


RESOURCE ARTICLE

A chromosome-scale reference genome and genome-wide genetic variations elucidate adaptation in yak

Qiu-mei Ji^{1,2} | Jin-wei Xin^{1,2} | Zhi-xin Chai³ | Cheng-fu Zhang^{1,2} | Yangla Dawa^{1,2} | Sang Luo^{1,2} | Qiang Zhang^{1,2} | Zhandui Pingcuo^{1,2} | Min-Sheng Peng⁴ | Yong Zhu^{1,2} | Han-wen Cao^{1,2} | Hui Wang³ | Jian-lin Han⁵ | Jin-cheng Zhong³ 

¹State Key Laboratory of Hullless Barley and Yak Germplasm Resources and Genetic Improvement, Lhasa, China

²Institute of Animal Science and Veterinary Research, Tibet Academy of Agricultural and Animal Husbandry Sciences, Lhasa, China

³Key Laboratory of Qinghai-Tibetan Plateau Animal Genetic Resource Reservation and Utilization, Sichuan Province and Ministry of Education, Southwest Minzu University, Chengdu, China

⁴State Key Laboratory of Genetic Resources and Evolution & Yunnan Laboratory of Molecular Biology of Domestic Animals, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China

⁵CAAS-ILRI Joint Laboratory on Livestock and Forage Genetic Resources, Institute of Animal Science, Chinese Academy of Agriculture Sciences (CAAS), Beijing, China

Correspondence

Jin-cheng Zhong, Key Laboratory of Qinghai-Tibetan Plateau Animal Genetic Resource Reservation and Utilization, Sichuan Province and Ministry of Education, Southwest Minzu University, Chengdu, China.
Email: zhongjincheng518@126.com

Funding information

This study was supported by Tibetan Autonomous Region special grants for 'Functional Genomics and Germplasm Enhancement of Yak', the Key Research and Development Projects in Tibet: Preservation of Characteristic Biological Germplasm Resources and Utilization of Gene Technology in Tibet (Grant No. XZ202001ZY0016N), the Second Tibetan Plateau Scientific Expedition and Research Program (STEP) (Grant No. 2019QZKK0501) and the Program of National Beef Cattle and Yak Industrial Technology System (No. CARS-37).

Abstract

Yak is an important livestock animal for the people indigenous to the harsh, oxygen-limited Qinghai-Tibetan Plateau and Hindu Kush ranges of the Himalayas. The yak genome was sequenced in 2012, but its assembly was fragmented because of the inherent limitations of the Illumina sequencing technology used to analyse it. An accurate and complete reference genome is essential for the study of genetic variations in this species. Long-read sequences are more complete than their short-read counterparts and have been successfully applied towards high-quality genome assembly for various species. In this study, we present a high-quality chromosome-scale yak genome assembly (BosGru_PB_v1.0) constructed with long-read sequencing and chromatin interaction technologies. Compared to an existing yak genome assembly (BosGru_v2.0), BosGru_PB_v1.0 shows substantially improved chromosome sequence continuity, reduced repetitive structure ambiguity, and gene model completeness. To characterize genetic variation in yak, we generated de novo genome assemblies based on Illumina short reads for seven recognized domestic yak breeds in Tibet and Sichuan and one wild yak from Hoh Xil. We compared these eight assemblies to the BosGru_PB_v1.0 genome, obtained a comprehensive map of yak genetic diversity at the whole-genome level, and identified several protein-coding genes absent from the BosGru_PB_v1.0 assembly. Despite the genetic bottleneck experienced by wild yak, their diversity was nonetheless higher than that of domestic

Qiu-mei Ji, Jin-wei Xin, Zhi-xin Chai, and Cheng-fu Zhang contributed equally to this work.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2020 The Authors. Molecular Ecology Resources published by John Wiley & Sons Ltd

yak. Here, we identified breed-specific sequences and genes by whole-genome alignment, which may facilitate yak breed identification.

KEYWORDS

genetic introgression, regions of homozygosity, single-nucleotide polymorphisms, whole-genome sequencing, yak

1 | INTRODUCTION

Domestic yak (*Bos grunniens*) is a vital livestock animal in the harsh, oxygen-limited Qinghai-Tibetan Plateau and Hindu Kush ranges of the Himalayas. Yak is an important source of food (meat and milk), transportation, shelter and fuel for the people indigenous to these regions (Wiener, JianLin, & Ruijun, 2003).

In 2012, the genome of a female domestic yak was sequenced by the Illumina-based (HiSeq 2000 platform; read length ~100 bp) whole-genome shotgun method (Qiu et al., 2012). However, genome assembly with short reads may yield incomplete key genomic regions/elements. Moreover, this approach prevents accurate gene annotation and is not amenable to profound comparative genomic analyses (Alkan, Sajjadian, & Eichler, 2010). Certain regions of the coding exons were not assembled in the previously available yak genome. For example, the highly polymorphic and repetitive major histocompatibility complex (MHC) is difficult to assemble using short reads. MHC participates in overall host pathogen resistance/susceptibility in the host and is, therefore, critical to immunity and reproductive success. Hence, an accurate and complete yak reference genome is required to facilitate genetic studies and develop improvement programs for this species.

To this end, we exploited sequencing technologies including PacBio long-read sequences coupled with Illumina short reads, BioNano and Hi-C to assemble the genome of a female domestic yak. This strategy has already been implemented in the assembly of the genomes of humans (Burton et al., 2013), *Arabidopsis* (Wang et al., 2015) and goats (Bickhart et al., 2017).

There are twelve officially recognized domestic yak breeds in China. They were initially classified according to their origin, location and minor differences in production traits (Wiener et al., 2003). However, virtually no measurements of the magnitude of their genetic differences were reported. To investigate the genetic diversity among yak breeds, we generated de novo assemblies for seven recognized domestic varieties. Modern domestic yak is descended from wild yak. Although they live in similar environments, the domestic yak is smaller than the latter and they differ in terms of temperaments. Unlike other livestock, domesticated yak has a low domestication level and cannot be held in captivity as their domestication history is comparatively short. It is thought that yak domestication began in the late Stone Age (~10,000 y ago) in the *Changtang* region covering more than half of Tibet (Wiener et al., 2003). Based on the analyses of 59 domesticated and 13 wild yak, it was estimated that domestication started ~7,300 y ago (Qiu et al., 2015).

Genetic introgression between yak and Tibetan cattle has been previously reported (Medugorac et al., 2017). Yak have a low level of

domestication (Wiener et al., 2003), a high acute sense of smell, and strong vigilance. For these reasons, they are difficult to manage than the other domesticated animals in the same region. Introgression from Tibetan cattle to yak may have been a factor in yak domestication. Therefore, it is informative to investigate the similarities and differences in introgression from Tibetan cattle to the seven domestic yak breeds.

2 | MATERIALS AND METHODS

2.1 | Ethics statement

The Guidelines for Experimental Animal Management of Southwest Minzu University were followed throughout the present study. Animal care was performed according to the regulations of the Administration of Affairs Concerning Experimental Animals (Ministry of Science and Technology; revised June 2004) and approved by the Institutional Animal Care Committee of Southwest Minzu University, Chengdu, Sichuan, China.

2.2 | Genome sequencing and initial assembly

High-quality DNA was extracted from the peripheral blood of a female yak in Riwoqe County, Tibet. SMRT sequencing libraries were constructed with a Blood and Cell Culture DNA Mini Kit (Qiagen, Hilden, Germany). A total of 142 SMRT cells generated 184.6 Gbp of subread bases with a mean read length of 9.5 Kbp on a PacBio RS II instrument (Pacific Biosciences). The *FALCON* (version 0.5.0) pipeline was used for the initial assembly. First, all overlaps in the raw reads were identified. Then, the read error was corrected by leveraging the overlap information. Next overlaps in the corrected reads were detected. This step required no consensus calling. Finally, the string graph assembly and contig sequence output were generated in FASTA format. To improve assembly quality, 113.34 Gbp of Illumina short reads were generated from the same individual. Using *PILON* (version 1.23; Walker et al., 2014), 845,002 homozygous insertions, 166,908 deletions and 2,355,196 substitutions were identified and corrected.

