

The first complete T2T Assemblies of Cattle and Sheep Y-Chromosomes uncover remarkable divergence in structure and gene content

Timothy Smith

tim.smith2@usda.gov

USDA, ARS, U.S. Meat Animal Research Center (USMARC) <https://orcid.org/0000-0003-1611-6828>

Temitayo Olagunju

University of Idaho <https://orcid.org/0000-0002-1497-6173>

Benjamin Rosen

Agricultural Research Service USDA <https://orcid.org/0000-0001-9395-8346>

Holly Neibergs

Washington State University

Gabrielle Becker

University of Idaho <https://orcid.org/0000-0002-1455-6443>

Kimberly Davenport

Washington State University

Christine Elsik

University of Missouri

Tracy Hadfield

Utah State University

Sergey Koren

Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health
<https://orcid.org/0000-0002-1472-8962>

Kristen Kuhn

Agricultural Research Service USDA

Arang Rhie

Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA <https://orcid.org/0000-0002-9809-8127>

Katie Shira

University of Idaho

Amy Skibiel

University of Idaho

Morgan Stegemiller

University of Idaho

Jacob Thorne

Texas A&M

Patricia Villamediana

South Dakota State University

Noelle Cockett

Utah State University



Brenda Murdoch
University of Idaho

Article

Keywords:

Posted Date: April 3rd, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-4033388/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: Yes there is potential Competing Interest. SK has received travel funds to speak at events hosted by Oxford Nanopore Technologies.

Abstract

Reference genomes of cattle and sheep have lacked contiguous assemblies of the sex-determining Y chromosome. We assembled complete and gapless telomere to telomere (T2T) Y chromosomes for these species. The pseudo-autosomal regions were similar in length, but the total chromosome size was substantially different, with the cattle Y more than twice the length of the sheep Y. The length disparity was accounted for by expanded ampliconic region in cattle. The genic amplification in cattle contrasts with pseudogenization in sheep suggesting opposite evolutionary mechanisms since their divergence 18MYA. The centromeres also differed dramatically despite the close relationship between these species at the overall genome sequence level. These Y chromosome have been added to the current reference assemblies in GenBank opening new opportunities for the study of evolution and variation while supporting efforts to improve sustainability in these important livestock species that generally use sire-driven genetic improvement strategies.

INTRODUCTION

The suppression of recombination between the mammalian X- and Y-chromosomes^{1,2} outside the pseudo autosomal region (PAR) followed their separation from autosomes about 190MYA³. The X-chromosome gene content was maintained while the Y-chromosome rapidly lost genetic content^{4,5} while accumulating duplicated DNA elements and repeats. The loss of genes on the Y-chromosome was followed by the acquisition of male-specific genes which are critical for sex determination of an individual and play vital roles in spermatogenesis and male fertility⁶⁻¹⁰.

Obtaining a complete assembly of mammalian Y-chromosome has been elusive mainly due to the high repetitive DNA content and the inability of sequencing technologies and assembly tools to sufficiently tackle the challenges presented by the structure of this sex-determining chromosome. Very few mammals, including the *Bovidae* family to which cattle and sheep belong, have had a Y-chromosome assembly to date, and no member of the *Bovidae* family has a complete Y-chromosome. The majority of previous attempts to characterize the Y-chromosomes of the *Bovidae* family have been based on fluorescence *in situ* hybridization (FISH)¹¹. Relatively non-contiguous 40 megabase² (Mb) and 43.3Mb (NCBI accession GCA_000003205.6) assemblies of the cattle Y chromosome² have been produced from bacteria artificial chromosome (BAC) clones, in addition to a 16Mb Y chromosome assembled in 67 contigs¹² from long reads sequencing, out of the estimated 50Mb size¹³. A 10.8Mb male-specific Y (or MSY) assembly comprised of 50 contigs alongside 4.11Mb of the PAR has been produced for sheep¹⁴.

The first T2T assembly of the complete human genome¹⁵ did not have a complete Y-chromosome assembly due to the use of a cell line lacking a Y, although an assembly for this chromosome from another source was recently added¹⁵. The successes recorded in various T2T chromosome assemblies^{16,17} have been made possible by recent advances in long read sequencing technologies¹⁸ complemented by improvements in genome assembly algorithms¹⁹, effectively bridging the technological gap that previously hindered successful sequencing and assembly of Y-chromosomes. Furthermore, the use of parental data in genome assembly introduced with the trio-binning method²⁰ has been an invaluable technique to produce fully phased haplotype-resolved assemblies of diploid species.

We present a major leap forward for livestock assemblies with the first complete and gapless T2T Y-chromosomes of cattle and sheep. These sex chromosomes were obtained from haplotype-resolved whole-genome assemblies based on a combination of Illumina short reads²¹ and Pacific Biosciences²² and Oxford Nanopore Technology²³ long read sequencing technologies. We present a detailed structural analysis of these chromosomes highlighting novel features in hitherto hard to reach regions, and further elucidate the similarities and differences between them. These complete Y-chromosomes of cattle and sheep provide important resources for studying ruminant biology and mammals by extension. By interrogating these T2T Y-chromosomes we can begin to address salient long-standing biological questions around the structure and evolution of the Y-chromosomes of these two members of the *Bovidae* family.

RESULTS

Whole genome assemblies of cattle and sheep

The T2T Y-chromosome assemblies of cattle and sheep were obtained from draft versions of haplotype-phased whole genome draft assemblies (in progress) of the F1 individuals from the Sire_x_Dam crosses of Wagyu_x_Charolais cattle and Churro_x_Friesian sheep breeds respectively, using the Verkko¹⁹ assembler and parental data for phasing. The cattle Y-chromosome thus represents a Wagyu haplotype while the sheep Y-chromosome is from a Churro. The combination of the ONT ultra-long reads with the PacBio Hifi reads successfully resolved the highly repetitive telomeres, centromeric and heterochromatic regions of the Y-chromosomes. The Y-containing paternal haplotype assemblies were highly contiguous with contig NG50 of 96.68Mb (cattle) and 108.17Mb (sheep) (Supplementary Table S1). The reference-agnostic genome assembly completeness evaluation with Merqury²⁴ revealed high rate of k-mer survival of the raw reads in the final assemblies (Supplementary Figure S2) with 53.6-56.13 QV scores, while 98.11-99.78% complete BUSCOs were discovered from the *Cetatiodyctyla_ODB10* orthologous genes database (Supplementary Table S1). Visualization of the haplotype-colored assembly graphs with Bandage²⁵ showed that some of the autosomes were not gapless, but the cattle and the sheep Y-chromosomes were in single contigs (Supplementary Figure S1). The Y chromosome contigs were extracted from the paternal haplotype assemblies for in depth analysis. Merqury²⁴ QV scores for the cattle and sheep Y-chromosome contigs were higher than their paternal haplotype averages at 62.38 and 59.95 respectively (Supplementary Table S1). Telomere sequences were located at the distal ends of the two chromosomes – 14.6kb and 20.3kb from the p- and q- arms of the cattle and 19.2kb and 17.7kb from the p- and q- arms of the sheep, (Table 1, Figure 1A) indicating the completeness of the single-contig assemblies.

The cattle and sheep Y-chromosomes have substantial differences in structure

The total lengths of the complete cattle and sheep T2T Y-chromosome assemblies were substantially different at 59.4Mb and 25.9Mb, respectively. The 120.76kb sheep Y-chromosome centromere at 8.03Mb from the end of the short arm is 15 times smaller than the 2.52Mb-long centromere on the cattle located at 14.12Mb from the end of the shorter arm of the chromosome (Table 1, Figure 1A). Approximately half (50.54% for cattle and 48.55% for sheep) of the bases on the two chromosomes were annotated as repetitive DNA (Figure 1B, Supplementary Table S2), and the q-arms comprised a long region of a mosaic of high similarity repetitive DNA arrayed in either direct or inverted orientation to one another (Figure 1A). The distribution of the repeat elements indicates that LINE elements had the highest and similar proportion (about 31%) on both Y-chromosomes (Figure 1B, Supplementary Table S2). The proportion of other classes of repetitive DNA were higher in sheep except for satellites and simple repeats (low complexity region) (Figure 1B, Supplementary Table S2). This observation agrees with previous reports that LINE elements are the dominant retrotransposons in mammals^{26,27}.

The PAR on the Y-chromosome assemblies were identified by mapping long reads from female haplotypes described in²⁸ with further details in the methods section. The alignment track visualized with IGV²⁹ clearly defined the PAR boundaries with soft clipping of the long reads at the region lacking homology with the Y-chromosome, rendered as the highly colored reads segment coupled with a drastic drop of the coverage to zero on the coverage track (Supplementary Figure S9). The sheep Y-chromosome PAR (7,018,329bp) was 195.9kb longer than the cattle PAR (6,822,380bp) (Figure 1A, Table 1) despite having much lower total chromosome length. The rest of the Y-chromosome adjacent to and outside the PAR is the MSY region comprising the gene-rich euchromatin and gene-deficient heterochromatin of the chromosome (Figure 1A). This region is sub-classified into the X-degenerate region harboring the genes that ancestrally recombined with the X-chromosome, and the ampliconic region containing intrachromosomal duplication of genes which are expressed mainly in the testis³⁰.

