Wisent genome assembly uncovers extended runs of homozygosity and a large deletion that inactivates the thyroid hormone responsive gene

Chiara Bortoluzzi[1], Xena Marie Mapel[1], Stefan Neuenschwander[2], Fredi Janett[3], Hubert Pausch[1, #], Alexander S. Leonard[1, #]

Affiliations: [1]Animal Genomics, Department of Environmental Systems Science, Universitätstrasse 2, 8092 Zürich, Switzerland; [2]Animal Genetics, Department of Environmental Systems Science, Tannenstrasse 1, 8092 Zürich, Switzerland; [3]Clinic of Reproductive Medicine, Vetsuisse Faculty, University of Zürich, Winterthurerstrasse 260, 8057, Zürich, Switzerland

[#]Corresponding authors: Hubert Pausch (hubert.pausch@usys.ethz.ch), Alexander S. Leonard (alexander.leonard@usys.ethz.ch)

Abstract

The wisent (*Bison bonasus*) is Europe's largest land mammal. We produced a HiFi read-based wisent assembly with a contig N50 value of 91 Mb containing 99.7% of BUSCO genes which improves contiguity a thousand-fold over an existing assembly. Extended runs of homozygosity in the wisent genome compromised the separation of the HiFi reads into parental-specific read sets, which resulted in inferior haplotype assemblies. A bovine super-pangenome built with assemblies from wisent, bison, gaur, yak, taurine and indicine cattle identified a 1,580 bp deletion removing the protein-coding sequence of *THRSP* encoding thyroid hormone-responsive protein from the wisent and bison genomes. Analysis of 725 sequenced samples across the Bovinae subfamily showed that the deletion is fixed in both *Bison* species but absent in *Bos* and *Bubalus*. The *THRSP* transcript is abundant in adipose, fat, liver, muscle, and mammary gland tissue of *Bos* and *Bubalus*, but absent in bison indicating that the deletion inactivates *THRSP* possibly contributing to low bison milk and meat fat content. We show that super-pangenomes can reveal potentially trait-associated variation across phylogenies, but also demonstrate that haplotype assemblies from species that went through population bottlenecks warrant scrutiny, as they may have accumulated long runs of homozygosity that complicate phasing.

Keywords: *Bison bonasus,* wisent, endangered species, long read sequencing, trio-binning, genome assembly, gene loss, pangenomics

Introduction

The wisent (*Bison bonasus*), also known as the European bison, is a member of the *Bovidae* family that contains several domesticated livestock species, such as cattle, buffalo, yak, sheep, and goat[1]. The wisent went extinct in the wild in 1921. A restoration program established in 1942 from 12 captive

individuals by the International Society for the Preservation of the European Bison led to the creation of the lowland and lowland-Caucasian lines, which are still managed as separate populations[2,3]. As a result of these efforts, the wild wisent population has expanded to around 6,800 free-roaming individuals across 10 countries[4]. However, the wisent is still considered a near threatened species according to the International Union for the Conservation of Nature[5].

Wisent genetic research had long relied on mitochondrial genome sequences and microsatellite markers[3,6,7]. More recent investigations into genome-wide genetic diversity in wisent have utilized a taurine cattle (*Bos taurus taurus*) reference genome and bovine microarrays[8–10]. Considering that wisent and taurine cattle diverged between 1.7 and 0.85 million years ago (MYA)[9], this methodological approach likely introduces reference bias[11,12]. Low contiguity (contig N50: 14.53 Kb; scaffold N50: 4.69 Mb) and high fragmentation (29,074 scaffolds) complicate the wider application of a short read-based wisent assembly[13] for genetic investigations[14], therefore assembling the wisent genome with more recent approaches is warranted. A highly contiguous genome assembly of the wisent is also of great interest for the Bovine Pangenome Consortium[15] to investigate signatures of domestication, natural and artificial selection in the genomes of divergent lineages of the *Bovinae* subfamily.

The semi-automated construction of highly contiguous, near complete, and near error-free assemblies is feasible with current sequencing and assembly methods[16]. Trio binning is a widely used assembly method when parent-offspring trios are accessible[17–21]. The availability of parental sequencing data allows to bin the offspring's sequencing reads into haplotype-specific sets based on *k*-mers specific to either the paternal or maternal haplotype, thereby assembling maternal and paternal haplotypes[17,18].

Here, we assemble a wisent genome with Pacific Biosciences (PacBio) HiFi reads. We produce two haplotype assemblies with trio binning and a primary assembly. We show that the wisent genome contains many extensive runs of homozygosity (ROH), which complicates phasing and results in incomplete haplotype-resolved assemblies. We integrate the primary wisent assembly into a bovine super-pangenome and identify a putatively trait-associated deletion of the entire protein-coding sequence of *THRSP* which is fixed in wisent and bison.

