP3 families in 22 different human tissues, Pharmacogenet. Genomics 17 (2007) 731–742.
[88] M.J. Kuhn, A.K. Putman, L.M. Sordillo, Widespread basal cytochrome P450 expression in extrahepatic bovine tissues and isolated cells, J. Dairy Sci. 103 (2020) 625–637.
[89] J. Chen, S. Jiang, J. Wang, et al., A comprehensive review of cytochrome P450 2E1 for xenobiotic metabolism, Drug Metab. Rev. 51 (2019) 178–195.
[90] L. Andersson, L. Rask, Characterization of the MHC class II region in cattle. The number of DQ genes varies between haplotypes, Immunogenetics 27 (1988) 110–120.
[91] K.T. Ballingall, K. Dicks, P. Kyriazopoulou, et al., Allelic nomenclature for the duplicated MHC class II DQ genes in sheep, Immunogenetics 71 (2019) 347–351.
[92] K.T. Ballingall, S.A. Ellis, N.D. MacHugh, et al., The DY genes of the cattle MHC: expression and comparative analysis of an unusual class II MHC gene pair, Immunogenetics 55 (2004) 748–755.
[93] C. Li, R. Huang, F. Nie, et al., Organization of the addax major histocompatibility complex provides insights into ruminant evolution, Front. Immunol. 11 (2020) 260.
[94] M. Band, J.H. Larson, J.E. Womack, et al., A radiation hybrid map of BTA23: identification of a chromosomal rearrangement leading to separation of the cattle MHC class II subregions, Genomics 53 (1998) 269–275.
[95] C.P. Childers, H.L. Newkirk, D.A. Honeycutt, et al., Comparative analysis of the bovine MHC class IIb sequence identifies inversion breakpoints and three unexpected genes, Anim. Genet. 37 (2006) 121–129.
[96] C.L. Brinkmeyer-Langford, C.P. Childers, K.L. Fritz, et al., A high resolution RH map of the bovine major histocompatibility complex, BMC Genomics 10 (2009) 182.
[97] Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood, Mol. Biol. Evol. 24 (2007) 1586–1591.
[98] D. Dutta Banik, L.E. Martin, M. Freichel, et al., TRPM4 and TRPM5 are both required for normal signaling in taste receptor cells, Proc. Natl. Acad. Sci. U. S. A. 115 (2018) E772–E781.
[99] N. Gaowa, W. Li, B. Murphy, et al., The effects of artificially dosed adult rumen contents on abomasum transcriptome and associated microbial community structure in calves, Genes (Basel) 12 (2021).
[100] A. Plath, R. Einspanier, F. Peters, et al., Expression of transforming growth factors alpha and beta-1 messenger RNA in the bovine mammary gland during different stages of development and lactation, J. Endocrinol. 155 (1997) 501–511.
[101] L.D.D. Vries, T. Casey, H. Dover, et al., Effects of transforming growth factor-β on mammary remodeling during the dry period of dairy cows, J. Dairy Sci. 94 (2011) 6036–6046.
[102] K. Mense, J. Heidekorn-Dettmer, E. Wirthgen, et al., Increased concentrations of insulin-like growth factor binding protein (IGFBP)-2, IGFBP-3, and IGFBP-4 are associated with fetal mortality in pregnant cows, Front Endocrinol. (Lausanne) 9 (2018) 310.
[103] Q. Qiu, M. Bell, X. Lu, et al., Significance of IGFBP-4 in the development of fetal growth restriction, J. Clin. Endocrinol. Metab. 97 (2012) E1429–E1439.
[104] Y. Pandey, A.R. Pooja, H.L. Devi, et al., Expression and functional role of IGFs during early pregnancy in placenta of water buffalo, Theriogenology 161 (2021) 313–331.
[105] L.M. González, K. Estrada, R. Grande, et al., Comparative and functional genomics of the protozoan parasite Babesia divergens highlighting the invasion and egress processes, PLoS Negl. Trop. Dis. 13 (2019), e0007680.
[106] L. He, Q. Liu, B. Yao, et al., A historical overview of research on Babesia orientalis, a protozoan parasite infecting water Buffalo, Front. Microbiol. 8 (2017).
[107] G. Andersson, A.C. Svensson, N. Setterblad, et al., Retroelements in the human MHC class II region, Trends Genet. 14 (1998) 109–114.
[108] D. Taruscio, L. Manuelidis, Integration site preferences of endogenous retroviruses, Chromosoma 101 (1991) 141–156.
[109] G. Uilenberg, Babesia–a historical overview, Vet. Parasitol. 138 (2006) 3–10.
[110] J.L. Hurwitz, B.G. Jones, E. Charpentier, et al., Hypothesis: RNA and DNA viral sequence integration into the mammalian host genome supports long-term B cell and T cell adaptive immunity, Viral Immunol. 30 (2017) 628–632.
[111] Y.S. Mahmmod, Molecular detection of natural Babesia bovis infection from clinically infected and apparently healthy water Buffaloes (*Bubalus bubalis*) and crossbred cattle, J. Buffalo Sci. 1 (1) (2012).