3-9-2021

# De novo assembly and annotation of the North American bison (Bison bison) reference genome and subsequent variant identification

L K. Dobson

A Zimin

D Bayles

E Fritz-Waters

D Alt

*See next page for additional authors*

## Authors

L K. Dobson, A Zimin, D Bayles, E Fritz-Waters, D Alt, S Olsen, J Blanchong, J Reecy, T PL Smith, and J N. Derr

ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics

FULL PAPER

doi: 10.1111/age.13060

# *De novo* assembly and annotation of the North American bison (*Bison bison*) reference genome and subsequent variant identification

**L. K. Dobson*, A. Zimin†, D. Bayles‡, E. Fritz-Waters§, D. Alt‡, S. Olsen‡, J. Blanchong¶, J. Reecy¶, T. P. L. Smith** and J. N. Derr***

*Department of Veterinary Pathobiology, Texas A&M University, College Station, TX 77845, USA. †Department of Biomedical Engineering, Center for Computational Biology, Johns Hopkins University, Baltimore, MD 21205, USA. ‡Infectious Bacterial Diseases Research Unit, National Animal Disease Center, Agricultural Research Service, United States Department of Agriculture, Ames, IA 50010, USA. §Department of Animal Science, Iowa State University, Ames, IA 50011, USA. ¶Department of Natural Resource Ecology and Management, Iowa State University, Ames, IA 50011, USA. **U.S. Meat Animal Research Center, USDA-ARS, Clay Center, NE 68933, USA.

## Summary

Genomic tools have improved the ability to manage bison populations and enhanced efforts to conserve this iconic species. These tools have been particularly useful for detecting introgression of cattle genome within bison herds but are limited by the need to use the cattle genome as a surrogate for mapping reads. This complicates efforts to distinguish the species of origin of chromosomal segments in individual bison at the genomic level. An assembly (Bison_UMD1.0) based on 75X genome coverage by Illumina and 454 reads was generated using the MaSuRCA assembler, generating a 2.81 Gigbases *de novo* reference genome from American bison. Comparison of bison and domestic cattle references identified 28 443 364 single nucleotide variants and 2 627 645 insertions/deletions distinguishing the species. Sequence alignment of an additional 12 modern bison samples and two historic bison samples to domestic cattle and bison references provides a dataset of genomic variants defining the different species and within-species variation. This first annotated draft assembly represents a resource for the management and conservation of bison, as well as a means to study the effects on the genome of interspecies hybridization. The comparisons of historical bison sequences with the new bison reference identified genomic differences between modern and pre-population bottleneck bison. The results support the application of genomics to enhance future research on disease, the establishment of satellite conservation herds and insight into bison and cattle speciation. The first genome assembly for bison and dataset provides a foundation that can be built upon as genetic technologies improve over the years.

**Keywords** conservation, hybridization, management, population genomics

## Background

American bison (*Bison bison*) are an iconic species symbolizing the early colonization of North America. However, extensive over-hunting of the species in the late 1800s resulted in the almost complete decimation of the species, producing a population bottleneck that greatly reduced genetic diversity (Coder, 1975; Dary, 1989). Further decline of the species came from efforts to hybridize bison with cattle (*Bos taurus*) in attempts to improve the hardiness of beef cattle raised on the Great Plains (Goodnight, 1912; Coder, 1975). These efforts created the current bison genome which now defines a hybrid species with both bison and cattle genetics, as evidenced by a reduction in body size and the identification of cattle mitochondrial sequence in bison (Verkaar *et al.*, 2003; Derr *et al.*, 2012). The decreased genetic diversity among bison and the introgression of cattle DNA into the species present challenges in the management and conservation of the

Address for correspondence

L. K. Dobson, Department of Veterinary Pathobiology, Texas A&M University, College Station, TX 77845, USA.
E-mail: ldobson@cvm.tamu.edu

*Accepted for publication 09 March 2021*

American bison today, which is imperative as bison is the national mammal of the United States.

Approximately 30 000 of the 500 000 bison in North America are found in conservation herds, with the remainder found in private production herds (Boyd, 2003; Halbert *et al*., 2005). Many of the North American bison herds have been shown to carry traces of cattle genomes as a result of hybridization, such that differentiation between hybrid and non-introgressed bison within a population is difficult (Polziehn *et al*., 1995; Ward *et al*., 1999; Ward, 2000; Halbert *et al*., 2005; Douglas *et al*., 2010). Recent studies have proven that certain conservation plains bison herds, as well as those found to have introgression with domestic cattle, have distinct genetic compositions owing to their having unique bison alleles and allelic distributions (Halbert, 2003; Freese *et al*., 2007). Whereas the primary focus for bison conservation has been on herds that are potentially hybridization free or have low levels of domestic cattle introgression, a large number of bison herds have conservation value owing to their historical genetic makeup (Freese *et al*., 2007).

The recent development of new genetic and genomic tools has improved the ability to manage bison populations and enhanced efforts to conserve the species. Accurate parentage testing and identification of QTL have improved population management to increase profitability and conservation efforts through population relationship assessments and cattle introgression detection (Polziehn *et al*., 1995; Ward *et al*., 1999; Schnabel *et al*., 2000; Halbert & Derr, 2007). The current technologies being used to test for domestic cattle genetics, in both the mitochondrial and nuclear DNA in bison populations, have been useful; however, they lack the resolution that is needed to detect cattle introgression in individual bison at the genomic level (Polziehn *et al*., 1995; Ward *et al*., 1999; Ward, 2000; Halbert, 2003; Halbert *et al*., 2005).

The use of whole-genome sequencing technology provides the next step in advancing bison management and conservation. However, the only bison assembly available is of the European wisent, *Bison bonasus*, from an animal whose mitochondrial sequence more closely resembles cattle than *B. bison* (Wang *et al*., 2017). Although the cattle genome sequence is available, using it as a guide to assemble a bison reference sequence would create domestic cattle reads in the bison sequence and lead to inconsistent alignments and misplaced reads while comparing sequences, and would not reflect all of the novelty of the bison genome (Gnerre *et al*., 2009). Therefore, providing a bison *de novo* reference assembly will allow researchers to not be limited by the need to use the cattle or wisent genome as a surrogate for mapping reads and allow for an unbiased genomic sequence determination.

In the present study, we selected an animal from the Yellowstone National Park (YNP) herd, named Templeton, based on molecular and cytogenetic data and park records, to represent the bison genome reference sequence. We present an annotated draft genome assembly, Bison_UMD1.0, of this animal and characterize the assembly for variants within bison and compared with the cattle reference assembly. The study also compares other re-sequenced bison from different herds with both the domestic cattle and bison reference genomes to provide a genomic variant list to be used for future studies. With this information on the bison genome, conservation management can be improved by identifying those herds that have high levels of genetic diversity, unique or historical lineages and low levels of domestic cattle introgression for the establishment of new bison herds on native ranges. The genome assembly and population data represent a resource for the management and conservation of bison, as well as a means of studying the effects on the genome of inter-species hybridization.