The Y-chromosome assemblies were submitted to NCBI and annotated with their Eukaryotic Genome Annotation Pipeline (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/) for use in downstream analyses. The number of protein-coding genes on the cattle Y-chromosome was more than 3 fold higher than that found on the sheep Y-chromosome (352 to

109) (Supplementary Tables S5 and S6). However, more pseudogenes (150 to 79) were annotated on the sheep Y compared to the cattle. This difference in the number of annotated genes, with an approximate 1:2 ratio is proportionate to the chromosome sizes. All the previously reported mammalian PAR genes identified through sequence analysis³¹ and physical mapping with FISH³² could be identified on each T2T assembly. The most proximal protein-coding gene in the PAR of both Y chromosomes was *PLCXD1*, located at 33.74kb on the sheep and 34.57kb on the cattle from the telomeres on the p-arms of the respective Y-chromosomes while the most distal gene was *GPR143* located about 3.50kb from the PAR boundaries of both Y-chromosomes (Supplementary Tables S5 and S6). The PAR on the sheep harbored a total of 45 protein-coding genes comprising 33 single-copy genes and four multi-copy genes (*CSF2RA:2*, *ASMTL:2*, *OBP:4* and *BOS2D:4*) (Supplementary Table S6) while 27 single-copy and three multi-copy genes (*ASMTL:2*, *OBP:3* and *BOS2D:2*), a total of 34 protein-coding genes, were located on the cattle PAR (Supplementary Table S5). These included an uncharacterized gene *LOC112445918* (Supplementary Tables 5 and 7) in the cattle Y PAR, and copies of Splicing Factor 3A Subunit 2 (*SF3A2*) and Proline Rich Protein BstNI Subfamily 1 and 3 (*PRB1* and *PRB3*) (Supplementary Table 6) on the sheep Y PAR which were not in the PAR of their respective X-chromosomes (Supplementary Table 7). The overall conservation of the mammalian PAR genes on both cattle and sheep Y-chromosomes is expected since crossing over still occurs between the X- and Y-chromosomes.

The MSY contains the sex-determining gene *SRY* and is further divided into sub-regions named the X-degenerate (X-d) and ampliconic regions. The sex-determining gene, *SRY* is located at 58,122,906bp and 17,199,770bp on the cattle and sheep Y-chromosomes, respectively (Supplementary Tables S5 and S6). The X-d region contains genes that still maintain some level of homology with the X-chromosome outside the PAR. The cattle X-d (spanning about 1.9Mb at 6.8Mbp and 23.0Mbp from the p-arm) contained 15 protein-coding genes with 3 having pseudogenes (*EIF1AX:1*, *UBA:2*, *SHROOM2:3*) for a total of 27 pseudogenes (Supplementary Table S5). Out of the 18 protein-coding genes in the sheep X-d region (about 1.7Mb long at 7.0Mbp from the p-arm), only *USP9X* is multi-copy with just 2 copies; *EIF1AX* is the only protein-coding gene having pseudogenes with 4 copies out of the total 23 pseudogenes in this region (Supplementary Table S6).

The putative ampliconic gene families were extracted from the NCBI annotations including *RBMY*, *HSFY*, and *TSPY* that are conserved across mammals³³ as well as the bovine specific genes *PRAME*, *ZNF280A* and *ZNF280B*^{34,35}. There were approximately 4-fold more cattle protein-coding ampliconic genes (187) compared to the sheep (46) (Figure 2A, Supplementary Table S8). However, like the overall observation on the gene content, the pseudogenes on the sheep (127) were more than double the number on the cattle (50) (Supplementary Table S8). These figures highlight significant divergence in the copy number of the ampliconic genes between cattle and sheep. Bearing in mind that the size and content of the PAR of the two chromosomes are similar (Figure 1A, Table1), and the number of X-degenerate genes are also similar (Supplementary Tables S5 and S6), the ampliconic genes, occupying an approximate 40MB block of DNA on the cattle compared to about 17MB on the sheep, appear to be responsible for the large chromosome size difference.

Ampliconic genes are important MSY genes which are required for spermatogenesis and fertility³⁶⁻³⁸. Variation in their copy number within a population has been previously reported in cattle for *TSPY*^{39,40}, *PRAME*⁴¹ and *ZNF280A*^{42,43}, and copy number has been implicated in male fertility^{41,44,45}. Ampliconic genes were present mainly in tandem arrays on the cattle, but not on the sheep. For instance, *HSFY*, *RBMY*, *PRAME* and *TSPY* were arranged in tandem on the same strand on the cattle, whereas only 1 copy each of *RBMY* and *PRAME* were on the sheep (Figure 2B and Supplementary Figure S11). An island of 44 protein-coding copies (out of the total 84 copies) of *TSPY* is located on the p-arm of the cattle spanning 1.21Mb while the rest of the copies are distributed on the ampliconic region of the chromosome (Supplementary Figure S11A). *TSPY* has been reported as the largest tandem protein-coding array present on the human genome^{33,46,47} and the active array constitute the active copies, but they have been more amplified on cattle relative to human⁴³.

Pseudogenization of ampliconic genes suggests an evolutionary process on the Y-chromosomes

A lower ratio of protein-coding genes to pseudogenes in the ampliconic gene family was observed on the sheep (0.07 to 0.85) compared to the cattle Y-chromosome (0.66 to 31) (Figure 2A, Supplementary Table S8). The *TSPY* gene family had

the lowest ratio (0.07) in sheep. Only *ZNF280B* ampliconic gene had more protein-coding copies (1.80 ratio) than pseudogenes (Figure 2B, Supplementary Table S8). With no protein-coding copy of *ZNF280A* annotated on either Y-chromosome, 15 and 22 copies of the pseudogene were present on the cattle and the sheep respectively. In contrast, no *RBMY* pseudogenes were found on either chromosome (Figure 2B, Supplementary Table S8 and Supplementary Figure S11).

A previous study on the evolution of the bovine *MSY* genes, based on the expansion of the ampliconic genes⁴³, reported that the amplification of the *TSPY* and *ZNF280B* genes predated the amplification of *HSFY* genes around the same time that sheep diverged from cattle⁴⁸. Since the amplification of *TSPY* occurred before the divergence of the two species, it is therefore expected that a relatively high copy number similar to the number found on the cattle (98) should also be found on the sheep. However, the preponderance of pseudogenes (71) to protein-coding copies (5) (Figure 2A & 2B, Supplementary Table S8) suggests the decay of their protein-coding copies on the sheep Y in line with chromosome evolution theory of gene loss through pseudogenization³³. This may also explain the observed higher pseudogene copies of *HSFY* and *PRAME* than the protein-coding ones (Figure 2A & 2B, Supplementary Table S8). In contrast to the other ampliconic gene family members, *ZNF280B* still maintained more protein-coding copies than pseudogenes (Figure 2A & 2B, Supplementary Table S8).

A cross-species phylogenetic comparison of the protein-coding copies of the ampliconic gene families revealed a general high intra-species clustering between the genes with clearly separated clades for the cattle and the sheep copies (Supplementary Figure 10). The protein-coding copies of *ZNF280B* did not show a clear separation from the ancestral autosomal copy³⁵ (Supplementary Figure 10A), while the higher copies on sheep may have resulted from interlocus gene conversion events. In contrast, the *HSFY* and *TSPY* copies on cattle and sheep branched at the human copy-rooted trees (Supplementary Figures 10B & 10C). The fewer protein-coding copies on sheep compared to cattle is suggestive of loss of the ancestral copies on sheep in response to evolutionary pressures since sheep diverged from cattle⁴⁸. *RBMY* and *PRAME* were not included in this cross-species phylogenetic comparison since the sheep Y had just one copy of each.

The transcriptional activity of the ampliconic gene families was investigated with transcript reads from RNA-Seq experiments of different tissues and developmental stages (Supplementary Tables S9 and S10) and indicated that most of the protein-coding copies of these genes had no evidence of expression. On the cattle Y-chromosome, *TSPY* with the highest protein-coding copies (84) had only 7 copies with observed transcripts while a single copy of *TSPY3* showed considerable levels of transcriptional activity. Only 1 out of the 11 copies in the *RBMY* array had evidence for transcriptional activity (Supplementary Table S9). Similarly, 18 protein-coding copies of the total 31 for *PRAME* recorded reads while only 9 copies were at substantial levels. *ZNF280B* was the only gene with transcriptional activity across all the tissues analyzed with transcripts corresponding to 14 out of 23 gene copies and only 3 of these displaying relatively high transcriptional activity (Supplementary Table S9). It is noteworthy that the ancestral copies of the ampliconic genes showed less transcriptional activity compared to the bovine specific *PRAME* and *ZNF280B*. The significance of this is not understood yet but will require further studies.