2.3 | BioNano assembly

DNA from the same female yak used for PacBio sequencing was extracted and processed according to BioNano Genomics guidelines.

Raw data were assembled with the BIONANO SOLVE (version 3.1.00) assembly pipeline (BioNano Genomics). Combining this assembly with the initial one yielded a superior assembly with a scaffold N50 of 65.67 Mbp and a maximum scaffold length of 128.62 Mbp.

2.4 | Hi-C sequencing and assembly

Blood cells were extracted from the same aforementioned female yak. Hi-C libraries were created from yak whole-blood cells, 2–5 million cells were cross-linked and digested with the restriction enzyme *HindIII*. The sticky ends of all fragments were biotinylated, ligated to each other to form chimeric circles, enriched, sheared and processed into sequencing libraries wherein the individual templates were chimeras of the physically associated DNA molecules from the original cross-linking. The libraries were sequenced on Illumina HiSeq X Ten (Illumina) following the manufacturer's instructions. A total of 272.01 Gbp of Hi-C reads was generated. Paired-end reads were uniquely mapped onto the BioNano assembly and classified into 30 groups using 3d-DNA (20,180,922) as the final assembly, and referred to as BosGru_PB_v1.0. The exact locations of each scaffold in the 30 groups were based on the collinearity between yak and cattle (UMD3.1.1).

2.5 | Assembly quality assessment

The Illumina short reads were aligned to the BosGru_PB_v1.0 assembly with BWA (version 0.7.15; Li & Durbin, 2010). SAMTOOLS (version 1.9; Li, 2011) was used to detect single-nucleotide polymorphisms (SNPs). The filtering conditions were as follows: quality score >50; site sequencing depth >4; and mean copy number of nearby regions <2. LUMPY-SV (version 0.2.13; Layer, Chiang, Quinlan, & Hall, 2014) and the reference-free assembly validation software FRCBAM (version 1.3.0; Vezzi, Narzisi, & Mishra, 2012) evaluated assembly correctness. Erroneous bases in the BosGru_PB_v1.0 assembly were identified with the variant-calling software FREEBAYES (version 1.0.0; Garrison & Marth, 2012). Genome assembly quality was determined by BUSCO assessment using the 4,104 single-copy orthologous genes in BosGru_PB_v1.0 and BosGru_v2.0.

2.6 | Repeat and gene annotation

The same pipeline was used for the BosGru_v2.0 and BosGru_PB_v1.0 yak genome assemblies. Transposable elements in the yak genome were identified by a combination of homology-based and de novo approaches. Tandem repeats were identified with Tandem Repeat Finder (Benson, 1999). Interspersed repeats were characterized by homology-based identification with REPEATMASKER (version 4.0.3; Smit, Hubley, & Green, 1996) and the repeat database Repbase (20,170,127) (Bao, Kojima, & Kohany, 2015). Repeated proteins were identified with RepeatProteinMask (Smit et al., 1996) and

a transposable element protein database. De novo-identified interspersed repeats were annotated with REPEATMODELER (version 1.0.4; Price, Jones, & Pevzner, 2005). LTR_FINDER (version 1.0.5; Xu & Wang, 2007) identified the long terminal repeats (LTRs). These results were used to generate the de novo repeat libraries and REPEATMASKER was run against them. All repeats identified in this manner were added to the total count of interspersed repeats.

Yak protein-encoding genes were annotated using a combination of homology prediction, de novo gene prediction and RNA-seq based evidence. For homology gene prediction, the protein sequences from six close species (BosGru_v2.0, *Bos taurus*, *Camelus ferus*, *Capra hircus*, *Pantholops hodgsonii* and *Sus scrofa*) were mapped to BosGru_PB_v1.0 with TBLASTN (version 2.2.28+; Altschul, Gish, Miller, Myers, & Lipman, 1990). Exonerate (version 2.2.0; Birney, Clamp, & Durbin, 2004) predicted the gene model based on the alignment results. De novo gene prediction was performed with GENSCAN (Burge & Karlin, 1997), AUGUSTUS (version 2.7; Stanke et al., 2006) and GLIMMERHMM (version 3.0.1; Majoros, Pertea, & Salzberg, 2004) based on the repeat-masked genome. RNA-seq data were downloaded from NCBI SRA under SRP009200. RNA-seq reads were aligned to the genome with Tophat and the alignments served as Cufflinks inputs. MAKER (version 2.28; Cantarel et al., 2008) was then used to integrate the predicted genes. Manual integration was performed to construct the final gene set, which served as a search query against the KEGG (Kanehisa & Goto, 2000), SWISS-PROT (Bairoch & Apweiler, 2000) and TREMBL (Bairoch & Apweiler, 2000) protein databases to identify gene functions.

2.7 | Gap resolution

Whole-genome alignment of BosGru_v2.0 against BosGru_PB_v1.0 was performed with NUCMER (v3.9.4alpha; Kurtz et al., 2004). After filtering the isolated fragment alignments with custom scripts, a linear block alignment was created between the scaffolds of the two assemblies. Gaps precisely located in the region between two neighbouring alignment blocks were considered closed and synteny type (Figure S1).

All 500-bp fragments upstream and downstream of each gap region in BosGru_v2.0 were extracted with BEDTOOLS (<http://bedtools.readthedocs.io/en/latest/>) and aligned to BosGru_PB_v1.0 with BWA MEM (0.7.15-r1140) (Li & Durbin, 2010). The gap region-corresponding sequence in BosGru_PB_v1.0 was identified by linear block alignment and then the BWA alignment was checked. If both fragments were successfully aligned based on the credibility interval and the two flanking sequences were mapped between the two assemblies in a consistent order, the gap was considered a paired-end type. The other gap with two flanking sequences mapped in the opposite order was considered an overlap type. In the latter case, the gap may have actually been the consequence of incorrect assembly. When both fragments aligned on the same scaffold and across the gap credibility interval, the gap was considered an over-cross type. Moreover, a closed gap was considered 'long length' when the closed

sequence was >10 kb, that is, only two gaps greater than 10 Kbp in BosGru_v2.0. For fragments aligned to two separate scaffolds, the region was considered unmapped. During identification, if gap flanking sequences were aligned to ≥ 2 regions, all pair combinations were analysed to detect closed gaps and to determine their type based on the target location and the credibility interval.

2.8 | Yak-specific gene families' analysis

Genome and annotation data for *Pantholops hodgsonii* (GCF_000400835.1), *Camelus ferus* (GCF_000311805.1), *Bos taurus* (GCF_000003055.6), *Capra hircus* (GCF_001704415.1), *Homo sapiens* (GRCh37) and *Ovis aries* (GCF_000298735.2) were downloaded from the National Center for Biotechnology Information (NCBI) database. An all-versus-all blastp alignment with the protein sequences from yak (BosGru_PB_v1.0) and the aforementioned mammalian species and gene family clustering were performed with ORTHOMCL version 2.0.9 (Li, Stoeckert, & Roos, 2003; v2.0.9). Yak-specific gene families were identified in the clustering output.

2.9 | MHC analysis

To identify MHC-related scaffolds in the yak assemblies, the yak chromosome sequences were aligned with NUCMER using its default parameters against the taurine cattle genome sequences in the range of 1 bp–30.7 Mbp on chr23 (UMD3.1.73; Ensembl release 73) containing bovine leucocyte antigen (BoLA). The best hits with the longest 'matches' in raw results were retained while low-quality and redundant alignments were manually removed. MHC class I and II genes and the anchor genes of MHC regions were identified by their gene names. Nucleotide-level synteny of taurine cattle and yak was inferred with BLASTN (Altschul et al., 1990; Camacho et al., 2009) [20, 30] (Altschul et al., 1990; Camacho et al., 2009) High-scoring sequence pairs > 500 bp were plotted.