## Materials and methods

### Collection of DNA samples/isolation of DNA

The reference bison (aka Templeton) has a well-documented history, showing that it originated from YNP. This bison was utilized in a collaborative research project on brucellosis (Forgacs *et al*., 2016). When sampled, he was being managed as part of a brucellosis-free herd on a private ranch in Montana. In March 2011, blood, hair and tissue samples were collected (Appendix S1). DNA was isolated from 15 ml of blood using a standard phenol–chloroform–isoamyl alcohol extraction protocol (Sambrook *et al*., 1989).

### Whole-genome sequencing, assembly and annotation

The American bison genome (Bison_UMD1.0/Templeton) was assembled using a *de novo* assembly method that utilizes hybrid Illumina and 454 sequencing platforms. Using 30 μg of genomic DNA, four sequencing libraries were generated with approximately 20 kb paired-end single-stranded libraries and were circularized using a 'titanium' 42 bp linker for sequencing on a 454 GS-FLX Titanium™ sequencer following the manufacturer's protocol (GS FLX Titanium Series; Roche Applied Sciences). Ten paired-end libraries, with an approximately 390 bp insert size, were prepared following the manufacturer's protocols and a 5 kb Nextera jump mate-pair library was prepared for sequencing on Illumina HiSeq 2000™ (Illumina; 100 bp paired-end reads).

DNA sequence files were used to produce an approximately 75× coverage of a *de novo* reference assembly. The reference assembly was performed using the MaSuRCA assembler version 1.8.3 (Zimin *et al*., 2013). The MaSuRCA assembler is based on the idea of using a combination of the de Bruijn graph and the overlap-layout consensus (Celera assembler, version 6.1) methods. This is achieved by

reducing the most numerous and high-coverage Illumina paired-end reads to a much smaller set of long consensus super-reads. The super-reads are then assembled using the overlap-layout consensus method along with the error-corrected and filtered Illumina linking mate pair reads and the 454 paired-end reads. The assembly is followed by scaffold gap filling with subassemblies of Illumina reads as described in Zimin *et al.* (2013). For the American bison genome nearly 200 billion bases in close to 2 billion 101 bp paired-end Illumina reds data were reduced to about 7.2 billion bases in 26.7 million super-reads with an average length of 269 bases. Utilization of the super-reads reduced the assembly problem by a factor of 75 for the Illumina data.

This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under accession no. JPYT00000000. The version described in this paper is JPYT00000000.1 (JPYT01000000 (nig.ac.jp)). Annotation of the *de novo* bison reference genome sequence was completed using the assembled bison reference sequence and RNA sequences provided to NCBI and followed the NCBI pipeline using software version 6.2 (Appendix S1; Thibaud-Nissen *et al.*, 2013; http://www.ncbi.nlm.nih.gov/genome/annotation_euk/Bison_bison_bison/100/). The bison genome reference assembly can be found with the assembly accession no. GCF_000754665.1 and assembly name Bison_UMD1.0 at http://www.ncbi.nlm.nih.gov/assembly/GCF_000754665.1/. The database link can be found using the BioProject ID: PRJNA257088 and the BioSample ID: SAMN02947321 (NCBI). BUSCO analysis was also performed as a second analysis to check for the integrity of the genome and gene sets (Seppey *et al.*, 2019).

## Bison reference sequence alignment to domestic cattle reference and identification of genetic variants and analysis

Both the paired-end and mate-pair sequences of the bison reference raw reads were trimmed using FASTQ-MCF, which trimmed bases with a quality score less than 20 from each individual read and reads with a remaining sequence length of less than 70 bases (Aronesty, 2011). The filtered reads were mapped to the domestic cattle UMD3.1 reference sequence using BURROWS–WHEELER ALIGNMENT version 0.6.2 (BWA-MEM; Li, 2013) using the default settings. The resulting BAM (binary short DNA sequence read alignment; Li *et al.*, 2009) files were combined using the merge option of the SEQUENCE ALIGNMENT/MAP (SAM) tools 0.1.18 software package (Li *et al.*, 2009). Read group information was added using the AddOrRelpaceReadGroups option of PICARDTOOLS 1.7.1 (https://github.com/broadinstitute/picard/releases/tag/1.128). The GENOME ANALYSIS TOOLKIT 3.1.1 (GATK; McKenna *et al.*, 2010) option RealignerTargetCreator was used to realign and account for indel shifted coordinates to create a realigned and sorted BAM file of mapped reads to UMD3.1 reference. SAMTOOLS view and flagstat options (Li *et al.*, 2009) were used to obtain statistics of the alignment of the bison reference genome to the domestic cattle reference genome.

Genetic variants, SNPs and indels were called against the cattle reference for mapped reads and were filtered according to the GATK Best Practices recommendations (Depristo *et al.*, 2011; Van der Auwera *et al.*, 2013). The resultant variants were placed into VCF files. The VCFTOOLS 0.1.11 vcf-stats (Danecek *et al.*, 2011) option was used to determine basic statistics and counts of the SNPs and indels. These identified variants were then annotated using SNPEFF 4.1 software (Cingolani *et al.*, 2012) against the UMD3.1.76 reference from Ensembl. VCF files were deposited into the European Variation Archive (https://www.ebi.ac.uk/eva/?Home) with accession nos.

## Pseudo-chromosome mapping

Pseudo-chromosomes were produced using the UMD3.1.76 gff (http://useast.ensembl.org/Bos_taurus/Info/Annotation) chromosome file from Ensembl (Flicek *et al.*, 2014) and scaffolds of bison reference sequence to create synteny blocks using the software SYMAP 4.2 (Soderlund *et al.*, 2006). Figure S1 presents alignment using MINIMAP2 and visualization using the D-GENIES tool with the more recently produced European bison (wisent) assembly that indicates that no substantial error was introduced by using the higher-quality, but different species, assembly for chromosome scaffolding (Cabanettes & Klopp, 2018; Li, 2018).

## Whole-genome re-sequencing of historic samples, EIW, CCSP and YNP bison

Illumina paired-end libraries were prepared for sequencing for 14 bison samples on Illumina HiSeq 2000™ Next-Gen from the extracted DNA (Appendix S1; Table S1) for whole-genome resequencing using the Nextera DNA Sample Preparation Kit (Illumina). Historic samples (S6 and S9) were not combined owing to the lower quality DNA and libraries were prepared using the NEXTflex Illumina ChiP-Seq Library Prep Kit by Bioo Scientific protocol and run on one lane with the normal High Output 2×100 mode (Illumina). For each of the four samples from Caprock Canyons State Park (CCSP) and Elk Island National Park (EIW) the genomic libraries were indexed with adapters and four samples were run together on 2 HiSeq lanes using the 2×100 normal mode. Illumina TruSeq Nano libraries for the four samples from YNP were prepared using the Illumina TruSeq Nano DNA Sample Preparation Kit (Illumina), and run on four separate lanes on 2×100 mode. The sequence data for these samples has been deposited in NCBI within Bioproject PRJNA658430.