All the copies of the protein-coding ampliconic genes on the sheep generally exhibited high levels of transcriptional activity except 2 of the 27 copies of *ZNF280B* which registered no reads across all the samples analyzed (Supplementary Table S10). These results indicate that only 10% of the protein-coding copies of the ampliconic genes on the cattle registered transcriptional activity in sharp contrast to 91% on the sheep. Low gene expression levels of ampliconic gene families have been linked to their nonessentiality on the Y-chromosome of apes⁴⁹, evolution of new function⁵⁰ or compensation by the X-chromosome paralog⁵⁰. While there is no sufficient data in this study to draw any of these conclusions on the observed differences in the transcriptional activities of the ampliconic genes, the preliminary observation is suggestive of some ongoing evolutionary mechanism on the Y chromosomes.

Novel insights into the cattle and sheep centromere structure and organization

The physical coupling of chromosomes to spindle fibers during cell division is facilitated by assembly of the kinetochore proteins at the centromere. The centromeres of most eukaryotic chromosomes are characterized by satellite DNA repeats

which are generally involved in a higher order repeat (HOR) structure of composite elements⁵¹ as the genetic signal for kinetochore attachment. A characteristic epigenetic methylation dip at the centromeric dip region (CDR) due to the binding of the centromere protein A (CENPA) has also been observed⁵². A generally high CpG methylation within the active centromeric alpha satellite array relative to the adjacent regions is interrupted by local reduced CpG methylation dips at the CDRs⁵³.

The cattle centromere was defined by a well-organized tandem array of a HOR containing a 73bp monomeric satellite repeat unit (Figure 1C). This is the first time that this sequence has been identified in cattle. The 73bp monomeric satellite unit was arranged into a higher-order repeat (HOR) structure wherein a total of 364 copies of the 3.7kb unit spans the 2.5Mb centromeric region (Figure 3A). The individual HORs had 95-98% sequence identity between them and were organized into four segments based on the homogenization of the tandem monomeric repeat unit – the first and third segments comprised 71-98% identical 5-mer and 6-mer tandem copies of the 73bp monomers respectively, while the second and fourth segments contained more diverged copies of the monomer and non-satellite DNA (Figure 3A). There were other copies of the HOR found on the cattle outside the centromere, but they were more diverged from the copies within it (Supplementary Figure S5). Similarly, more divergent copies of the 73bp monomeric unit were evenly distributed along the entire length of the cattle MSY region (Supplementary Figure S6A). Using cytosine residue methylation status predicted from the long sequence read data produced by the ONT PromethION and PacBio Revio platforms, the CDRs supporting the annotation of the centromere locus on the cattle were revealed at the flanks of the centromeric HOR array and coincided with an enrichment of signal from the inner kinetochore CENP-A (Supplementary Figure S3).

The alignment of the cattle 73bp centromeric monomer to the whole genome assemblies (comprising the two haplotypes each) of the Churro and the Wagyu suggests that the sequence is specific to only the sex chromosomes but with much fewer and more diverged copies on the cattle and sheep X-chromosomes and the sheep Y-chromosome (Supplementary Figure S6B, Supplementary Table S11).

The sheep Y-chromosome centromere was organized differently than the cattle as a tandem array of 47 units of a 2.51kb composite structure comprising two ruminant-specific transposable elements (TEs) *BOV-A2* and *BovB* interspersed with spacer sequences (Figure 1C). The composite structure can be subdivided into three segments – a 299bp segment1, 171bp segment2 and 2,043bp segment3 (Figure 3B). Segment2 contained RepeatMasker-annotated *BOV-A2* SINE and *BovB* LINE TEs. This tandem co-location of *BOV-A2* and *BovB* is not surprising since LINEs have been reported to facilitate the transposition of SINE elements⁵⁴. A C-rich 58bp simple repeat sequence at the end of segment2 was the only other annotated repeat, segments 1 and 3 were lacking any annotation. The methylated cytosine data supported identification of the characteristic CDR which defines the site of kinetochore assembly on eukaryotic centromeres as epigenetic support for the annotation of the sheep Y-chromosome centromere (Supplementary Figure S4).

The presence of TEs within the repeat unit led us to hypothesize that the TE may have been inserted into a piece of DNA before replication and homogenization at the centromere. This hypothesis was tested by mapping the sequence surrounding the TEs to the whole genome to locate a possible origin of the repeat unit. However, there was no significant mapping and as such the origin of the repeat unit could not be ascertained in this manner.

Most of the putative bovine satellite DNA sequences have been reported to be related to one another with short segments⁵⁵ arising from their evolutionary trajectory. For instance, *SAT1.706*, *SAT1.711a* and *SAT1.720* share a 23bp sub-repeat unit while *SAT1.711b* and *SAT1.715* also share another 31bp sub repeat unit. Based on this, we sought to identify any relatedness of the cattle and sheep centromeric repeat sequences (Supplementary Table 16) to any of the putative bovine satellites by aligning them, but none was observed. This suggests that they are novel sequences and bear no evolutionary relationship to the known bovine satellites.

The T2T Y-chromosomes compared with publicly available cattle and sheep assemblies

Our T2T Y-chromosome assemblies were compared with publicly available assemblies of cattle and sheep on NCBI (<https://www.ncbi.nlm.nih.gov/>) to assess the concordance of the available assemblies with our T2T assemblies. A 43.3Mb assembly of Hereford cattle Y chromosome (BTAU5-Y) with the accession GCA_000003205.6 and a Hu sheep assembly with accession CM022046.1 from a previous study¹⁴ were obtained from NCBI.

The first 2.5Mb of the BTAU5-Y was absent and only about 100kb sequence overlapped between the two Y-chromosome assemblies at the PAR (Supplementary Figure S7A, Supplementary Table 12). The next alignment segment was 0.5Mb long and was followed by a large gap spanning 7.8Mb on the BTAU5-Y up to 16.6Mb distal to the centromere on the T2T cattle, implying an absence of the centromere on the BTAU5-Y. The lack of centromere on the BTAU5-Y was confirmed with the absence of the novel cattle centromeric HOR (with just 4 copies) despite the abundance of the 73bp monomer (Supplementary Figure S7B) across the length of the chromosome. The region spanned by this BTAU5-Y gap on the T2T cattle contains the TSPY1 array (Supplementary Table 17), a repeats-enriched 4.1Mb segment and is adjacent to the centromeric repeat (Figure 1). The next 6Mb of DNA on the T2T cattle aligned with only about 400kb of the BTAU5-Y indicating more than 5Mb of sequence harboring 33 ampliconic genes (Supplementary Table 17) is missing relative to the T2T cattle Y-chromosome (Supplementary Figure S7A, Supplementary Table 12). Alignment contiguity between the two Y-chromosomes reduced from 23.8Mb to the end of the T2T cattle within the highly repetitive ampliconic region. This comparison has accounted for more than 15Mb of sequence missing in the BTAU5-Y relative to the T2T cattle Y-chromosome.

The alignment between the Hu Sheep and the T2T sheep showed a broad sequence concordance up to about 10Mb out of the 10.6Mb of the Hu sheep MSY extending from the end of the T2T sheep PAR to 19.67Mb (Supplementary Figure S8, Supplementary Table 13). However there was a 1.85Mb region on the T2T sheep that exhibited lower alignment contiguity with an average segment length of about 150kb as well as a few sequence inversions. The rest of the Hu sheep MSY, about 600kb long, mapped to the T2T sheep in segments of about 50kb extending from 17.6Mb within the ampliconic region to its end (Supplementary Figure S8, Supplementary Table 13). These comparisons have revealed remarkable gaps in the public assemblies, mainly in the highly repetitive regions of the chromosomes, which have been filled by the T2T Y-chromosome assemblies.

DISCUSSION

The combination of long reads for contig creation and short reads for haplotype phasing has produced single contiguous high-quality Y-chromosome assemblies of two members of the *Bovidae* family – cattle and sheep – spanning the highly repetitive DNA content. Following detailed analysis, we present the structure and organization of the important genomic structures contained in these two chromosomes.