2.10 | Segmental duplication (SD) detection

Whole-genome assembly comparison methods were used to call segmental duplicates (SDs). Self-alignment of BosGru_v2.0 or BosGru_PB_v1.0 was achieved with LASTZ (version 1.04.00) using the parameters $T = 2$ and $Y = 9,400$. Segmental duplication was defined as two sequences >1 kbp and with >90% identity for BosGru_v2.0 and BosGru_PB_v1.0.

2.11 | Genome sequencing and assembly for eight different yaks

Four Tibetan (SB, PL, SZ and NY), two Sichuan (JL and MW) and one Gansu (TZ) domestic yak breeds were selected for whole-genome

sequencing and assembly. DNA was extracted from the ears of the Tibetan and Gansu breeds, the blood of the Sichuan breeds, and the skin of the wild yak from Kunlun Spring, Hoh Xil. A whole-genome shotgun strategy and next-generation sequencing (NGS) technology were run on the Illumina HiSeq 2500 platform (Illumina). Each genome was sequenced using short-insert (180 and 500 bp) and long-insert (2 and 5 kbp) DNA libraries. SOAPDENOV0 (version 2.04) was used to assemble each genome. The protein-coding gene annotation was the same as that used for BosGru_PB_v1.0.

2.12 | Identification of single-nucleotide polymorphisms (SNPs) and regions of homozygosity (RoH)

BWA MEM mapped high-quality, paired-end, short-insert reads for the seven domesticated and one wild yak onto the BosGru_PB_v1.0 genome. SAMtools merged and sorted the alignment results. Through HaplotypeCaller analysis, GATK called the SNPs based on the local de novo assembler and the HMM likelihood function. RoHs were defined as continuous or uninterrupted stretches of DNA sequences without heterozygosity in the diploid state. PLINK (version 1.9) was used to identify regions ≥ 200 kbp encompassing 50 homozygous alternate SNPs relative to BosGru_PB_v1.0 and a maximum of one heterozygous SNP.

2.13 | Missing sequences and absent genes in the new BosGru_PB_v1.0 reference genome

Whole-genome alignments of assemblies for the seven domestic and one wild yak were run against BosGru_PB_v1.0 with NUCMER (Kurtz et al., 2004). The unmapped genome regions of the eight different yaks were selected as candidate missing regions of BosGru_PB_v1.0. The all short-insert reads for the eight assemblies were mapped to the BosGru_PB_v1.0 reference genome and to their own assemblies. If neither end of a read mapped to BosGru_PB_v1.0 but did map to its own assembly, those mapped regions from the eight assemblies were assumed to be missing from the BosGru_PB_v1.0 reference genome. The overlapped candidate regions were combined into one sequence missing from BosGru_PB_v1.0. Genes present in these missing sequences were annotated as described in the preceding subsections and compared with the missing regions in the BosGru_PB_v1.0 reference genome. Genes with overlapping exons were considered absent in BosGru_PB_v1.0.

2.14 | Detection of introgressed haplotypes

The combined data set comprising both Tibetan cattle and domestic yak was processed by read-aware phasing with SHAPEIT (v2. r904; Delaneau, Marchini, & Zagury, 2011). ChromoPainter in FINESTRUCTURE (version 2.1.3; Copenhaver, Lawson, Hellenthal, Myers, &

Falush, 2012) identified migrant tracts resulting from introgression from the Tibetan cattle to domestic yak. The program runs a hidden Markov model to estimate the probability of Tibetan cattle and domestic yak ancestry at each variable position along the genome according to haplotype similarity patterns. A probability threshold of 0.85 was set for the detection of introgressed tracts.

3 | RESULTS

3.1 | Yak genome sequencing and assembly

We generated 184.9 Gbp of PacBio sequencing data (Table S1) with N50 read length = 14 kbp and used them to generate the initial assembly in FALCON (Chin et al., 2016). The total size of the corrected initial assembly without gaps was 2.77 Gbp. The longest contig was 90 Mbp and the contig N50 length was 14.89 Mbp. These results constituted a tenfold improvement over the previous scaffold N50 length and a 652-fold improvement over the former contig N50 length (Table S3). The 210 longest contigs accounted for 90% of the total genomic length, the corresponding N90 length was 2.3 Mbp (Table 1). We used BioNano data to scaffold the PacBio contigs. The N50 scaffold length was 65.67 Mbp which was $\sim 4.41 \times$ longer than the initial PacBio contig N50 length (Table 1). We used the Hi-C data to link the BioNano scaffolds to a chromosome scale. We aligned the Hi-C reads (Table S2) to the BioNano-based assembly and applied 3d-DNA (Dudchenko et al., 2017) to cluster the contigs. After the parameter adjustment, 94.97% of the contigs in the initial assembly ($n = 480$) were clustered into 30 groups. To compensate for inconsistent Hi-C sequence direction and order, we performed whole-genome alignment against a taurine cattle (*Bos taurus*) reference genome and reorganized the scaffolds in each group. Overall, the yak genome assembly had high quality and good collinearity with the taurine cattle reference genome (Figure 1b).

3.2 | New assembly quality assessment

We assessed the quality of our new yak genome assembly (BosGru_PB_v1.0) by comparing it with the yak BosGru_v2.0 reference assembly (Qiu et al., 2012). The latter was assembled with SOAPdenovo2 using Illumina short reads. Relative to BosGru_v2.0, BosGru_PB_v1.0 was more complete and accurate. Completeness was evaluated according to read mapping percentage, additional assembled sequencing in BosGru_PB_v1.0, and gaps in BosGru_v2.0.

Accuracy was assessed on the basis of assembly error, gene structure and MHC/SD.

The Illumina short reads used for the BosGru_v2.0 assembly were aligned to both assemblies. There were 9.6% more reads mapped to BosGru_PB_v1.0 than BosGru_v2.0 (Table S5). Therefore, BosGru_PB_v1.0 was more complete than BosGru_v2.0. The total contig length of BosGru_PB_v1.0 was 246.79 Mbp greater than that of BosGru_v2.0. We used BUSCO (Simao, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) to determine the completeness of the assemblies. BosGru_v2.0 and BosGru_PB_v1.0 had 91.6% and 92.5% complete orthologous genes, respectively (Table S4).

PacBio sequencing technology supported the assembly of additional repetitive regions in the yak genome. In total, 1,150 Mbp (42.82%) of the total BosGru_v2.0 genome was annotated as repetitive sequences. In contrast, the novel BosGru_PB_v1.0 assembly included another 269 Mbp of repetitive sequences that were missing in the BosGru_v2.0 assembly. Comparison of the repetitive landscapes of the BosGru_v2.0 and BosGru_PB_v1.0 assemblies revealed that recently inserted transposable elements were assembled in far greater numbers in the BosGru_PB_v1.0 than the BosGru_v2.0 assembly (Figures S2 and S3). The overall number of repetitive elements increased at all divergence levels and especially at the lower ones (Figure S4). The total length of the satellite/centr repeat sequences (189.71 Mbp) in BosGru_PB_v1.0 was 97.66-fold longer than that in BosGru_v2.0 (1.92 Mbp) (Figure S2). Very few satellite/centr-related contigs could be localized to specific chromosomes.