### Variant identification in relation to the domestic cattle and bison reference genomes

Prior to mapping the reads of the 14 re-sequenced bison to both the bison (Bison_UMD1.0) and domestic cattle reference sequences, raw reads were trimmed using the same method as previously described. Re-sequenced bison samples were individually mapped to the reference bison scaffolds and separately to the domestic cattle reference sequence using Burrows-Wheeler Alignment 0.6.2 (BWA-MEM; Li, 2013) and variants were called and annotated using the methods described above. VCFTOOLS was also used to identify shared or informative variants within populations to the bison reference to identify potential subpopulation variants.

In order to annotate the identified variants for the bison populations to the bison reference, the SyMap pseudochromosomes that were generated previously were used to change the scaffold IDs in the combined bison population VCF to actual chromosome numbers based on position. This allowed for the scaffolds in the combined VCF files for each population, or in the case of the historical samples individually, to be replaced by chromosome based on the positions created in the SyMap anchor file and using a perl script. The changed VCF files were then annotated in SNPEFF using UMD3.1.76 as a reference as the pseudochromosomes were generated from synteny blocks to the UMD31.76 reference. Custom script can be found in Appendix S2.

### Phylogenetic analysis

SNPHYLO version 20140701 (Lee *et al*., 2014) was used to generate a phylogenetic tree using the combined VCF file to domestic cattle (UMD3.1). The VCF file to UMD 3.1 was chosen for this analysis and not the combined VCF file for Bison_UMD1.0 so that the bison reference (Templeton) would be included in the phylogenetic analysis.

## Results

### Preliminary analysis

Current technologies available in our laboratory to assess for domestic cattle introgression (14 nuclear markers and TPW and 16S mitochondrial markers) and an additional 26 polymorphic markers were genotyped from the reference animal prior to the genome sequencing to ensure that the selected sample did not have detectable domestic cattle introgression (Appendix S1; Ward *et al*., 1999; Schnabel, Ward & Derr, 2000; Halbert, 2003); Fig. S2a). Karyotyping and FISH were also performed to ensure that normal chromosomes were obtained. Templeton was found to have bison mitochondrial DNA genotype and no domestic cattle introgression alleles were detected in the nuclear DNA. Alleles for microsatellites can be found in Fig. S2a. Templeton's main genetic composition when compared with the eight core US federal bison herds was as expected, with 91.0% of his genome coming from YNP (Fig. S2b).

Templeton was found to have normal chromosomes, a diploid number of $2n = 60$ and normal X and Y chromosomes (Fig. S2c,d). Cattle BAC containing PAR sequences was mapped to the short arm of the bison acrocentric Y chromosome, showing that the Y chromosome is structurally different from the *Bos taurus* Y chromosome, which is submetacentric (Di Meo *et al*., 2005; Das et al., 2009).

### Annotation

We generated a total of approximately 75× genome coverage by reads from two sequencing technologies: 454 sequencing by Roche and Illumina sequencing. Table 1 shows the library sizes, read lengths and the coverage for each library. The MASURCA assembler version 1.8.3 was used to assemble the sequencing data. The Bison_UMD1.0 (Templeton) assembly contained approximately 2.83 Gb of total sequence and was composed of 128 431 scaffolds with N50 contig size of 19.97 kb (L50 37 835) and scaffolds with an N50 scaffold size of 7.2 Mb (L50 116; Table 2).

Global statistics for the bison annotation and results of a BUSCO analysis are summarized in Table 3. When compared with the domestic cattle (UMD3.1) and human reference genome annotations (both HuRef_1 and HuRef2 (GRCh38)) the bison reference total sequence length was slightly larger than the cattle annotation and smaller than the two human reference genome annotations (Table 4). The bison genome had fewer genes and pseudogenes (combined together) when compared with either the bovine or human annotations, but it had more protein-coding genes than both species (Table 4).

### Bison reference sequence alignment to domestic cattle reference and identification of genetic variants and analysis

BWA mem (Li, 2013) was used to align raw bison DNA sequence paired-end and mate pair reads (1 008 038 624

**Table 1** Statistics for Illumina (paired-end and mate pair) and 454 paired-end libraries used for *de novo* bison reference sequence (Bison_UMD1.0/Templeton).

| Library | Average read length | Number of reads (millions) | Library mean size (bp) | Library standard deviation (bp) |
|---|---|---|---|---|
| Illumina | | | | |
| Paired-end | 101 | 1115 | 300 | 40 |
| Mate pair | 101 | 85 | 4000 | 800 |
| | 101 | 239 | 4500 | 900 |
| | 101 | 531 | 6000 | 1000 |
| 454 | | | | |
| Paired-end | 398 | 25.6 | 15 000 | 3500 |

**Table 2** Global statistics (in bp) for Bison_UMD1.0 (Templeton; NCBI).

| | Bison_UMD1.0 |
|---|---|
| Total sequence length | 2 828 031 685 |
| Total assembly gap length | 195 767 988 |
| Gaps within scaffolds | 341 984 |
| Number of scaffolds | 128 431 |
| Scaffold N50 | 7 192 658 |
| Number of contigs | 470 415 |
| Contig N50 | 19 971 |

**Table 3** Bison (UMD1.0/Templeton) reference genome annotation summary for gene and feature statistics from NCBI as well as results of BUSCO analysis.

| Feature | Bison_UMD1.0 |
|---|---|
| Total sequence length (bp) | 2 828 031 685 |
| Total number of chromosomes and organelles | 31 |
| Genes and pseudogenes | 26 001 |
| Protein-coding | 20 782 |
| Non-coding | 1677 |
| Pseudogenes | 3542 |
| Genes with variants | 6158 |
| mtDNA size | 16 319 |
| C, 86.5% [S, 85.6%; D, 0.9%], F, 3.9%, M, 9.6%, *n*, 9226 | |
| Complete BUSCO (C) | 7980 |
| Complete and single-copy BUSCOs (S) | 7899 |
| Complete and duplicated BUSCOs (D) | 81 |
| Fragmented BUSCOs (F) | 359 |
| Missing BUSCOs (M) | 887 |
| Total BUSCO groups searched | 9226 |

sequencing reads) against the UMD3.1 domestic cattle reference (Zimin *et al.* 2013). The SAMtools options view and flagstat (Li *et al.*, 2009) were used to obtain statistics of the bison Illumina paired-end reads mapped to the UMD_3.1 domestic cattle sequence. A total of 993 981 233 of the 1 008 038 624 (98.6%) bison reads mapped to the UMD3.1 cattle assembly (Zimin *et al.*, 2013), with 944 493 355 (93.7%) reads properly mapped or correctly oriented.

After identification of genomic variants, a total of 28 443 364 SNPs were discovered between Bison_UMD1.0 (Templeton) and the domestic cattle reference. Variant identification for SNPs and indels is summarized in Table 5. Some SNPs identified occurred when the variant was heterozygous for the reference (cattle) and for a bison variant allele, or when a cattle allele was found with a bison allele for that variant. It was expected that some positions in the bison genome would contain the same genomic sequences as observed in the cattle genome assembly because these species derived from a common ancestor. Previously it was believed that the split between domestic cattle and bison was approximately 0.5–2 mya in Eurasia (McDonald, 1981). With the advancements in genomics, however, this estimation ranges more to around 0.5–6.4 mya (Wang *et al.*, 2018). Overall, there was one SNP detected every 93 bases and 32 086 858 genome region and coding effects that were the result of the SNPs discovered.