The first striking observation comparing the two chromosomes is the substantially smaller size of the sheep Y-chromosome compared to the cattle Y-chromosome. The 59.4Mb cattle Y-chromosome was more than double the 25.9Mb length of the sheep. Although previous reports from chromosome banding techniques of a smaller sheep Y chromosome compared to cattle⁵⁶ and a significant variation of Y chromosome within the *Bovini* tribe⁵⁷ have been made, our observation presents a clearer picture of this size difference between cattle and sheep. Greater variation in the size of the Y-chromosome relative to similar X-chromosome size across primates has recently been reported⁵⁸. This suggests a higher susceptibility of the Y-chromosome to changes in response to evolutionary pressures than the X-chromosome. The extent of this Y chromosome polymorphism within the *Bovidae* family would be better elucidated with the availability of more complete Y-chromosome assemblies. The relatively shorter sheep Y is still nearly twice as long as the 14.9Mb length of the longest previously reported Y-chromosome sequence of the domestic sheep¹⁴. Similarly, the T2T cattle Y-chromosome is 9.4Mb more than the previously reported 50Mb estimate¹².

Gene annotation of the T2T Y chromosomes indicates that the PAR region, which still maintains recombination with the X-chromosome, is generally well conserved between cattle and sheep in agreement with the gene content of the mammalian PAR. The uncharacterized *LOC112445918* gene located on the cattle Y-chromosome is in a region missing from the X-chromosome (Supplementary Table S7) assembly and may have been present if the assembly was complete. On the sheep Y-chromosome however, the three previously unreported genes did not have homologs on the X-chromosome PAR, nor on the autosomes.

Since the number of bases covered by the PAR on the two chromosomes are similar, and the extent of the X-degenerate regions is similar in both chromosomes, the ampliconic gene families appear to be responsible for the substantial size difference between the cattle and sheep Y chromosomes. Previous studies of the ampliconic region have reported increased tendency of drastic differences in gene content and copy number between closely related species^{36,59} and even within a population^{39-43,58}. It has previously been suggested that ampliconic genes were involved in the diversification of *Bovidae*⁴³, but the extent of their contribution to the wide divergence between the cattle and sheep Y-chromosome sizes is remarkable. Studies within a clade, such as being carried out in ruminants by the Ruminant T2T (RT2T) Consortium⁶⁰, will help to elucidate Y-chromosome polymorphism between the species in that clade.

The amplification patterns of the protein-coding ampliconic genes observed on the cattle and the pseudogenization of some copies on the sheep suggests different evolutionary mechanisms taking place within the two Y-chromosomes since the divergence of sheep from cattle about 19MYA⁴⁸. Higher copy number for genes which are associated with important traits such as fertility is correlated with higher gene expression⁶¹, however, the transcriptional activity analysis indicated that most (90%) of the ampliconic genes had no evidence of transcriptional activity. With higher copy numbers and reduced expression in a region devoid of crossing over with the X-chromosome, the amplification might be due to gene conversion as a means to conserve gene function^{33,62}. These observations are preliminary since this study was not designed to answer these questions and will require further population level studies across breeds and species to establish the evolutionary mechanism shaping the Y-chromosomes of cattle and sheep.

Analysis of our T2T Y-chromosome assemblies of cattle and sheep have provided novel insights into the content and organization of the centromeric region of the chromosomes. Bovine satellite DNAs have been well characterized with FISH especially in the centromeric and pericentromeric regions on the autosomes of cattle and sheep, but none of these satellites has been located on either of the sex chromosomes (reviewed in⁵⁵). For the first time the monomeric (73bp) satellite as well as the HOR sequence (3.7kb) which characterizes the centromeres of cattle as well as the composite repeat unit at the centromeres of sheep have been identified. The copies of the cattle 73bp cattle monomer located on the sheep Y indicate that this monomer predates the divergence of cattle and sheep. The few copies on the sheep Y which were not proximal to the centromere were highly diverged in sequence compared to the copies on the cattle Y. It is yet to be determined whether a copy of this monomer was amplified and adopted as the Y centromere in cattle or lost and degraded in sheep. This new knowledge of the structure and organization of the centromeres on the cattle and sheep Y-chromosomes is invaluable to the study of chromosome biology and evolution since centromeric satellites are regarded as the fastest evolving DNA elements^{63,64}.

The preliminary knowledge on cattle and sheep Y-chromosomes which has been revealed by these T2T assemblies have provided the foundation for further exploration of different aspects of Y-chromosome biology at population scale. In a recent study, the copy number of the *TSPY* array was reported to vary between human, which had 44 tandem copies, and non-human primates, with an average copy number of 18⁵⁸. It is yet to be determined what kind of evolutionary relationship exists between the human orthologs and the tandem copies found also on the T2T cattle Y-chromosome. Furthermore, studying the gene content on other species in the *Bovidae* family or the ruminantia sub-order⁶⁰ would enrich our knowledge of the evolution of the sex-determining chromosomes within and between lineages.

CONCLUSIONS

The first complete T2T Y-chromosome assemblies of cattle and sheep from the *Bovidae* family have provided a holistic insight into the structure and organization of these important sex-determining chromosomes. Comparison of the two assemblies revealed inter-chromosomal similarities and differences in their genetic components. Remarkable differences were noticed in the size and organization of the centromeres and the overall chromosome length. The difference in the chromosome lengths, specifically in the MSY, could be ascribed to loss of copies of ampliconic genes on the sheep Y since the cattle Y is the ancestral copy, and is suggestive of different evolutionary processes on the two chromosomes. These new assemblies which are important resources for ruminant chromosome biology have been added to the current reference assemblies of cattle and sheep on NCBI and present new opportunities to answer pertinent biological questions on the sex-determining Y-chromosome.

METHODS

Genome assembly quality control

GfaStats⁶⁵ was used to produce the assembly statistics used to evaluate the contiguity of the whole genome assemblies using default parameters. The reference-free k-mer-based genome assembly completeness evaluation tool Merqury²⁴ was run as recommended by the developers on the diploid assemblies from the F1 crosses of the cattle and sheep species to produce q-value statistics and plots to check the k-mer distribution between the raw and the final assemblies for assembly completeness assessment. Compleasm⁶⁶ (formerly miniBUSCO) was also used to evaluate assembly completeness using the number of orthologous genes identified in the draft assemblies. Compleasm was run with default parameters and lineage="cetartiodactyla_odb10".

Seqtk telo module from Seqtkv1.4 (<https://github.com/lh3/seqtk>) was used to identify the telomere sequence at the ends of the Y-chromosome assemblies. Although the program identified the telomeric repeat sequence at the ends of the Y-chromosomes, it is difficult to ascertain if the full length of the telomeres were assembled.

Defining the PAR boundaries

We defined the pseudo autosomal region (PAR) on the Y-chromosome assemblies by mapping raw PacBio long reads (HiFi for cattle and CCS for sheep) from female haplotypes to the cattle and the sheep Y-chromosome assemblies as previously described²⁸. Briefly, the PAR on the X-chromosomes of the cattle and sheep reference assemblies were hard-masked using the current annotation coordinates. Long reads from a female individual were mapped to the assemblies using minimap2⁶⁷ *minimap2 -a -t \$threads \$inFile \$inFile2 \$inFile3 \$inFile4 > \$inputTAG.sam*. The alignment file was sorted and indexed with Samtools⁶⁸ before visualization with IGV²⁹. The alignment track (Supplementary Figure S9) clearly showed the PAR boundaries with the soft clipping of the long reads at the region lacking homology with the Y-chromosome rendered as the highly colored reads segment, coupled with a drastic drop of the reads coverage to zero as shown in the coverage track (Supplementary Figure S9).

To further finely define the PAR region for the cattle and sheep sex chromosomes, we aligned the Y- and X-chromosome assemblies to each other with Mashmap⁶⁹ using the parameters segment length *-s 10000* and minimum identity *-pi 95*. For cattle, the PAR region on the Y-chromosome extended from the beginning of the chromosomes to 6,819,999bp on the Y-chromosome 6,810,542 Mb on the X-chromosome with 99.38% sequence identity between them. The sequence identity of the next alignment block between the two sex chromosomes spanning 9,999bp dropped drastically to 92.79%. The sheep PAR region spanned 7,019,999bp in two blocks from the start of the Y-chromosome and between 179bp and 6,987,958 on the X-chromosome, at an average of 98.85% sequence identity between them, while the next alignment block of 9,999bp dropped to 95.07 sequence identity.

X-d regions

The X-degenerate region of the Y-chromosomes comprise homologous regions with the X-chromosome outside the PAR and without active recombination with it. This region was defined by aligning the Y-chromosome to the X-chromosome with Mashmap. Mashmap was run with segment length “-s 5000” and percentage identity “-pi 65”.

Gene annotation

The protein-coding genes on the Y-chromosomes were initially annotated manually in two ways - using Liftoff v1.6.3⁷⁰ and by homology search with protein sequences from cattle and sheep, as well as from the well-annotated human and mouse genomes. Homology-based annotation was necessary to supplement the lack of Y chromosome-specific genes from the liftoff of the current cattle and sheep reference assemblies since the reference had no annotated Y. Refined annotation became available upon submission to NCBI through the application of their RefSeq annotation pipeline, revealing only slight disparities in the number and loci of genes from our manual annotations (Supplementary Tables S3 and S4). However, we used the NCBI annotations for the downstream analysis.