The total gap length in BosGru_v2.0 was 118 Mbp. We aligned BosGru_PB_v1.0 with BosGru_v2.0 to determine the number of gaps filled in the former. In BosGru_PB_v1.0, 93.2% of the 192,006 gaps detected in BosGru_v2.0 were closed. Of these, 49% and 31% were closed by synteny and paired-end relationships, respectively, 10% may have been the result of overlapping sequence assembly error, and 3% had low accuracy because of cross-synteny block (Figure S5). Gap closure added ≥ 101.19 Mbp of euchromatic sequences that dramatically improved the repetitive element and gene annotations (Table S7; Figures S6 and S7). Most of the filled gaps (72.39%) were mapped to LINE/BovB and then to LINE/L1 (Adelson, Raison, & Edgar, 2009). However, 73.97% of the gaps 5–10 kbp in length were mapped only to LINE/L1 (Figure S8).

We called single-nucleotide polymorphisms (SNPs) with a pipeline based on short-read alignment. There were fewer homozygous SNPs (70,744 vs. 462,128) but more heterozygous SNPs (3,847,016 vs. 3,188,054) in BosGru_PB_v1.0 than in BosGru_v2.0 (Table S6). BosGru_PB_v1.0 had four-fold fewer structural variations (SV) per 100 Mbp (891 vs. 3,255) and eleven-fold fewer deletions (4,976 vs.

TABLE 1 Genome assembly statistics

Assembly	Number of contigs	Number of scaffolds	Unplaced contigs	Contig NG50/Mb	Scaffold NG50/Mb	Total size/Gb
PacBio	5,553	—	—	14.89	—	2.77
PacBio + BioNano	5,475	5,168	—	14.74	65.67	2.80
PacBio + BioNano + Hi-C	5,468	30	5,063	14.74	101.61	2.80

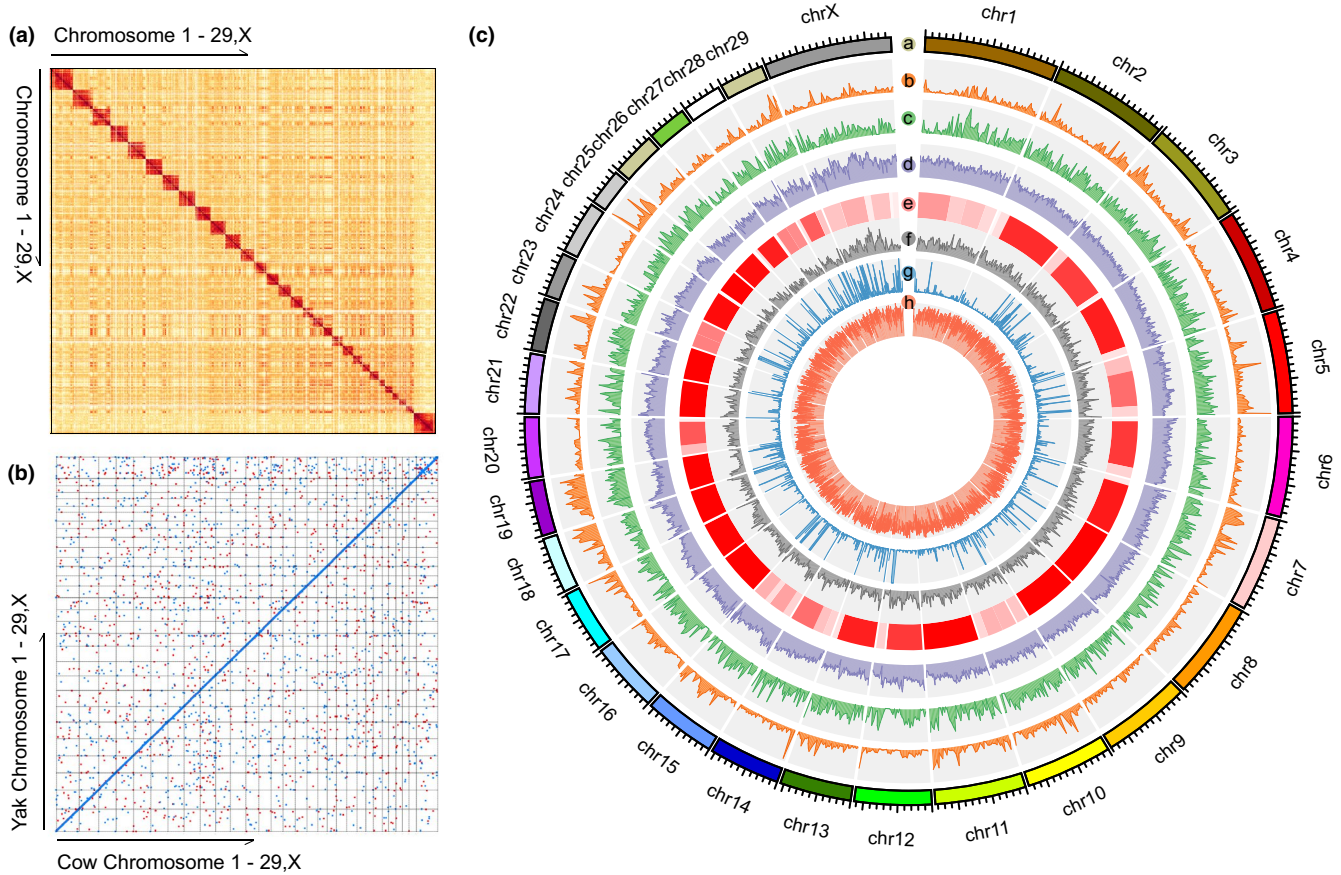


FIGURE 1 Landscape of high-quality yak chromosomes. (A) Hi-C heatmap for each group. (B) Whole-genome alignment revealing strong collinearity between yak and taurine cattle chromosomes. (C) Yak genome circos. (a) Yak chromosomes. (b) Exon coverage ratio for each 2-Mbp window. (c) Intron coverage ratio for each 2-Mbp window. (d) Intron coverage ratio for each 1-Mbp window. (e) Super-scaffold heatmap in chromosomes. Colour depth increases with super-scaffold chromosome length ratio. (f) Filled gap distribution versus BosGru_v2.0. (g) SD distribution. (h) Distribution of chromosome region identity between yak and taurine cattle [Colour figure can be viewed at wileyonlinelibrary.com]

56,115) than BosGru_v2.0 (Table S8). Hence, the BosGru_PB_v1.0 assembly represented a significant improvement over the BosGru_v2.0 assembly.

To predict the genes in the BosGru_PB_v1.0 assembly, we aligned the gene sequences (including BosGru_v2.0) for six closely related species to BosGru_PB_v1.0 and combined the output of RNA-seq and ab initio gene prediction. We predicted that 22,477 genes (Table S9) were similar to the previously predicted 22,282 genes. To illustrate the difference between BosGru_PB_v1.0 and BosGru_v2.0, we plotted line charts based on the distribution of the numbers of different lengths of these gene structures (Figure S9). The genes in BosGru_PB_v1.0 were more complete than those in BosGru_v2.0 because the former filled 42,922 gaps in the introns of 10,764 genes and 5,588 gaps overlapping the exons of 5,180 genes (Figure S10).

We compared gene coverage among the taurine cattle, BosGru_PB_v1.0, and BosGru_v2.0 genomes to evaluate the assemblies. BosGru_PB_v1.0 covered 99.1% of the genes whereas BosGru_v2.0 covered only 93.5% of them. In contrast, the taurine cattle genome covered 96.6% and 97.2% of the genes in BosGru_PB_v1.0 and BosGru_v2.0, respectively. A functional enrichment was performed

on 264 common noncovered genes revealed that the 'glycerolipid metabolism' and 'regulation of actin cytoskeleton' pathways and the 'secretion', 'triglyceride lipase activity', and 'signal release' functions were enriched. All of these are yak-specific energy metabolism and intracellular stability mechanisms.