There were 2 627 645 indels discovered when Bison_UMD1.0 (Templeton) was mapped against the domestic cattle reference, with 1 233 140 (46.9%) insertions and 1 394 505 (53.1%) deletions. All indels were annotated with SnpEff, and results are summarized in Fig. 1a. Chromosomal variant counts for both SNPs and indels for bison onto domestic cattle can be found in Fig. 1b, with chromosome 1 having the most detected variants. Figue 1c shows the count of variants with corresponding quality scores of the SNPs and indels annotated with SnpEff after filtering.

## Pseudo-chromosome mapping

Given that bison and domestic cattle shared a common ancestor and have the same number of chromosomes, we used the domestic cattle reference to generate pseudo-chromosomes to provide gene placements on chromosomes. SyMAP 4.2 (Soderlund *et al.*, 2006) was used to produce a synteny alignment between the Bison_UMD1.0 (Templeton) scaffolds and chromosomes from the UMD3.1.76 domestic cattle reference. SyMAP was able to create 447 synteny anchors and mapped only a total of 447 scaffolds to the 29 autosomes and the X chromosome of domestic cattle. However, in total, Bison_UMD1.0 scaffolds covered approximately 2 283 389 917 (85.5%) Gb of the 2 670 424 944 Gb UMD3.1.76 cattle reference genome. Table S3 depicts Bison scaffolds sorted by chromosome

**Table 4** Bison (UMD1.0) reference genome (Templeton) annotation comparison to domestic cattle (UMD3.1) and human [HuRef_1 and HuRef2 (GRCh38)] reference genome annotations.

| Feature | Bison_UMD1.0 | Cattle_UMD3.1 | HuRef_1 | HuRef_2 (GRCh38) |
|---|---|---|---|---|
| Total sequence length (bp) | 2 828 031 685 | 2 670 422 299 | 2 844 000 504 | 3 209 286 105 |
| Total number of chromosomes and organelles | 31 | 31 | 24 | 25 |
| Genes and pseudogenes | 26 001 | 26 740 | 39 480 | 41 722 |
| Protein-coding | 20 782 | 19 994 | 19 691 | 20 246 |
| Non-coding | 1677 | 3825 | 8555 | 9153 |
| Pseudogenes | 3542 | 797 | 11 234 | 12 323 |
| Genes with variants | 6158 | 2581 | 9563 | 14 632 |
| mtDNA size | 16 319 | 16 338 | | 16 569 |

**Table 5** Summary statistics for SNPs and indels found in Bison_UMD1.0 (Templeton) compared with domestic cattle reference (UMD3.1).

| | SNPs | Indels |
|---|---|---|
| Number of lines (input file) | 28 443 364 | 2 598 155 |
| Number of variants (before filter) | 28 483 599 | 2 627 645 |
| Homozygous for variant allele | 22 073 944 | 2 208 623 |
| Heterozygous (one reference one variant) | 6 329 185 | 360 038 |
| Reference alleles | 6 329 185 | 360 038 |
| Number of multi-allelic VCF entries | 40 235 | 29 494 |
| Number of effects | 32 086 858 | 2 976 475 |
| Genome total length | 2 670 424 944 | 2 670 423 585 |
| Genome effective length | 2 660 909 050 | 2 660 907 691 |
| Variant rate | 1 variant every 93 bases | 1 variant every 1012 bases |

placements, synteny block assigned, scaffold start and end position, and domestic cattle start and end position. Chromosome 1 was found to have the most scaffolds mapped to it with 30 synteny blocks anchored, whereas chromosome 26 was found to have the least amount ($n = 6$) of scaffolds placed on it (Fig. 2a). There were also a low number of bison scaffolds placed on the domestic cattle X-chromosome. Synteny blocks (in black) anchored to

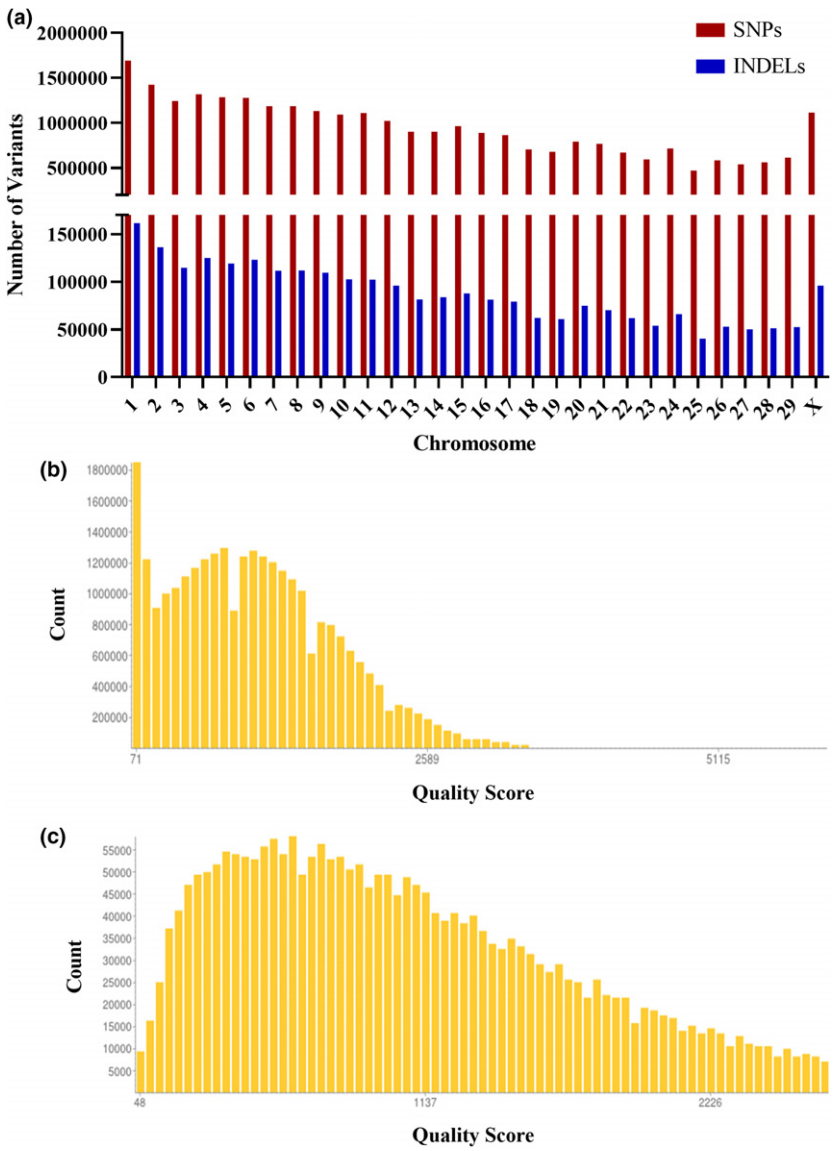


**Figure 1** Variant SNPs and indels) counts and quality scores of the bison reference to domestic cattle reference. (a) SNP and indels found for each chromosome from Bison_UMD1.0 (Templeton) aligned to domestic cattle. (b) Quality scores of annotated SNPs for Bison_UMD1.0 to Cattle_UMD3.1 (c) Quality scores of annotated indels for Bison_UMD1.0 to Cattle_UMD3.1.