Liftoff was run with the parameters -flank 0.0 -s 0.4 -exclude_partial -copies -sc 0.98 -cds on the Y-chromosome and X-chromosome assemblies of both species using the X-chromosomes of the current assemblies of the Rambouillet ARS-UI-Ramb_v2.0⁷¹ for sheep and the ARS_UCD_1.3⁷² for cattle as reference. The liftoff result was filtered for only protein-coding genes (Supplementary Tables S3 and S4). Homology-based gene annotation on the Y-chromosomes was done using protein sequences from cattle, sheep, mouse and the human T2T genome assemblies. Miniprot⁷³ was used with the -outc=0.7 option to output only hits with at least 70% coverage between the query and the target. The result was filtered with the alignment score (Supplementary Tables S3 and S4) before manual curation with Apollo⁷⁴, a genome annotation plug-in for JBrowse⁷⁵.

Repeat elements annotation

Repeats analysis was done on the Y-chromosomes with RepeatMasker version 4.1.2-pl⁷⁶ using a combined repeats library produced from the *Dfam* library (dfam.h5 version 3.7) and an older library (RepeatMasker.lib) in order to have a comprehensive database of repeat elements. RepeatMasker was run with the options -species “bos taurus” -xsmall -no_is -paths -threads -gff.

Comparison of the T2T Y-chromosomes with publicly available assemblies

We collected all the cattle and sheep genome assemblies where male samples were reported sequenced from NCBI. Out of a total of eleven assemblies published for cattle, seven assemblies came from male samples, but only Btaurus_INIA1 was not assembled to chromosome level. All the chromosome level assemblies reported partial Y-chromosome assemblies except Bos_taurus_UMD_3.1.1 which had none (Supplementary Table 15). The current cattle reference genome ARS_UCD1.3, does not have a Y-chromosome assembly since it is from a Hereford cow. We thus selected the Y chromosome from the Btau_5.0.1 (GCA_000003205.6) assembly being the longest at 43.3Mb.

For the publicly available sheep assemblies, from 56 whole genome assemblies NWAUFU_Friesian_1.0 and ASM2243283v1 were the two chromosome-level assemblies out of the 6 produced from male animals (Supplementary Table 14). Take note that assemblies ASM2243283v1, ASM2132593v1 and ASM2132590v on one hand, and assemblies ASM2270250v1 and ASM2270250v on the other hand were produced from the same biotypes and were thus not treated as unique assemblies. We selected the Hu Sheep assembly¹⁴ which covered only the MSY of the sheep Y chromosome for comparison with our T2T Y-chromosome.

We aligned the T2T Y-chromosome assemblies to the publicly available assemblies with Mashmap⁶⁹ at a minimum identity of 90% and segment length of 5kb.

Defining the centromere

The centromeres of the Y-chromosomes were defined using satellite DNA repeats annotation Tandem Repeat Finder (TRF)⁷⁷, CENP-A enrichment analysis from CUT&RUN data, and epigenetic information of methylated Cytosine residues at the centromeric region.

Satellite DNA annotation with Tandem Repeat Finder (TRF)

TRF was run on the two Y-chromosome assemblies with the following parameters:

```
trf $inFile $match $mismatch $indel $PM $PI $minscore $maxperiod -d -h -ngs > trfOutput.txt
```

match=2, mismatch=7, indel=7, PM=80, PI=5, minscore=200, maxperiod=2000

The output of TRF was converted into a bed file and visually inspected on JBrowse to identify regions with tandem arrays of repeat elements. The 73bp repeat sequence which was identified in this manner from the centromeric tandem array was aligned back to the cattle (Wagyu and Charolais) and sheep (Churro and Friesian) assemblies using Blastn⁷⁸ to estimate their relative abundance. The blast results were filtered with a minimum sequence coverage of 80% between the query and the target, and a minimum e-value of 1E-3.

CENP-A enrichment analysis

CENP-A pull-down assay was produced from bovine satellite cells obtained from semimembranosus muscle of a 53-day old commercial male Angus calf using the GeneTex antibody with catalog number GTX13939 and sequenced. Adapter sequences were trimmed from the raw reads using Cutadapt⁷⁹. The trimmed high-quality reads were then aligned to the whole genome sequence of the Y chromosome-containing Wagyu cattle using Bowtie2⁸⁰ with default parameters asides "-k 100". The alignments were sorted with Samtools⁶⁸ and duplicates were removed with Picard toolkit (<https://broadinstitute.github.io/picard/>). Bedgraph files were created from the alignment files as input into the SEACR⁸¹ peaks caller to call the CENP-A peaks. SEACR was run using the stringent and 10% filter options.

```
bash SEACR_1.3.sh target.bedgraph 0.1 non stringent output
```

Methylated Cytosines analysis (CpG Methylation)

CpG methylation data was produced with ONT and PacBio HiFi reads using the manufacturer-prescribed protocols.

Oxford Nanopore Technology (ONT)

ONT-based methylated CpG sites were called in four main steps:

1. Methylated bases were called bases with Dorado using the R9_v0.3.4_v3.3_5hmc-5mc_A100mig model
2. Fastq files were extracted from the *modbams* and the MM, ML tags were retained
3. The fastq files were aligned to the reference assembly with Winnowmap⁸². The resulting sequence alignment map (.sam) file was filtered, converted to a bam file and indexed.
4. The bam file was parsed into modbam2bed to produce the bed file containing the methylated sites filtered using the 20% threshold for unmethylated bases and 80% for methylated bases.

Pacific Biosciences (PacBio)

The prescribed PacBio pipeline for calling CpG methylated sites is as follows:

1. Primrose was used to estimate the probability of methylation from the raw reads at each of the CpG sites.
2. The resulting bam files were aligned to the reference assembly with pbmm2 using the command:

pbmm2 align \

-j 48 \

-sort \

-log-level INFO \

-preset HIFI \$1 \$2 \$3, where \$1 is the reference assembly, \$2 is the file of list of modbams, and \$3 is the output

3. *Aligned_bam_to_cpg_scores* from *pb-CpG-tools* version 2.3.2 was used to produce the percentage of methylated reads across the assembly:

aligned_bam_to_cpg_scores \

-bam \$3 \

-output-prefix \$4 \

-model pileup_calling_model.v1.tflite \

-modsites-mode reference \

-ref \$1 \

The CpG methylation calls obtained from these steps were post-processed by first changing all the

NANs in the methylation values to 0.00s. The aggregate methylation levels for 5kb and 10kb bins were calculated using custom python scripts. The resulting data were plotted with an R script in RStudio to visualize the tracks of the methylation pattern. With visual inspection of the methylation tracks, the Centromeric dip region (CDR) was defined according to⁵² as the region with a local dip in methylated Cytosines with respect to the surrounding sequence and flanking the active satellite array.

Transcript level quantification on the Y-chromosomes

We obtained publicly available RNA-Seq data of transcripts quantification experiments from different tissues in order to measure the transcription activity of the genes on the Y-chromosomes.

Sheep: Two sets of transcriptome data were obtained for the transcription activity analysis on the sheep Y-chromosome.

- i. PRJNA552574: RNA-Seq was used to investigate whether testis development and gene expression is altered by the exposure of sheep to gossypol in utero and during lactation⁸³. We obtained RNA-Seq data from a total of 18 testes samples from 60-day old lambs with 9 in control and 9 in treated groups were used.
- ii. PRJNA437085: This study used RNA-Seq to investigate the effects of dietary energy levels on Hu sheep testicular development⁸⁴. RNA-Seq data from three replicates each of the two energy levels were obtained for our analysis.

Cattle: Two sets of transcriptome data were also used for the cattle.

- i. PRJNA565682: Six samples were retrieved from the project in an experiment to profile the transcriptome of mature and immature testes⁸⁵. Three replicates each of the mature (24-month old) and immature (1-day old) testes of Chinese Red Steppes cattle were used (Supplementary Table 9).
- ii. The second set was RNA-Seq data obtained from the Lung, Spleen, Muscle and Liver tissues of the Wagyu_x_Charolais F1 offspring from which the cattle Y-chromosome was obtained.

Quality control steps were carried out on the raw *.fastq* files to remove adapter, poly-X and poly-G sequences using fastp⁸⁶ with the parameters `-f 4 --trim_poly_g --trim_poly_x`

STAR splice-aware aligner⁸⁷ was then used to align the reads from the multi-tissue RNA-Seq data to the whole genome assemblies of the Wagyu and the Churro from where the Y-chromosomes were obtained. Before aligning the reads, the index of the whole genome assemblies were built with the command

```
STAR --runThreadN $threads --runMode genomeGenerate --genomeDir $pathToGenome \  
--genomeFastaFiles $pathToGenomeFasta --sjdbGTFfile $annotationPath \  
--sjdbGTFtagExonParentTranscript Parent --sjdbOverhang 100
```

The reads were then aligned with STAR in single or paired end reads mode depending on the data.