Previous studies subdivided bovine class II MHC genes into class IIa and IIb subregions (Balligall, Fardoe, & McKeever, 2008; Deverson et al., 1991; Escayg, Hickford, Bullock, Montgomery, & Dodds, 2009; van der Poel, Groenen, Dijkhof, Ruyter, & Giphart, 1990; Scott, Choi, & Brandon, 1987; Snibson, Maddox, Fabb, & Brandon, 1998). Class IIa is continuous with classes I and III and spans ~3.3 Mbp of chr23 in taurine cattle. Class IIb spans only ~400 kbp and is ~18.5 Mbp distant from class IIa (Gao et al., 2010). We identified and assembled potential MHC-related scaffolds of these regions in BosGru_PB_v1.0 and BosGru_v2.0 on the basis of alignment with taurine cattle MHC and conserved anchoring genes (Belov et al., 2006). The MHC length and number of genes in both yak assemblies were consistent with those of taurine cattle (Table S10). BosGru_PB_v1.0 and BosGru_v2.0 had class IIb genes assembled in one scaffold and presented with nearly the same syntenic relationship as taurine cattle MHC class IIb genes.

The class IIb regions of taurine cattle and yak differed by three inversions and two rearrangements (Figure 2a). These discrepancies were localized to the intergenic regions and introns. For the other MHC regions, BosGru_PB_v1.0 had only 15 high-quality scaffolds whereas BosGru_v2.0 had 138 scaffolds (Table S11). There were 127 and 116 protein-coding genes in the BosGru_PB_v1.0 and BosGru_v2.0 MHC regions, respectively. Most of the conserved genes between yak and taurine cattle were highly collinear. The BosGru_PB_v1.0 MHC bore relatively longer fragments and had certain insertions in its 0.3–1-Mbp region that were absent in the same region of BosGru_v2.0 MHC (Figure 2b). Comprehensive MHC gene characterization is essential for successful vaccine development and elucidation of mate choice and other social behaviours in yak.

Segmental duplications may play a role in the creation of new genes (Emanuel & Shaikh, 2001; Taylor & Raes, 2004). Using a whole-genome analysis comparison (WGAC) pipeline, we predicted

49 Mbp of SDs in BosGru_PB_v1.0. In contrast, WGAC disclosed only 45 Mbp of SDs for BosGru_v2.0. Although the BosGru_PB_v1.0 PacBio-based genome had only 4 Mbp more SDs than BosGru_v2.0, its SDs were relatively longer more continuous and complete than those in BosGru_v2.0 (Figure S11). The 827 genes in the SD regions had >80% coverage. Of these, 617 showed accredited transcription and homolog support. The SD-related genes were enriched in olfactory, immune and allograft rejection functions, and some were associated with disease-related pathways (Table S12). SDs facilitate yak adaptation to changes in their food sources and exposures to novel or altered infectious agents. A similar observation was reported for primates (Bailey & Eichler, 2006). Therefore, SDs are vital genome variations that enable mammals to adapt to different environments.

We compared our assembly with the recently completed haploid yak genome (Rice et al., 2020) using yak × cattle F1 hybrid and trio-binning approaches. We found that 99.53% of the maternal assembly was covered by BosGru_PB_v1.0 and the average identity

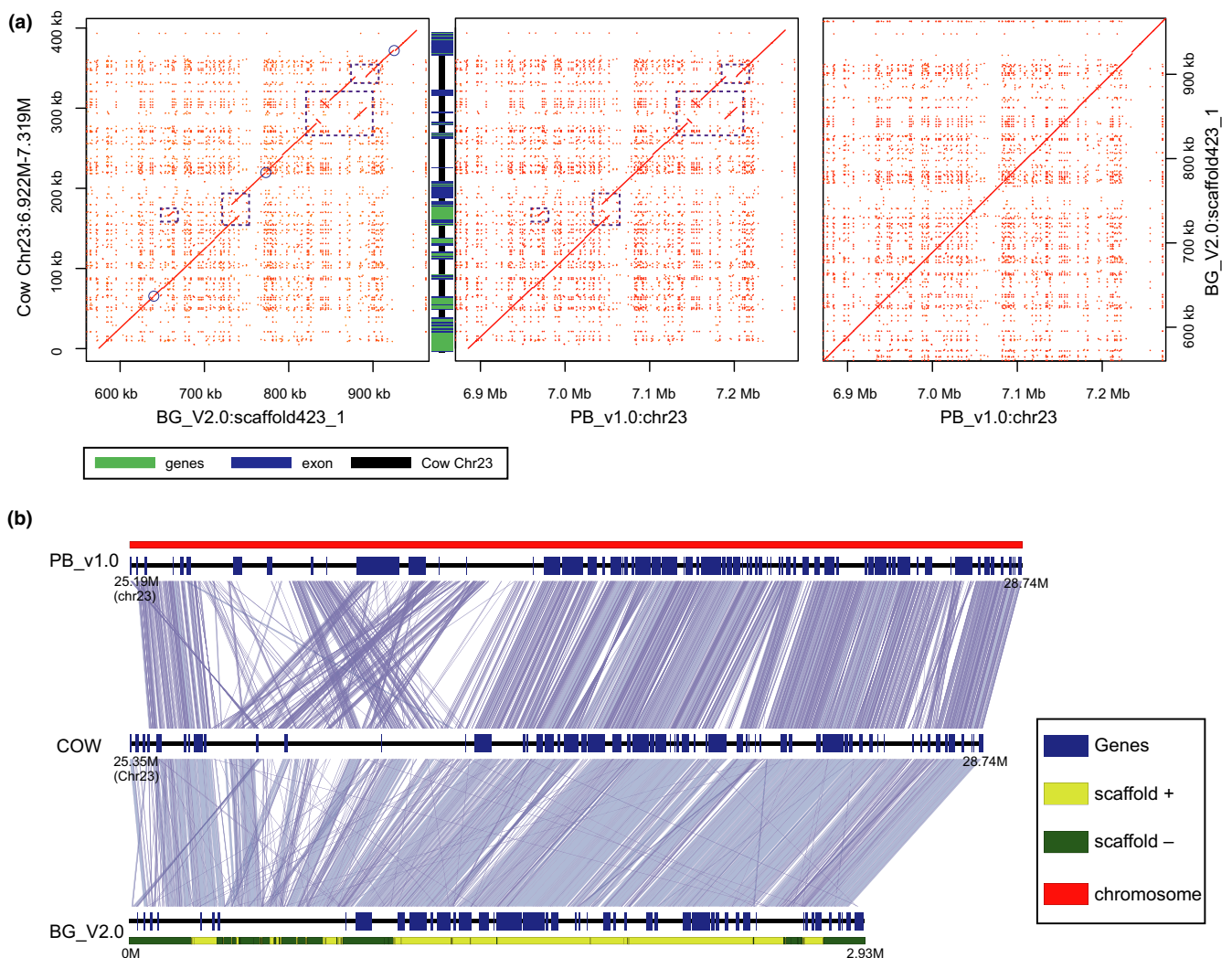


FIGURE 2 Comparative analysis of yak MHC. (a) Comparative plots of MHC-IIb sequences among taurine cattle, BosGru_v2.0, and BosGru_PB_v1.0 assemblies disclose several rearrangement sequences or discontinuities. Red rectangle highlights yak genome sequence inversions and insertions compared with taurine cattle. (b) Map of MHC class I-IIa for the three assemblies. Collinearity of taurine cattle MHC versus two yak genome MHC [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]