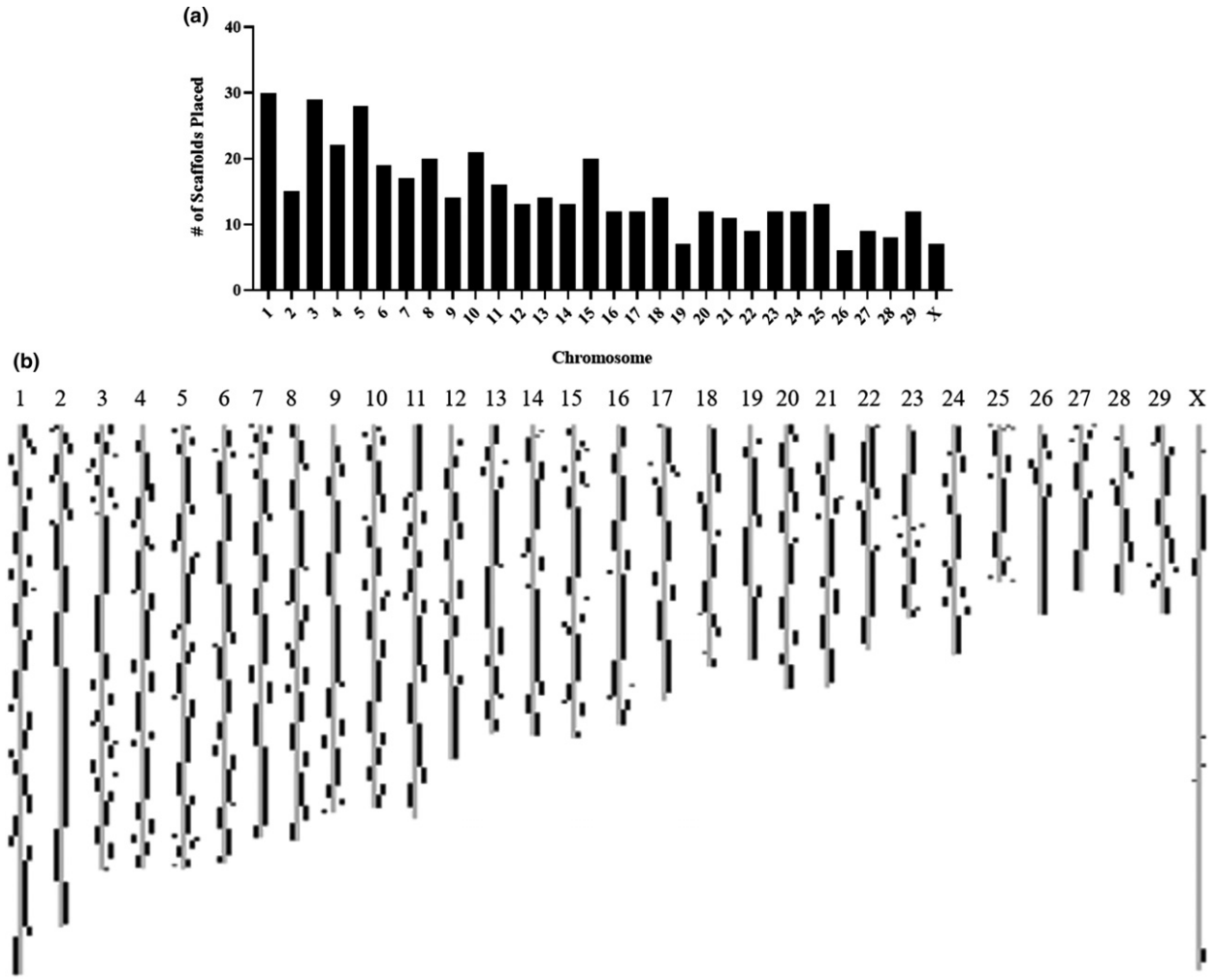


**Figure 2** Bison synteny to domestic cattle UMD3.1.76. (a) Bison scaffold placement on chromosomes from SyMap with number of bison scaffolds placed on each domestic cattle chromosome. (b) Bison anchor placement on domestic cattle chromosomes. Black anchors are those scaffolds that were found to have synteny with domestic cattle.

domestic cattle (in grey) for all chromosomes can be found in Fig. 2b. Even though these are different species they contain similar chromosomal arrangements and gene placement throughout their genomes.

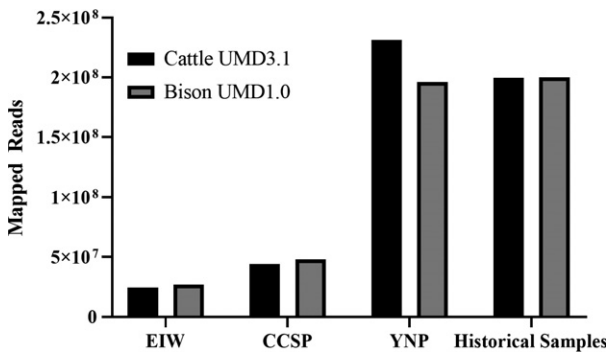


**Figure 3** SAMtools flagstat statistics of the 14 reseqenced bison mapped reads to domestic cattle (Cattle UMD3.1) and bison (Bison UMD1.0/Templeton) reference sequences (reads are in bp).

## Variant identification in relation to the domestic cattle and bison reference genomes for re-sequenced bison

Raw Illumina paired-end sequences for the 14 resequenced bison were individually mapped to the domestic cattle UMD3.1 sequence. Using the flagstat option in SAMtools, statistics of the 14 resequenced bison mapped to cattle_UMD3.1 and Bison_UMD1.0 (Templeton) reference sequences were obtained, and the mapped reads based on population to both can be found in Fig. 3. Figure 3 shows the amount of sequencing variation between individuals based on the sequencing methods used. Therefore, the analysis for variants was grouped together to consider the variants within populations for both references. All variants are summarized in Table 6, showing counts for SNPs, indels and total variants found for CCSP, EIW, YNP, the individual historical sequences, S6 and S9, to both references. Bison_UMD1.0 (Templeton) was added to the domestic cattle comparison.