For single end reads mode:

```
STAR --runThreadN $threads --runMode alignReads --genomeDir $pathToGenome \  
--readFilesIn $pathToRead1 \  
--readFilesCommand zcat --outFileNamePrefix $pathToOutputFileWithPrefix \  
--outSAMtype BAM SortedByCoordinate --quantMode GeneCounts
```

For paired-end reads mode:

```
STAR --runThreadN $threads --runMode alignReads --genomeDir $pathToGenome \  
--readFilesIn $pathToRead1 $pathToRead2 \  
--readFilesCommand zcat --outFileNamePrefix $pathToOutputFileWithPrefix \  
--outSAMtype BAM SortedByCoordinate --quantMode GeneCounts
```

The output is the number of reads aligned to each gene.

Phylogenetic analysis

The sequence for each member of the ampliconic gene families were extracted from the coordinates using *bedtools getfasta* module. Multiple sequence alignment was performed using Mafft⁸⁸. The output was imported to Unipro Ugene using the PhyML maximum likelihood tree building method with branch length optimized.

Dot plots

The self-alignment dotplots of the Y-chromosome assemblies were produced for visualization with Moddotplot (<https://github.com/marbl/ModDotPlot>) using the program parameters `--no-bed --identity 85 -s 52`.

DECLARATIONS

DATA AVAILABILITY

The T2T Y-chromosomes have also been added to the current reference assemblies of cattle and sheep on NCBI. They can also be found independently in the NCBI SRA under the accessions CP128563.1 for the cattle, and CP128831.1 for the

sheep. The RNA-Seq data used for the transcriptional activity analysis on the Y-chromosomes were retrieved from NCBI with the project accessions: sheep (PRJNA552574 and PRJNA437085) and cattle (PRJNA565682).

ETHICS STATEMENT

This study was carried out in accordance with the University of Idaho Institutional Animal Care and Use Committee (IACUC) approved protocol with number IACUC-2020-58 for sheep and IACUC-2021-21 for cattle.

FUNDING INFORMATION

This material is based upon work that is supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, award number USDA-NIFA-2021-67016-33416 and Hatch grant no. IDA01566 from the USDA National Institute of Food and Agriculture. SK and AR are supported, in part, by the Intramural Research Program of the National Human Genome Research Institute (NHGRI), National Institutes of Health (NIH).

CONFLICT OF INTEREST DECLARATION

SK has received travel funds to speak at events hosted by Oxford Nanopore Technologies.

REFERENCES

1. Ross, M. T. *et al.* The DNA sequence of the human X chromosome. *Nature* **434**, 325–337 (2005).
2. Hughes, J. F. *et al.* Sequence analysis in *Bos taurus* reveals pervasiveness of X-Y arms races in mammalian lineages. (2020) doi:10.1101/gr.269902.120.
3. Luo, Z.-X., Yuan, C.-X., Meng, Q.-J. & Ji, Q. A Jurassic eutherian mammal and divergence of marsupials and placentals. *Nature* (2011) doi:10.1038/nature10291.
4. Bellott, D. W. *et al.* Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**, 494–499 (2014).
5. Charlesworth, B., Harvey, P. H., Charlesworth, B. & Charlesworth, D. The degeneration of Y chromosomes. *Philos Trans R Soc Lond B Biol Sci* **355**, 1563–1572 (2000).
6. Delbridge, M. L. & Graves, J. A. M. Mammalian Y chromosome evolution and the male-specific functions of Y chromosome-borne genes. *Reviews of Reproduction* vol. 4 101–109 Preprint at <https://doi.org/10.1530/ror.0.0040101> (1999).
7. Liu, W. S. *et al.* A novel testis-specific protein, PRAMEY, is involved in spermatogenesis in cattle. *Reproduction* **153**, 847–863 (2017).
8. Yamauchi, Y. *et al.* Loss of mouse Y chromosome gene *Zfy1* and *Zfy2* leads to spermatogenesis impairment, sperm defects, and infertility. *Biol Reprod* **106**, 1312–1326 (2022).
9. Pacheco, H. A., Rezende, F. M. & Peñagaricano, F. Gene mapping and genomic prediction of bull fertility using sex chromosome markers. *J Dairy Sci* **103**, 3304–3311 (2020).
10. Yue, X. P. *et al.* Copy number variations of the extensively amplified Y-linked genes, *HSFY* and *ZNF280BY*, in cattle and their association with male reproductive traits in Holstein bulls. *BMC Genomics* **15**, 1–12 (2014).
11. Rossetti, C. *et al.* State of the art on the physical mapping of the Y-chromosome in the Bovidae and comparison with other species – A review. *Anim Biosci* **35**, 1289–1302 (2022).
12. Liu, R. *et al.* New insights into mammalian sex chromosome structure and evolution using high-quality sequences from bovine X and Y chromosomes. *BMC Genomics* **20**, 1000 (2019).
13. Liu, W.-S. & Len, F. Mapping of the Bovine Y Chromosome. *Electronic Journal of Biology* **3**, (2007).

14. Li, R. *et al.* A Hu sheep genome with the first ovine Y chromosome reveal introgression history after sheep domestication. **64**, 1116–1130 (2021).
15. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* (2023) doi:10.1038/s41586-023-06457-y.
16. Chen, J. *et al.* A complete telomere-to-telomere assembly of the maize genome. *Nat Genet* **55**, 1221–1231 (2023).
17. Nurk, S. *et al.* The complete sequence of a human genome. *Science* (1979) **376**, 44–53 (2022).
18. Marx, V. Method of the year: long-read sequencing. *Nature Methods* vol. 20 6–11 Preprint at <https://doi.org/10.1038/s41592-022-01730-w> (2023).
19. Rautiainen, M. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* (2023) doi:10.1038/s41587-023-01662-6.
20. Koren, S. *et al.* De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol* **36**, 1174–1182 (2018).
21. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
22. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37**, 1155–1162 (2019).
23. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**, 338–345 (2018).
24. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 1–27 (2020).
25. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).
26. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
27. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
28. Johnson, T. *et al.* Short communication: Identification of the pseudoautosomal region in the Hereford bovine reference genome assembly ARS-UCD1.2. *J Dairy Sci* **102**, 3254–3258 (2019).
29. Robinson, J. T. *et al.* Integrative genomics viewer. (2011) doi:10.1038/nbt0111-24.
30. Hughes, J. F. *et al.* Strict evolutionary conservation followed rapid gene loss on human and rhesus y chromosomes. *Nature* **483**, 82–87 (2012).
31. Liu, R. *et al.* New insights into mammalian sex chromosome structure and evolution using high-quality sequences from bovine X and Y chromosomes. *BMC Genomics* **20**, 1–11 (2019).
32. Rossetti, C. *et al.* State of the art on the physical mapping of the Y-chromosome in the Bovidae and comparison with other species – A review. *Anim Biosci* **35**, 1289–1302 (2022).
33. Skaletsky, H. *et al.* *The Male-Specific Region of the Human Y Chromosome Is a Mosaic of Discrete Sequence Classes.* www.nature.com/nature (2003).
34. Liu, W.-S. *et al.* A novel testis-specific protein, PRAMEY, is involved in spermatogenesis in cattle. *Reproduction* **153**, 847–863 (2017).
35. Yang, Y. *et al.* ZNF280BY and ZNF280AY: Autosome derived Y-chromosome gene families in Bovidae. *BMC Genomics* **12**, 9–11 (2011).
36. Soh, Y. Q. S. *et al.* Sequencing the Mouse Y Chromosome Reveals Convergent Gene Acquisition and Amplification on Both Sex Chromosomes. *Cell* **159**, 800–813 (2014).
37. Paria, N. *et al.* A Gene Catalogue of the Euchromatic Male-Specific Region of the Horse Y Chromosome: Comparison with Human and Other Mammals. *PLoS One* **6**, e21374 (2011).