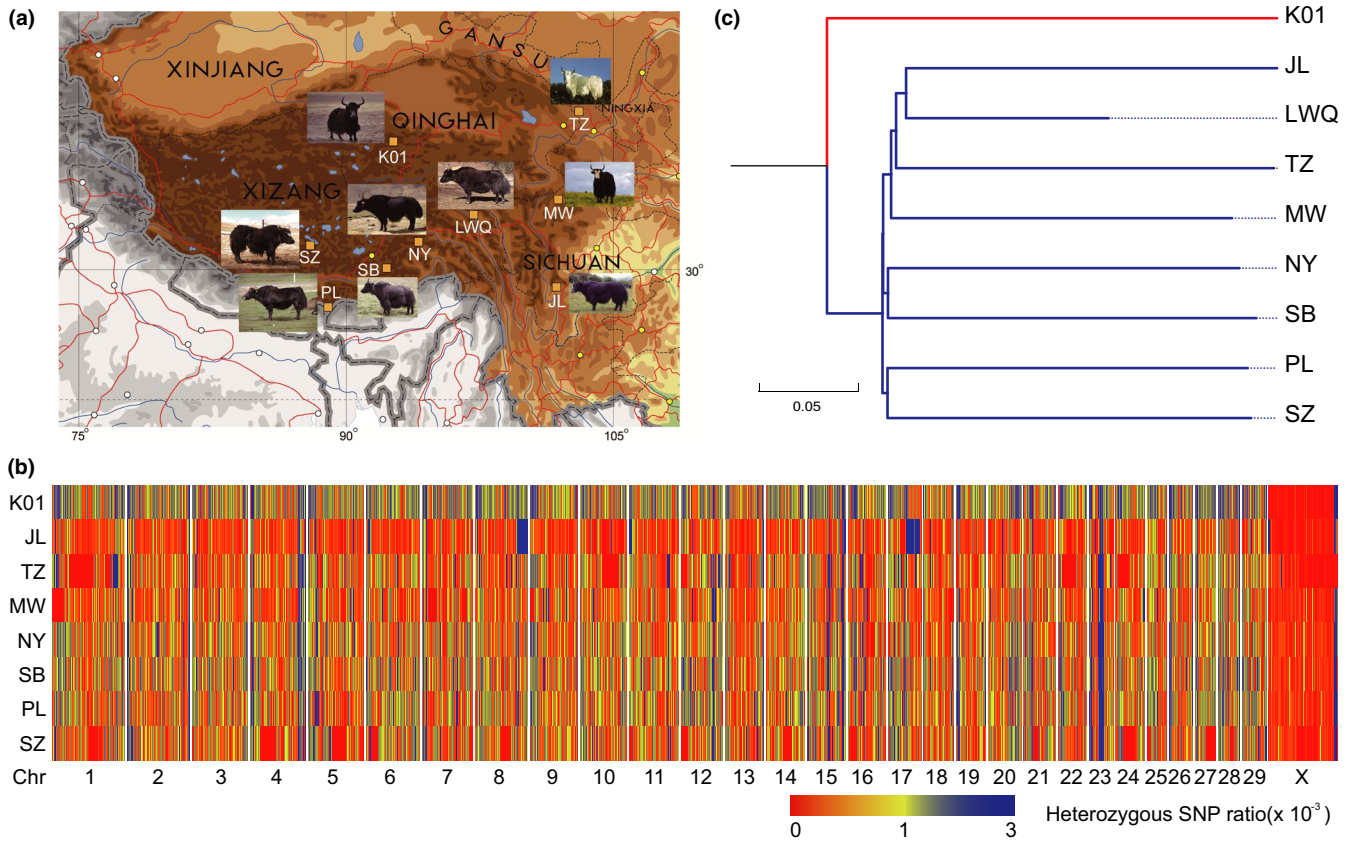


FIGURE 3 SNP callings for seven domestic and one wild yak. (a) Geographical distribution of seven domestic yak, and one wild yak used for BosGru_PB_v1.0 genome. (b) Heatmap of heterozygous SNP ratios for eight yak assemblies within each nonoverlapping 1-Mbp window across the BosGru_PB_v1.0 genome. (c) SNP-based phylogenetic tree showing relationships among all nine yak [Colour figure can be viewed at wileyonlinelibrary.com]

was 99.71%. Thus, there was strong agreement between the two assemblies. However, there were two inversions in ChrX for the assembly synteny (Figure S12). As ChrX of BosGru_PB_v1.0 was identical with UMD3.1 and UCD1.2 of cattle, we propose that our result is relatively more accurate.

3.3 | Yak-specific gene family analysis

We compared the protein-coding genes of cattle, humans, camels, sheep, goats and Tibetan antelope and built 18,176 orthologous clusters. Of these, 68% (12,388) were shared among all six mammals. Yak had a higher proportion of lineage-specific genes (3,034/22,477) than other organisms (Table S13). The same was true for Tibetan wild boar. Yak-specific gene families were significantly ($p < .01$) enriched in GO functions related to ribosome, translation, protein complex and cellular metabolic process (Table S14). Functions related to iron cations such as 'iron ion transport' (GO:0006826), 'ferric iron binding' (GO:0008199) and 'cellular iron ion homeostasis' (GO:0006879) were markedly enriched at GO levels >5. This finding may reflect the specialized adaptation of yak to the low atmospheric oxygen levels characteristic of high altitudes.

3.4 | Sequencing of one wild and seven domestic yak and characterization of their genetic variations

We used the Illumina sequencing platform and a whole-genome shotgun strategy to sequence the genomes of seven females from recognized indigenous yak breeds and that of a wild yak from Hoh Xil, Tibet (Figure 3a; Table S15). The scaffold N50 sizes of the seven domestic yak assemblies ranged from 100 to 270 kbp and their total lengths ranged from 2.56 to 2.65 Gbp. The scaffold N50 length of the wild yak was 28 kbp (Table S16). The assemblies were validated by whole-genome alignment to the BosGru_PB_v1.0 genome using mummer (v3.9.4alpha; Kurtz et al., 2004). The identity ranged from 99.31% to 99.66% (Table S18).

The protein-coding genes for each genome were annotated by integrating homology searches, mRNA expression evidence and ab initio prediction. There were 19,259–20,614 genes annotated for domesticated yak but 17,265 genes annotated for the wild yak (Figure S13, Table S17). In the latter case, the assembly may have been relatively fragmented.

By aligning the paired-end reads of each breed to the BosGru_PB_v1.0 genome with BWA and GATK, we identified 4.20–4.85 million SNPs in the seven domestic yak and 5.63 million SNPs in the wild yak

(Table 2; Figures S14 and S15, Table S19). We found only five regions of homozygosity (ROH) with a total size of 1.27 Mbp in the wild yak genome. The number and total length of these ROH were the lowest off all eight assemblies (Table 2; Figure S16). Although the wild yak underwent genetic bottlenecks, its diversity was still higher than that of domestic yak and it had the highest density of heterozygous SNPs (Figure 3b). In contrast, the JL yak presented with the lowest heterozygous SNP density, possibly as a consequence of a severe Rinderpest outbreak 150 y ago (Wiener et al., 2003). The JL samples generally had the fewest heterozygous SNPs but presented with two peaks clearly distinct from the other species (Figure 3b) at the end of chromosomes 8 (90–113 Mbp) and 17 (42–72 Mbp). There were 434 genes (126 in chromosome 8 and 308 in chromosome 17) affected by the heterozygous SNPs in the exons. The GO function enrichment analysis (Figure S17) indicated that certain genes were associated with immunity and disease.

We constructed a phylogenetic tree using 8,194,406 high-quality SNPs from the BosGru_PB_v1.0 genome and the eight yak genome assemblies (Figure 3c). The topology of the phylogenetic tree mirrored the sampling locations. The breeds in closest geographic

proximity had the shortest genetic distances between them. After separation from wild yak, the domestic yaks were segregated into three lineages: one for PL and SZ in eastern Tibet, one for SB and NY, and one for JL, LWQ, TZ and MW. These finding corroborated those reported for a previous study based on yak mitochondrial genomes (Wang et al., 2010). As domestication occurred in the Eastern Qinghai-Tibetan Plateau and Northern Tibetan areas and thence the yaks migrated in three different directions (Figure 4).

We combined assembly-versus-assembly alignment and short-read mapping and identified breed/species-specific sequences absent from BosGru_PB_v1.0 genome. These included DOCK2, RFTN1, ZFP37, HA2DMHCII and LIRA5 (immune response), as well as N2DL1, ASTN2, PTN1, PTPRA and ASTN2 (neuromodulation).