**Table 6** Variant summary (SNP, insertion and deletion) for populations and individuals to either the domestic cattle (UMD3.1) or bison (UMD1.0/Templeton) reference sequences.

| | Variant type | | | |
|---|---|---|---|---|
| | SNP | Insertion | Deletion | Total |
| CCSP | | | | |
| UMD3.1 | 15 617 914 | 55 773 | 61 535 | 15 735 222 |
| UMD1.0 | 3 877 737 | 14 769 | 13 683 | 3 906 189 |
| EIW | | | | |
| UMD3.1 | 9 590 819 | 22 994 | 24 532 | 9 638 345 |
| UMD1.0 | 2 192 618 | 6408 | 5593 | 2 204 619 |
| YNP | | | | |
| UMD3.1 | 30 538 894 | 1 101 381 | 1 230 864 | 32 871 139 |
| UMD1.0 | 9 157 950 | 208 771 | 202 350 | 9 569 071 |
| S6 | | | | |
| UMD3.1 | 24 955 527 | 385 125 | 456 563 | 25 797 215 |
| UMD1.0 | 11 857 832 | 112 949 | 134 501 | 12 105 282 |
| S9 | | | | |
| UMD3.1 | 16 951 692 | 162 921 | 226 079 | 17 340 692 |
| UMD1.0 | 6 635 219 | 35 791 | 49 406 | 6 720 416 |
| Templeton | | | | |
| UMD3.1 | 22 073 944 | 1 233 140 | 1 394 505 | 24 701 589 |

CCSP, Caprock Canyons State Park; EIW, Elk Island National Park; YNP, Yellowstone National Park.

There were a total of 50 746 586 variants found between these 15 bison and domestic cattle reference, with 47 514 082 SNPs, 1 492 303 insertions and 1 740 200 deletions. All variants to domestic cattle refence were annotated to give biological functions of the genes associated with the variants, and most of the variants were associated for protein-coding genes (Table S4).

As these populations were representing different subspecies of bison, populations were analyzed for informative variants within their representative populations (Table 7). Variants were determined to be informative if all of the samples in a population shared those variants. These unique variants between populations can be used to verify the taxonomic status of these bison populations. However, future validation still needs to be done for the variants reported in this research.

In order to annotate the detected SNPs of the 14 samples above, the scaffolds in the VCF file were replaced by a chromosome number. Using the previously produced SyMap pseudochromosomes, the bison scaffolds used for the alignment were anchored to positions on respective chromosomes. This allowed for the scaffolds in the combined VCF files for each population, or in the case of the historical samples individually, to be replaced by chromosome numbers based on the positions created in the SyMap anchor file and using a perl script (Appendix S2). The changed VCF files were then annotated in SnpEff.

When comparing the identified variants for each population from either using the VCF files that contained Templeton's scaffolds or Templeton's pseudo-chromosomes, a reduced number of variants detected can be seen (Table S5). So as not to exclude the variants detected to Bison_UMD1.0 (Templeton) for each population, the variants detected for both Templeton's scaffolds (previous analysis results) and pseudo-chromosomes were annotated. The annotated variants followed the same trend as the variants identified to the domestic cattle reference, where the majority of the genes annotated were protein-coding genes (Table S6).

## Phylogenetic tree

Using the combined VCF file to domestic cattle for all 15 bison samples, SNPHYLO was used to generate a phylogenetic tree (Fig. 4). Based on this tree, the historic samples were placed next to the YNP samples, with the YNP samples first and then Templeton (bison) and the CCSP samples also placed close to the historic samples. Historic samples were collected in what would become YNP, so this placement was expected. A split between EIW and the historic samples was seen as the EIW samples were placed next to the YNP samples, which with the moving of YNP bison into EIW, could be expected. What was also expected is the split between the CCSP and EIW populations as they represent the southern plains and wood buffalo populations. As these are three different subpopulations being compared, it is interesting that the samples that represent southern plains bison comprise the farthest bison population from the EIW population. What was not expected was to see a split between one of the CCSP samples and the others. The placement of CCSP 50-5792 within the EIW samples will

**Table 7** Number of common (informative) variants to bison reference (Bison_UD1.0/Templeton) found between all individuals by population.

| | CCSP | | EIW | | YNP | | Historic samples | |
|---|---|---|---|---|---|---|---|---|
| Shared between | SNPs | Indels | SNPs | Indels | SNPs | Indels | SNPs | Indels |
| 4 | 103 125 | 1394 | 21 683 | 496 | 741 721 | | | |
| 1 | 2 577 386 | 24 106 | 1 879 071 | 10 109 | 3 834 283 | 292 466 | 17 023 108 | 308 559 |
| 3 | 289 403 | 749 | 37 500 | 400 | 1 969 047 | 22 802 | | |
| 2 | 902 866 | 2084 | 251 115 | 926 | 2 612 899 | 93 107 | 733 561 | 11 633 |
| Count | 3 872 780 | 28 333 | 2 189 369 | 11 931 | 9 157 950 | 408 375 | 17 756 669 | 320 192 |

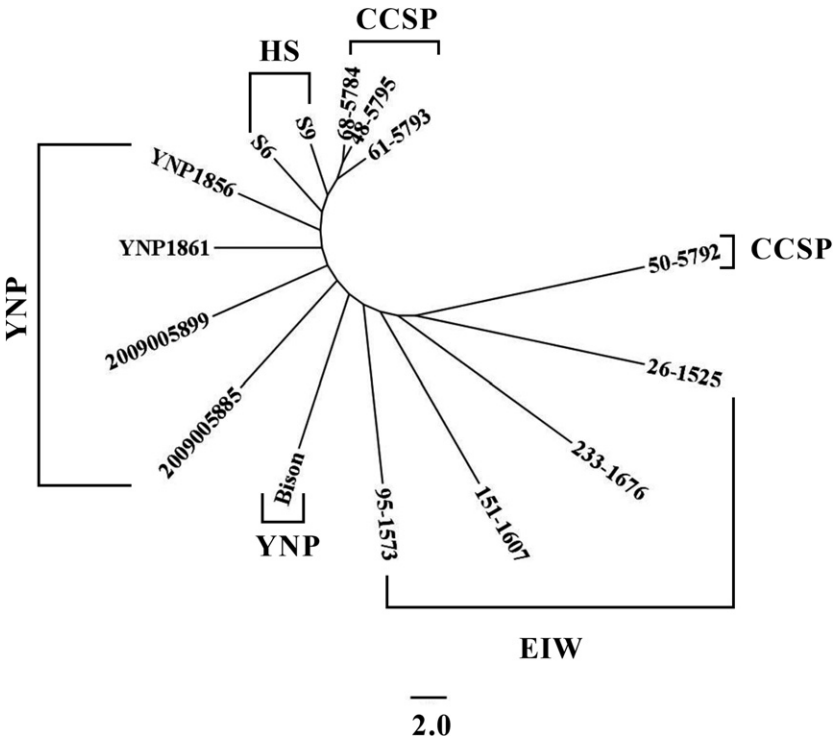


**Figure 4** Phylogenetic tree from the combined VCF file for all 15 bison samples to UMD3.1. Bison = Templeton. Population abbreviations: CCSP, Caprock Canyons State Park; EIW, Elk Island National Park; YNP, Yellowstone National Park; HS, Historic Samples.

require further analysis to evaluate this split of the CCSP bison samples. The phylogenetic tree offers a rough estimate of where subpopulations can be placed based on variant calls using whole-genome sequencing and read mapping. This method has moved bison phylogeny into genomics such as: distance trees, ABBA/BABA analysis and admixture events that have recently utilized bison genomic sequencing (Wang *et al.*, 2018; Zhang *et al.*, 2020).