38. Ghenu, A.-H., Bolker, B. M., Melnick, D. J. & Evans, B. J. Multicopy gene family evolution on primate Y chromosomes. *BMC Genomics* **17**, 157 (2016).
39. Verkaar, E. L. C. *et al.* Organization and concerted evolution of the ampliconic Y-chromosomal TSPY genes from cattle. *Genomics* **84**, 468–474 (2004).
40. Hamilton, C. K. *et al.* Copy Number Variation of Testis-Specific Protein, Y-Encoded &i>(TSPY)&i> in 14 Different Breeds of Cattle &i>(Bos taurus)&i>; *Sexual Development* **3**, 205–213 (2009).
41. Yue, X. P. *et al.* Copy number variation of PRAMEY across breeds and its association with male fertility in Holstein sires. *J Dairy Sci* **96**, 8024–8034 (2013).
42. Pei, S. W., Qin, F., Li, W. H., Li, F. D. & Yue, X. P. Copy number variation of ZNF280AY across 21 cattle breeds and its association with the reproductive traits of Holstein and Simmental bulls. *J Dairy Sci* **102**, 7226–7236 (2019).
43. Chang, T. C., Yang, Y., Retzel, E. F. & Liu, W. S. Male-specific region of the bovine Y chromosome is gene rich with a high transcriptomic activity in testis development. *Proc Natl Acad Sci U S A* **110**, 12373–12378 (2013).
44. Hamilton, C. K., Verduzco-Gmez, A. R., Favetta, L. A., Blondin, P. & King, W. A. Testis-specific protein Y-encoded copy number is correlated to its expression and the field fertility of Canadian Holstein bulls. *Sexual Development* **6**, 231–239 (2012).
45. Mukherjee, A. *et al.* Copy Number Differences of Y Chromosomal Genes Between Superior and Inferior Quality Semen Producing Crossbred (Bos taurus × Bos indicus) Bulls. *Anim Biotechnol* **26**, 65–72 (2015).
46. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Arkarachai Fungtammasan* **10**.
47. Ye, D. *et al.* High Levels of Copy Number Variation of Ampliconic Genes across Major Human Y Haplogroups. *Genome Biol Evol* **10**, 1333–1350 (2018).
48. Jiang, Y. *et al.* The sheep genome illuminates biology of the rumen and lipid metabolism. *Science (1979)* **344**, 1168–1173 (2014).
49. Vegesna, R. *et al.* Ampliconic Genes on the Great Ape Y Chromosomes: Rapid Evolution of Copy Number but Conservation of Expression Levels. *Genome Biol Evol* **12**, 842–859 (2020).
50. Vegesna, R., Tomaszkiwicz, M., Medvedev, P. & Makova, K. D. Dosage regulation, and variation in gene expression and copy number of human Y chromosome ampliconic genes. *PLoS Genet* **15**, e1008369 (2019).
51. Miga, K. H. & Alexandrov, I. A. Variation and Evolution of Human Centromeres: A Field Guide and Perspective. *Annual Review of Genetics* vol. 55 583–602 Preprint at <https://doi.org/10.1146/annurev-genet-071719-020519> (2021).
52. Gershman, A. *et al.* Epigenetic patterns in a complete human genome. *Science (1979)* **376**, (2022).
53. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
54. Ohshima, K. & Okada, N. SINES and LINES: Symbionts of eukaryotic genomes with a common tail. *Cytogenet Genome Res* **110**, 475–490 (2005).
55. Escudeiro, A. *et al.* Bovine satellite DNAs—a history of the evolution of complexity and its impact in the Bovidae family. *European Zoological Journal* **86**, 20–37 (2019).
56. Evans, H. J., Buckland, R. A. & Sumner, A. T. Chromosome homology and heterochromatin in goat, sheep and ox studied by banding techniques. *Chromosoma* **42**, 383–402 (1973).
57. Gallagher Jr., D. S. *et al.* A Molecular Cytogenetic Analysis of the Tribe Bovini (Artiodactyla: Bovidae: Bovinae) with an Emphasis on Sex Chromosome Morphology and NOR Distribution. *Chromosome Research* **7**, 481–492 (1999).
58. Makova, K. D. *et al.* The Complete Sequence and Comparative Analysis of Ape Sex Chromosomes. *bioRxiv* 2023.11.30.569198 (2023) doi:10.1101/2023.11.30.569198.
59. Cortez, D. *et al.* Origins and functional evolution of Y chromosomes across mammals. (2014) doi:10.1038/nature13151.
60. Kalbfleisch, T. *et al.* *RT2T: A Global Collaborative Project to Study Chromosomal Evolution in the Suborder Ruminantia.* (2024) doi:10.21203/rs.3.rs-3918604/v1.

61. Marais, G. A. B., Campos, P. R. A. & Gordo, I. Can Intra-Y Gene Conversion Oppose the Degeneration of the Human Y Chromosome?: A Simulation Study. *Genome Biol Evol* **2**, 347–357 (2010).
62. Trombetta, B., D'Atanasio, E. & Cruciani, F. Patterns of inter-chromosomal gene conversion on the male-specific region of the human Y chromosome. *Frontiers in Genetics* vol. 8 Preprint at <https://doi.org/10.3389/fgene.2017.00054> (2017).
63. Melters, D. P. *et al.* Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol* **14**, 1–20 (2013).
64. Henikoff, S., Ahmad, K. & Malik, H. S. The Centromere Paradox: Stable Inheritance with Rapidly Evolving DNA. *Science (1979)* **293**, 1098–1102 (2001).
65. Formenti, G. *et al.* Gfastats: Conversion, evaluation and manipulation of genome sequences using assembly graphs. *Bioinformatics* **38**, 4214–4216 (2022).
66. Huang, N. & Li, H. miniBUSCO: a faster and more accurate reimplement of BUSCO. *bioRxiv* 2023.06.03.543588 (2023) doi:10.1101/2023.06.03.543588.
67. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
68. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
69. Jain, C., Koren, S., Dilthey, A., Phillippy, A. M. & Aluru, S. A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics* **34**, i748–i756 (2018).
70. Shumate, A. & Salzberg, S. L. Liftoff: Accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
71. Davenport, K. M. *et al.* An improved ovine reference genome assembly to facilitate in-depth functional annotation of the sheep genome. *Gigascience* **11**, (2022).
72. Rosen, B. D. *et al.* De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* **9**, (2020).
73. Li, H. Protein-to-genome alignment with minimap2. *Bioinformatics* **39**, (2023).
74. Dunn, N. A. *et al.* Apollo: Democratizing genome annotation. *PLoS Comput Biol* **15**, e1006790 (2019).
75. Buels, R. *et al.* JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol* **17**, 66 (2016).
76. Smit AFA, Hubley R & Green P. RepeatMasker Open-4.0. (2013).
77. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–580 (1999).
78. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).
79. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* **17**, 10 (2011).
80. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
81. Meers, M. P., Tenenbaum, D. & Henikoff, S. Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. *Epigenetics Chromatin* **12**, 42 (2019).
82. Jain, C. *et al.* Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**, i111–i118 (2020).
83. Louvandini, H. *et al.* Gestational and lactational exposure to gossypol alters the testis transcriptome. *BMC Genomics* **21**, 59 (2020).
84. Fan, Y. X. *et al.* Effect of dietary energy restriction and subsequent compensatory feeding on testicular transcriptome in developing rams. *Theriogenology* **119**, 198–207 (2018).
85. Fang, X. *et al.* Comprehensive Analysis of miRNAs and Target mRNAs between Immature and Mature Testis Tissue in Chinese Red Steppes Cattle. *Animals* **11**, 3024 (2021).
86. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
87. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

88. Yamada, K. D., Tomii, K. & Katoh, K. Application of the MAFFT sequence alignment program to large data—reexamination of the usefulness of chained guide trees. *Bioinformatics* **32**, 3246–3251 (2016).

Table

Table 1: The structure and gene content of the cattle and sheep Y-chromosomes

Cattle				Sheep			
Chromosome structure							
Feature	length (bp)	start	end	length (bp)	start	end	
Telomere-p	14,684	1	14,685	19,266	1	19,267	
PAR	6,807,694	14,686	6,822,380	6,999,061	19,268	7,018,329	
MSY	52,633,132	6,822,831	59,455,963	18,880,972	7,018,330	25,899,302	
Telomere-q	20,325	59,455,964	59,476,289	17,775	25,899,303	25,917,078	
centromere	2,521,056	14,124,633	16,645,689	120,763	8,038,393	8,159,156	
Annotation							
	Protein-coding (PAR, X-d, Amp)	236 (34, 15, 187)		109 (45, 18, 46)			
	Pseudogenes (PAR, X-d, Amp)	50 (2, 20, 28)		127 (0, 23, 104)			
	lncRNA	14		29			
	tRNA	9		7			
	snRNA	5		5			
	snoRNA	1		1			
Multi-copy genes							
PAR genes	<i>ASMTL</i>	2		2			
	<i>OBP</i>	3		4			
	<i>BOS2D</i>	2		4			
	<i>CSF2RA</i>	1		2			
X-d genes	<i>USP9X</i>	1		2			
Ampliconic genes	<i>HSFY (Protein-coding, pseudogene)</i>	40 (37, 3)		26 (12, 14)			
	<i>HSFY2</i>	2 (2, 0)		0 (0, 0)			
	<i>PRAME</i>	31 (31, 0)		6 (1, 5)			
	<i>RBMY</i>	11 (11, 0)		1 (1, 0)			
	<i>TSPY1</i>	82 (68, 14)		52 (0, 52)			
	<i>TSPY3</i>	16 (16, 0)		24 (5, 19)			
	<i>ZNF280B</i>	40 (22, 18)		42 (27, 15)			
	<i>ZNF280A</i>	15 (0, 15)		22 (0, 22)			