3.5 | Introgression from Tibetan cattle to one wild and seven domestic yak breeds

We combined the genomes of one wild and seven domestic yak breeds with those previously published for Tibetan cattle to identify introgressions. We used ChromoPainter (Copenhaver et al., 2012) to localize introgressed genome tracts and plotted phylogenetic trees for all individuals in each genomic segment to match the genetic introgression signals. There was a wide range in the total introgression size for the seven domesticated yaks (Table 2). There were relatively more introgression regions in the breeds indigenous to the region east of the Qinghai-Tibet Plateau than there were in those native to Tibet. Hence, Tibetan cattle distribution and introgression occurred after domestication. The JL breed had the most introgression regions, mainly distribute on chromosome 8 and chromosome 17, which is identical with heterozygous SNP peaks, and they were distributed mainly on chromosomes 8 and 17. This finding aligned with that for the heterozygous SNP peaks.

TABLE 2 SNPs, ROH and introgressed lengths for seven domestic and one wild yak

Sample	Total SNP	Number of ROH	Total length of ROH (Mb)	Introgressed length (Mb)
JL	4,205,242	39	10.072	39.183
NY	4,659,962	21	5.560	11.774
K01	5,626,934	5	1.268	5.054
MW	4,234,076	42	10.276	14.123
PL	4,795,549	30	7.521	11.783
SB	4,854,301	22	5.329	13.754
SZ	4,657,495	73	17.502	13.587
TZ	4,566,618	100	26.194	24.909

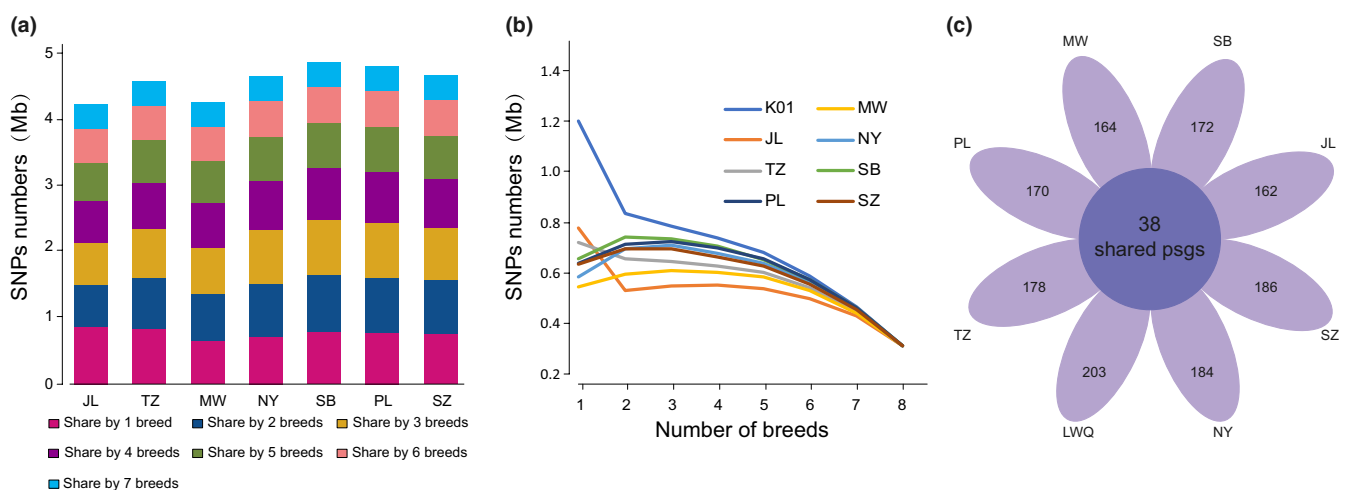


FIGURE 4 SNP distinction and positive selection in yak. (a) SNPs shared by seven domestic yak. (b) SNP sharing model for seven domestic and one wild yak. (c) Venn diagram of PSG numbers showing $k_a/k_s > 1$ between orthologous gene pairs for each domestic yak and K01 wild yak K01 [Colour figure can be viewed at wileyonlinelibrary.com]

4 | DISCUSSION

In the present study, we sequenced the yak genome using third generation PacBio long-read sequencing technology and combined this method with chromatin interaction, BioNano, and short-read sequence techniques to generate a high-quality yak genome assembly independent of existing reference genomes. This novel assembly yielded a more complete picture of the yak genome than its predecessor constructed from second-generation Illumina short-read sequences. Our BosGru_PB_v1.0 genome had a relatively low error rate and a comparatively high proportion of complete sequences.

We detected novel features related to yak adaptation to high altitudes. By comparing our newly assembled yak genome with those from other mammals, we confirmed that yak-specific gene families were significantly enriched in GO functions related to iron cations. Iron is an essential element for blood production and most of it resides in haemoglobin and myoglobin. Haemoglobin plays a vital role in transferring oxygen from the lungs to the tissues and cells whereas myoglobin accepts, stores, transports and releases oxygen.

To elucidate genetic diversity among the various officially recognized yak breeds, we sequenced and assembled seven other domestic varieties as well as a wild yak. The genome of the latter was used to clarify yak domestication. Most of the identified SNPs were shared among breeds. There was relatively low diversity among strains as yak have a short domestication history. Mutation and ROH analyses disclosed that wild yak has higher diversity than domestic yak.

The JL variety presented with a bottleneck effect caused by Rinderpest 150 y ago. Introgression increased JL diversity and helped it break through the bottleneck effect. Previous studies (Medugorac et al., 2017) reported that genes associated with coat coloration were under the strongest selection during cattle domestication. The TZ yak are famous for their white coat colour and had the second greatest number of introgression regions of all breeds studied here. Hence, the coat colour of TZ yak may be associated with introgression.

The extent to which the observed differences among yak varieties are genetically based remains to be determined. It is unknown whether any of these variations are attributable to the relative differences in the environmental conditions of the regions to which these yak breeds are indigenous. Nevertheless, the present study generated abundant genomic data that could stimulate and facilitate research on yak domestication, breeding and breed identification and clarify the relationships between genotype and phenotype.

AUTHORS' CONTRIBUTIONS

Z.J., J.Q., X.J., C.Z., Z.C., P.W. and H.J. planned and coordinated the study and wrote the manuscript. Dawayangla and Luosang performed long-read sequencing and the initial long-read assembly. J.Q., Z.Q. and Pingcuozhangdui designed the Hi-C experiments and produced assembly scaffolds from the data. Z.Y. and C.H. performed

downstream analysis of the data and assisted in the generation of additional files for the manuscript. All authors read and approved the final manuscript.

DATA AVAILABILITY STATEMENT

The assembly and raw data have been deposited to NCBI. The chromosome-scale reference genome-related data are under bioproject PRJNA507936, the data related to seven breeds are under bioproject PRJNA508860, and the wild yak-related data are under PRJNA508864.