## Discussion

The history and restoration of the North American bison is considered one of the first conservation success stories and a model of natural resource conservation (Ward 2000). With the completion of the first *de novo* reference assembly of the American bison genome, bison genetic research has now advanced into the genomic technology era. This study utilized technologies that are currently available and can be compared with new and improved genomic technologies to improve our dataset.

With the annotation of the bison *de novo* reference genome we were able to identify a total of 26 001 genes and pseudogenes with 20 782 genes being protein-coding genes. The bison reference also provided a means of detecting new genetic variants, including SNPs and indels, following alignment to the domestic cattle reference. Using the UMD3.1 and 1000 Bulls genome allowed us to compare the bison sequences with multiple domestic cattle sequences instead of just one to give a more thorough evaluation of the genomic differences between these two species. This allowed for the detection of approximately 50 000 000 new variants (both SNPs and indels combined) between bison and domestic cattle, vastly expanding the number of variants that define the genomic differences between the two species.

This study utilized detected genetic variants to complete an annotation that determined gene functions and can identify the biological pathways they affect. These pathways can be investigated in the future to analyze the physical, biological and adaptive genomic differences between bison and cattle, specifically in their response to disease, weather and even nutrition. Identifying the genomic regions responsible for these differences can help researchers narrow their focus on candidate genes that control these responses for future research. For example, focusing on the brucellosis and tuberculosis status of some bison populations is imperative for bison populations to ensure their health for future generations as well as improving herd management practices. Identified genomic variants associated with disease or immunity could be obtained from healthy, affected and exposed animals to compare phenotypes and genotypes, allowing the production of better vaccines for bison diseases (Zwane *et al.*, 2019). Using genomic technology as part of vaccine development against diseases that are detrimental to bison can greatly improve the management and relationships of sympatric livestock and wildlife populations.

We were also able to identify genomic components that are similar between the bison and the domestic cattle reference genomes, identifying parts of the genome

stemming from a shared common ancestor. Ancestral parts of the genome could be determined between bison and domestic cattle using statistical analyses and then used to evaluate what parts of the bison genome could have come from introgression of domestic cattle, similarly to recent research done to evaluate the lineage of the wisent and how hybridization could have played a role (Wang *et al*., 2018). This shared genomic information also gave us the ability to use the domestic cattle reference to provide presumptive chromosomal assignments of the bison reference scaffolds. We were able to anchor these genes to 'pseudo-chromosomes' for bison using synteny blocks between the bison scaffolds and the domestic cattle chromosomes, which provided the location for unknown bison genes.

Providing whole-genome sequencing of historic bison samples allowed for comparison of pre-population bottleneck bison with modern bison. Comparing these historic bison sequences with the bison reference sequence, we were able to evaluate the ancestral alleles/genomic regions that have been conserved over time from the historical samples to the modern samples. Even with the ability to detect approximately 12 and 9 million genomic variants between the historic samples S6 and S9 respectively to the bison reference, the total percentage of the genome that has detected variants throughout is only 0.43 and 0.24% respectively. Therefore, the majority of the bison genome when modern and historic bison sequences are compared remains quite similar and has been conserved in these regions.

To ensure that conservation management of bison continues to move forward using the newly available resources, more bison populations need to be evaluated to determine the genomic importance of different bison herds. Following similar strategies from this study, we can evaluate multiple bison herds to assess the presence of domestic cattle genomics and determine the genetic diversity and uniqueness of these herds, which can act as candidates to establish new conservation satellite herds. Indeed, a recent study made use of our sequencing, and a similar strategy to identify additional SNOs and to create a SNP-based parentage test and subpopulation composition in the Canadian bison industry (Yang *et al*., 2020). Following similar criteria and additional samples to better understand the genomic relationships between bison herds will aid with the relocation of bison herds when parks or herds have reached their carrying capacity.

In this study, we have reconstructed the first WGS of the North American bison. This genome will be an important resource for future investigations into their ecology, evolution and conservation. Whole-genome sequencing has allowed us to greatly increase the genomic variant information between bison and domestic cattle to identify differences between them. The results from this study provide the foundation for bison genomic research, which future studies can expand upon and used for comparison as genomic technologies improve over the years.

## Acknowledgements

Sequencing was completed at ISU Sequencing facility, http://www.dna.iastate.edu/. Analysis was done using the Texas A&M Institute for Genome Sciences & Society High Performance Computing Cluster. The authors thank Kranti Konganti for computational support. We also thank Bob Lee for technical assistance and Derek Brickhart for computational support. TPLS was supported by the US Department of Agriculture under USDA-ARS Project Plan no. 3040-31000-100-00D. The mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendations or endorsement by the US Department of Agriculture. USDA is an equal opportunity provider and employer.

## Conflict of interest

The authors declare no conflict of interest associated with this research.

### Data Availability Statement

This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under accession no. JPYT00000000. The bison genome reference assembly can be found with the assembly accession no. GCF_000754665.1 and assembly name Bison_UMD1.0 at NCBI. The resequenced bison sample data have been deposited in NCBI within Bioproject PRJNA658430. The VCF files have been deposited into European Variation Archive under the accession.

## References

Boyd D.P. (2003) *'Conservation of North American bison: Status and recommendations', ProQuest Dissertations and Theses*, p.222.

Cabanettes F. & Klopp C. (2018) D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958.

Cingolani P., Platts A., Wang L.L., Coon M., Nguyen T., Wang L., Land S.J., Lu X. & Ruden D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92.

Coder G.D. (1975) *The National Movement to Preserve the American Buffalo in the United States and Canada Between 1880 and 1920*. Columbus, OH:The Ohio State University.

Danecek P., Auton A., Abecasis G. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics* **27**, 2156–8.

Dary D. (1989) *The Buffalo Book: The Full Saga of the American Animal*. Swallow Press/Ohio University Press, Chicago, IL.

Das P.J., Chowdhary B.P. & Raudsepp T. (2009) Characterization of the bovine pseudoautosomal region and comparison with sheep,

goat, and other mammalian pseudoautosomal regions. *Cytogenetic and Genome Research* **126**, 139–47.

Depristo M.A., Banks E., Poplin R. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491–501.

Derr J.N., Hedrick P.W., Halbert N.D., Plough L., Dobson L.K., King J., Duncan C., Hunter D.L., Cohen N.D. & Hedgecock D. (2012) Phenotypic effects of cattle mitochondrial DNA in American bison. *Conservation Biology* **26**, 1130–6.

Di Meo G.P., Perucatti A., Floriot S. *et al.* (2005) Chromosome evolution and improved cytogenetic maps of the Y chromosome in cattle, zebu, river buffalo, sheep and goat. *Chromosome Research* **13**, 349–55.