X-d = X-degenerate

Figures

Figure 1. Global structure of the cattle and sheep Y chromosomes

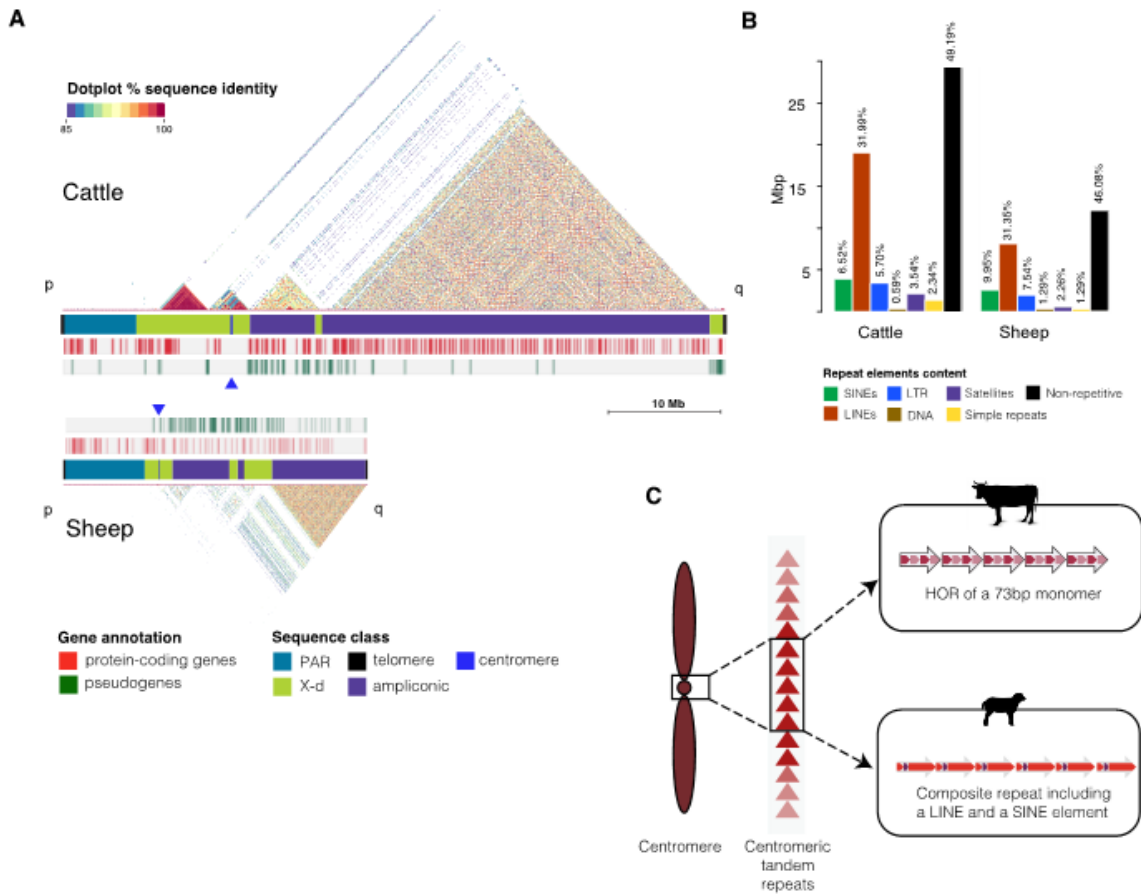


Figure 1: The structure of the cattle and the sheep Y-chromosomes showing the differences and the similarities between the two species. (A) The self-identity dotplots highlight the mosaic of repetitive sequence on the q-arms while the tracks below it show the different sequence classes and the gene annotation. (B) The repeat content annotation indicates that more than half of both Y-chromosomes is repetitive DNA. (C) Centromere content and organization shows that the two Y-chromosomes have tandemly arrayed repeat unit at the centromere; while the cattle centromeric repeat is organized into a higher order repeat (HOR) of a 73bp monomer, the sheep repeat unit is a composite of a tandem LINE and SINE element with spacer DNA between the copies.

Figure 1

See image above for figure legend.

Figure 2: Ampliconic genes content comparison between cattle and sheep

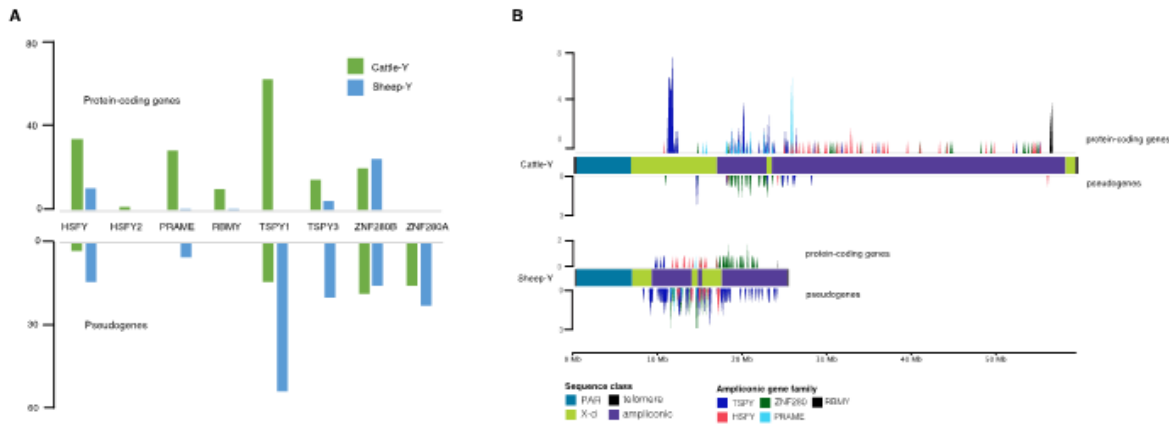


Figure 2: (A) The copy distribution of the protein-coding and the pseudogene copies of the ampliconic genes families on the cattle and sheep Y-chromosomes show remarkable differences. (B) The gene density of the ampliconic genes across the Y-chromosomes calculated in 100kb bin sizes highlighting the loci of the protein-coding genes on the upper tracks and the pseudogenes on the lower tracks. There are no protein-coding copies of TSPY1 on the Sheep-Y relative to the cattle-Y harboring a tandem array of 44 copies out of the total 68 protein-coding copies on the chromosome. More pseudogenes are found on the Sheep-Y relative to the protein-coding genes suggesting that some of the ampliconic genes are being pseudogenized, losing their protein-coding capabilities.

Figure 2

See image above for figure legend.

Figure 3: Content and organization of the Y-chromosome centromeres

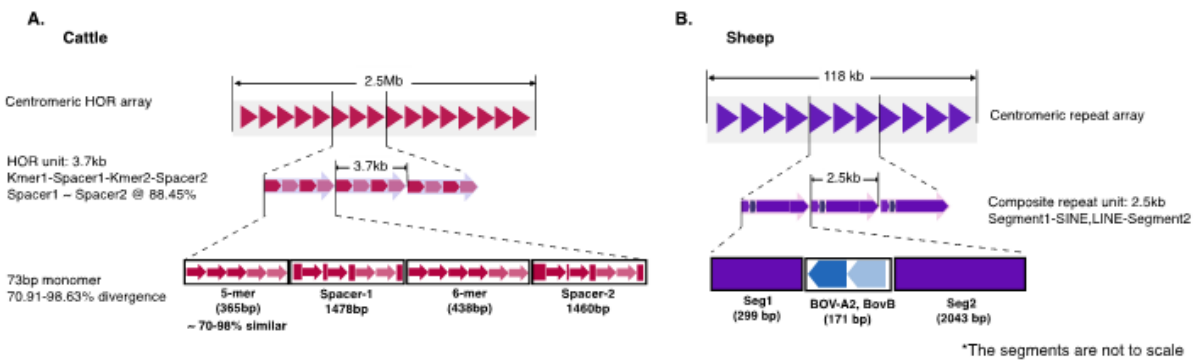


Figure 3: (A) The Cattle-Y centromere is organized as a tandem array of highly identical higher-order repeat (HOR) unit spanning 2.5Mb. The 3.7kb HOR contains copies of a 73bp monomeric unit arranged into four segments where segments 1 and 3 are tandem arrays of 5 copies (5-mer) and 6 copies (6-mer) respectively while segments 2 and 4 contain copies of the monomeric unit but not in tandem arrays as in segments 1 and 3. (B) The Sheep-Y centromere spanning about 118kb organized as an array of a 2.5kb composite repeat unit comprising a BOV-A2 SINE and a BovB LINE transposable elements embedded between two segments of DNA.

Figure 3

See image above for figure legend.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryTables.xlsx](#)
- [T2TYChromosomeManuscriptFiguresandTables.docx](#)