ORCID

Jin-cheng Zhong  <https://orcid.org/0000-0002-1734-5887>

REFERENCES

- Adelson, D. L., Raison, J. M., & Edgar, R. C. (2009). Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *Proceedings of the National Academy of Sciences USA*, 106, 12855–12860. <https://doi.org/10.1073/pnas.0901282106>
- Alkan, C., Sajjadian, S., & Eichler, E. E. (2010). Limitations of next-generation genome sequence assembly. *Nature Methods*, 8, 61–65. <https://doi.org/10.1038/nmeth.1527>
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- Bailey, J. A., & Eichler, E. E. (2006). Primate segmental duplications: Crucibles of evolution, diversity and disease. *Nature Reviews Genetics*, 7, 552–564. <https://doi.org/10.1038/nrg1895>
- Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28, 45–48. <https://doi.org/10.1093/nar/28.1.45>
- Ballingall, K. T., Fardoe, K., & McKeever, D. J. (2008). Genomic organization and allelic diversity within coding and non-coding regions of the Ovar-DRB1 locus. *Immunogenetics*, 60, 95–103. <https://doi.org/10.1007/s00251-008-0278-2>
- Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6, 11. <https://doi.org/10.1186/s13100-015-0041-9>
- Belov, K., Deakin, J. E., Papenfuss, A. T., Baker, M. L., Melman, S. D., Siddle, H. V., ... Miller, R. D. (2006). Reconstructing an ancestral mammalian immune supercomplex from a marsupial major histocompatibility complex. *PLoS Biology*, 4, e46. <https://doi.org/10.1371/journal.pbio.0040046>
- Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, 27, 573–580. <https://doi.org/10.1093/nar/27.2.573>
- Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., ... Smith, T. P. L. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*, 49, 643–650. <https://doi.org/10.1038/ng.3802>
- Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and genomewise. *Genome Research*, 14, 988–995. <https://doi.org/10.1101/gr.1865504>
- Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268, 78–94. <https://doi.org/10.1006/jmbi.1997.0951>
- Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., & Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, 31, 1119–1125. <https://doi.org/10.1038/nbt.2727>
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, 10, 421. <https://doi.org/10.1186/1471-2105-10-421>

- Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., ... Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, *18*, 188–196. <https://doi.org/10.1101/gr.6743907>
- Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., ... Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, *13*, 1050–1054. <https://doi.org/10.1038/nmeth.4035>
- Copenhaver, G. P., Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics*, *8*, e1002453.
- Delaneau, O., Marchini, J., & Zagury, J. F. (2011). A linear complexity phasing method for thousands of genomes. *Nature Methods*, *9*, 179–181. <https://doi.org/10.1038/nmeth.1785>
- Deverson, E. V., Wright, H., Watson, S., Ballingall, K., Huskisson, N., Diamond, A. G., & Howard, J. C. (1991). Class II major histocompatibility complex genes of the sheep. *Animal Genetics*, *22*, 211–225. <https://doi.org/10.1111/j.1365-2052.1991.tb00671.x>
- Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., ... Aiden, L. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*, 92–95.
- Emanuel, B. S., & Shaikh, T. H. (2001). Segmental duplications: An 'expanding' role in genomic instability and disease. *Nature Reviews Genetics*, *2*, 791–800. <https://doi.org/10.1038/35093500>
- Escayg, A. P., Hickford, J. G. H., Bullock, D. W., Montgomery, G. W., & Dodds, K. G. (2009). Polymorphism at the ovine major histocompatibility complex class II loci. *Animal Genetics*, *27*, 305–312. <https://doi.org/10.1111/j.1365-2052.1996.tb00974.x>
- Gao, J., Liu, K. A., Liu, H., Blair, H. T., Li, G., Chen, C., ... Ma, R. Z. (2010). A complete DNA sequence map of the ovine major histocompatibility complex. *BMC Genomics*, *11*, 466. <https://doi.org/10.1186/1471-2164-11-466>
- Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. In arXiv. <https://arxiv.org/abs/1207.3907v2>
- Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*, 27–30. <https://doi.org/10.1093/nar/28.1.27>
- Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, *5*, R12.
- Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*, R84. <https://doi.org/10.1186/gb-2014-15-6-r84>
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*, 2987–2993. <https://doi.org/10.1093/bioinformatics/btr509>
- Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*, 589–595. <https://doi.org/10.1093/bioinformatics/btp698>
- Li, L., Stoeckert, C. J. Jr., & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*, 2178–2189. <https://doi.org/10.1101/gr.1224503>
- Majoros, W. H., Pertea, M., & Salzberg, S. L. (2004). TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics*, *20*, 2878–2879. <https://doi.org/10.1093/bioinformatics/bth315>
- Medugorac, I., Graf, A., Grohs, C., Rothhammer, S., Zagdsuren, Y., Gladyr, E., ... Capitan, A. (2017). Whole-genome analysis of introgressive hybridization and characterization of the bovine legacy of Mongolian yaks. *Nature Genetics*, *49*, 470–475. <https://doi.org/10.1038/ng.3775>
- Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics*, *21*, i351–i358. <https://doi.org/10.1093/bioinformatics/bti1018>
- Qiu, Q., Wang, L., Wang, K., Yang, Y., Ma, T., Wang, Z., ... Liu, J. (2015). Yak whole-genome resequencing reveals domestication signatures and prehistoric population expansions. *Nature Communications*, *6*, 10283. <https://doi.org/10.1038/ncomms10283>
- Qiu, Q., Zhang, G., Ma, T., Qian, W., Wang, J., Ye, Z., ... Liu, J. (2012). The yak genome and adaptation to life at high altitude. *Nature Genetics*, *44*, 946–949. <https://doi.org/10.1038/ng.2343>
- Rice, E. S., Koren, S., Rhie, A., Heaton, M. P., Kalbfleisch, T. S., Hardy, T., ... Smith, T. P. L. (2020). Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. *Gigascience*, *9*. <https://doi.org/10.1093/gigascience/giaa029>
- Scott, P. C., Choi, C. L., & Brandon, M. R. (1987). Genetic organization of the ovine MHC class II region. *Immunogenetics*, *25*, 116–122. <https://doi.org/10.1007/BF00364277>
- Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*, 3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Smit, A. F., Hubley, R., & Green, P. (1996) *RepeatMasker Open-3.0*. <http://repeatmasker.org>
- Snibson, K. J., Maddox, J. F., Fabb, S. A., & Brandon, M. R. (1998). Allelic variation of ovine MHC class II DQA1 and DQA2 genes. *Animal Genetics*, *29*, 356–362. <https://doi.org/10.1046/j.1365-2052.1998.295351.x>
- Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Research*, *34*, W435–W439. <https://doi.org/10.1093/nar/gkl200>
- Taylor, J. S., & Raes, J. (2004). Duplication and divergence: The evolution of new genes and old ideas. *Annual Review of Genetics*, *38*, 615–643. <https://doi.org/10.1146/annurev.genet.38.072902.092831>
- van der Poel, J. J., Groenen, M. A., Dijkhof, R. J., Ruyter, D., & Giphart, M. J. (1990). The nucleotide sequence of the bovine MHC class II alpha genes: DRB, DOA, and DYA. *Immunogenetics*, *31*, 29–36.
- Vezi, F., Narzisi, G., & Mishra, B. (2012). Reevaluating assembly evaluations with feature response curves: GAGE and assemblathons. *PLoS One*, *7*, e52210. <https://doi.org/10.1371/journal.pone.0052210>
- Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., ... Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, *9*, e112963. <https://doi.org/10.1371/journal.pone.0112963>
- Wang, C., Liu, C., Roqueiro, D., Grimm, D., Schwab, R., Becker, C., ... Weigel, D. (2015). Genome-wide analysis of local chromatin packing in *Arabidopsis thaliana*. *Genome Research*, *25*, 246–256.
- Wang, Z., Shen, X., Liu, B., Su, J., Yonezawa, T., Yu, Y., ... Liu, J. (2010). Phylogeographical analyses of domestic and wild yaks based on mitochondrial DNA: New data and reappraisal. *Journal of Biogeography*, *37*, 2332–2344. <https://doi.org/10.1111/j.1365-2699.2010.02379.x>
- Wiener, G., JianLin, H., & Ruijun, L. (2003). *The Yak Second Edition*. Bangkok: FAO Regional Office for Asia and the Pacific.
- Xu, Z., & Wang, H. (2007). LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, *35*, W265–W268. <https://doi.org/10.1093/nar/gkm286>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Ji Q-M, Xin J-W, Chai Z-X, et al. A chromosome-scale reference genome and genome-wide genetic variations elucidate adaptation in yak. *Mol Ecol Resour*. 2021;21:201–211. <https://doi.org/10.1111/1755-0998.13236>