Douglas KC, Halbert ND, Kolenda C, Childers C, Hunter DL, Derr JN (2010) Complete mitochondrial DNA sequence analysis of Bison bison and bison-cattle hybrids: function and phylogeny. *Mitochondrion*. **11**, 166–75.

Aronesty, E. (2011). ea-utils : "Command-line tools for processing biological sequencing data". https://github.com/ExpressionAnalysis/ea-utils

Flicek, P. et al. (2014) 'Ensembl 2014.', *Nucleic acids research*, **42** (Database issue), pp. D749-55. https://doi.org/10.1093/nar/gkt1196

Forgacs D., Wallen R.L., Dobson L.K. & Derr J.N. (2016) Mitochondrial genome analysis reveals historical lineages in Yellowstone bison. *PLoS One* **11**, e0166081.

Freese C.H., Aune K.E., Boyd D.P. *et al.* (2007) Second chance for the plains bison. *Biological Conservation* **136**, 175–84.

Gnerre S., Lander E.S., Lindblad-Toh K. & Jaffe D.B. (2009) Assisted assembly: how to improve a de novo genome assembly by using related species. *Genome Biology* **10**, R88.

Goodnight C. (1912) My experience with bison hybrids. *Journal of Heredity* **III**, 197–9.

Halbert N.D. (2003) *The Utilization of Genetic Markers to Resolve Modern Management Issues in Historic Bison Populations: Implications for Species Conservation*. College Station, TX:Texas A&M University.

Halbert N.D. & Derr J.N. (2007) A comprehensive evaluation of cattle introgression into US federal bison herds. *Journal of Heredity* **98**, 1–12.

Halbert N.D., Ward T.J., Schnabel R.D., Taylor J.F. & Derr J.N. (2005) Conservation genomics: disequilibrium mapping of domestic cattle chromosomal segments in North American bison populations. *Molecular Ecology* **14**, 2343–62.

Lee T.H., Guo H., Wang X., Kim C. & Paterson A.H. (2014) SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* **15**, 162.

Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Available at: http://arxiv.org/abs/1303.3997 (Accessed: 23 July 2020).

Li H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* Edited by I. Birol. Oxford University Press, **34**, 3094–100.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G. & Durbin R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–9.

McDonald J.N. (1981) *North American Bison: Their Classification and Evolution*. University of California Press, Berkeley, CA.

McKenna A., Hanna M., Banks E. *et al.* (2010) 'The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–303.

Polziehn R.O., Strobeck C., Sheraton J. & Beech R. (1995) Bovine mtDNA discovered in North American bison populations. *Conservation Biology* **9**, 1638.

Sambrook J., Fritsch E.R. & Maniatis T. (1989) *Molecular Cloning*. Cold Spring Harbor Laboratory Press, New York, NY.

Schnabel R.D., Ward T.J. & Derr J.N. (2000) Validation of 15 microsatellites for parentage testing in North American bison, Bison bison and domestic cattle. *Animal Genetics* **31**, 360–6.

Seppey M., Manni M. & Zdobnov E.M. (2019) BUSCO: assessing genome assembly and annotation completeness. In: *Gene Prediction: Methods and Protocols* (Ed. by M. Kollmar), pp. 227–245. Springer New York, New York, NY.

Soderlund C., Nelson W., Shoemaker A. & Paterson A. (2006) SyMAP: a system for discovering and viewing syntenic regions of FPC maps. *Genome Research* **16**, 1159–68.

Thibaud-Nissen F., Souvorov A., Murphy T., DiCuccio M. & Kitts P. (2013) *Eukaryotic Genome Annotation Pipeline, The NCBI Handbook* [Internet], 2nd edn. National Center for Biotechnology Information (US). Available at: https://www.ncbi.nlm.nih.gov/sites/books/NBK169439/ (Accessed: 23 July 2020).

Van der Auwera G.A., Carneiro M.O., Hartl C. *et al.* (2013) From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics* **43**(Suppl 43), 11.10.1–11.10.33.

Verkaar E.L.C., Vervaecke H., Roden C., Romero Mendoza L., Barwegen M.W., Susilawati T., Nijman I.J. & Lenstra J.A. (2003) Paternally inherited markers in bovine hybrid populations. *Heredity* **91**, 565–9.

Wang K., Lenstra J.A., Liu L., Hu Q., Ma T., Qiu Q., Liu J. (2018) Incomplete lineage sorting rather than hybridization explains the inconsistent phylogeny of the wisent. *Communications Biology* **1**, 169. https://doi.org/10.1038/s42003-018-0176-6

Wang K., Wang L., Lenstra J.A. *et al.* (2017) The genome sequence of the wisent (*Bison bonasus*). *GigaScience* **6**, 1–5.

Ward T.J. (2000) *An Evaluation of the Outcome of Interspecific Hybridization Events Coincident With a Dramatic Demographic Decline in North American Bison*. College Station, TX:Texas A&M University.

Ward T.J., Bielawski J.P., Davis S.K., Templeton J.W. & Derr J.N. (1999) Identification of domestic cattle hybrids in wild cattle and bison species: a general approach using mtDNA markers and the parametric bootstrap. *Animal Conservation* **2**, 51–7.

Yang T., Miller M., Forgacs D., Derr J. & Stothard P. (2020) Development of SNP-based genomic tools for the Canadian Bison Industry: parentage verification and subspecies composition. *Frontiers in Genetics* **11**, 1–16.

Zhang K., Lenstra J.A., Zhang S., Liu W. & Liu J. (2020) Evolution and domestication of the *Bovini* species. *Animal Genetics* **51**, 637–57.

Zimin A.V., Marçais G., Puiu D., Roberts M., Salzberg S.L. & Yorke J.A. (2013) Genome analysis The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–77.

Zwane A.A., Schnabel R.D., Hoff J., Choudhury A., Makgahlela M.L., Maiwashe A., Van Marle-Koster E. & Taylor J.F. (2019)

Genome-wide SNP discovery in indigenous cattle breeds of South Africa. *Frontiers in Genetics* **10**, 1–16.

## Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Figure S1.** Alignment of European bison (wisent) and Bison_UMD1.0 (Templeton) assemblies.

**Figure S2.** Preliminary analysis for bison reference sample, Templeton.

**Figure S3.** Quality assessment of genomic DNA of historical bison samples.

**Table S1.** Sample information for bison used for sequencing.

**Table S2.** Partial BLAST report for historical sample S6.

**Table S3.** Bison reference scaffolds aligned to domestic cattle.

**Table S4 .** Biological functions of genes associated with annotated SNPs and indels in bison populations to domestic cattle reference (UMD3.1).

**Table S5 .** Comparison of variants identified for each population when analyzed from mapped reads to Templeton's scaffolds (S) or pseudo-chromosomes (C).

**Table S6 .** Biological functions of genes associated with annotated SNPs and indels in bison populations to bison reference (UMD1.0/Templeton).

**Appendix S1.** Supplemental methods for preliminary analysis, RNA purification, and resequencing sample collection and DNA extraction.

**Appendix S2.** Custom script used to make chromosome files for VCF files from mapped read alignment to Bison_UMD1.0 (Templeton) reference to annotate in SNPEFF.