Results and Discussion

Assembly of the wisent genome

A draft reference of the wisent genome was previously assembled with short sequencing reads[13] but the assembly is highly fragmented and therefore not suitable for detailed pangenome analyses[22]. To make the wisent genome amenable to pangenome analyses, such as identifying structural variants differing between wisent and other wild and domesticated members of the *Bovinae* subfamily, we sampled tissue from a parent-offspring trio to assemble a wisent genome with long reads through trio binning. We collected 131.1 Gb (or approximately 44x coverage) of HiFi reads, with a mean read length of 19.3 Kb and a mean quality value of 32.4 from a male wisent (hereafter referred to as F1) from a

captive population. We further collected approximately 40x, 38x, and 36x coverage of Illumina short reads from the F1, his sire, and dam, respectively. We used the parental short reads to trio-bin the long reads, assigning an unknown/paternal/maternal tag to each read. Overall, the "binnability" of the wisent sample reads was 61.9%, which is low compared to the 79.9%, 85.0%, or 99.2% binnability observed for intra-breed (Braunvieh x Braunvieh)[23], inter-breed (Rätisches Grauvieh x Simmental)[24], or inter-subspecies (Nellore [indicine] x Brown Swiss [taurine])[19] HiFi-sequenced *Bos taurus* crosses (**Fig. 1a-d**). The F1 sample also had higher variability in binnability along the genome than observed previously in other bovines, suggesting that in this wisent trio there is less parental-specific sequence used to assign haplotypes, but also a more uneven distribution of such sequence (**Fig. 1e; Supplementary Fig. 1**).

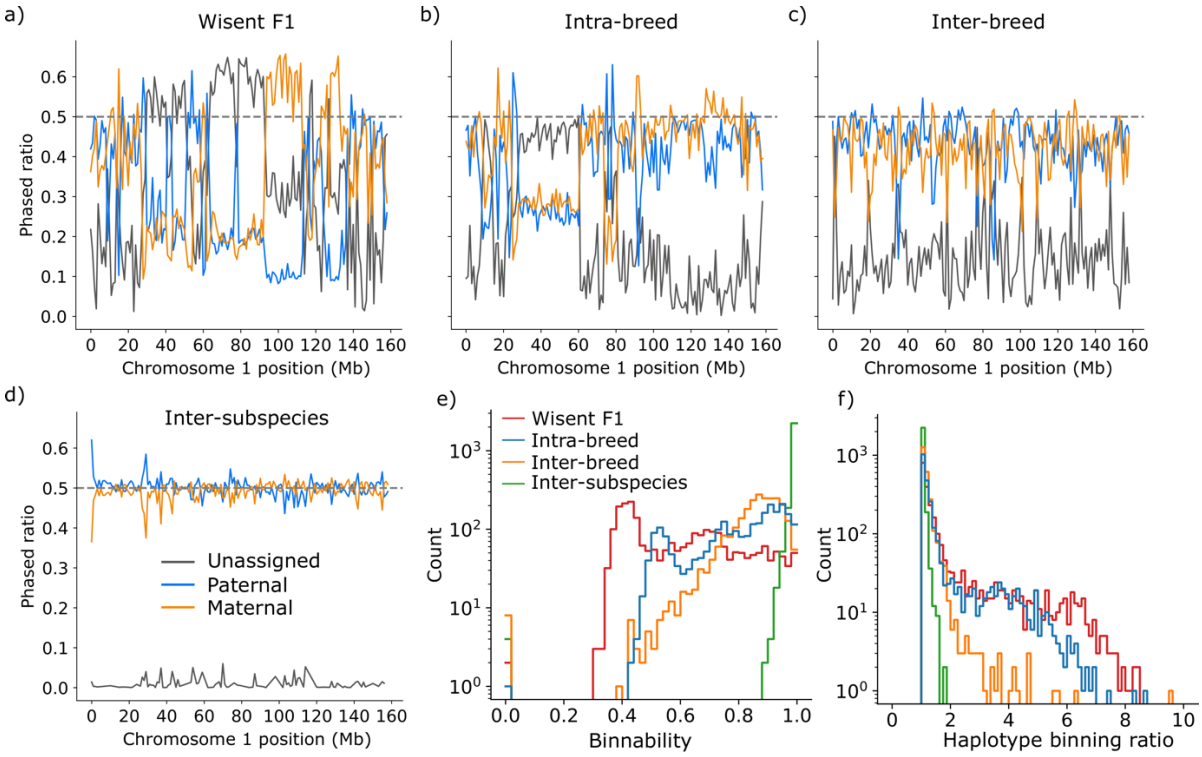

Figure 1. Trio binning of the wisent *(Bison bonasus)* genome. (**a-d**) Fraction of reads per 100 Kb window tagged as unassigned, paternal, or maternal haplotypes across chromosome 1 for the wisent sample, an intra-breed sample (BSWxBSW), and inter-breed sample (RGVxSIM), and an inter-subspecies sample (NELxBSW). (**e**) Histogram of proportion of parental-assigned reads (paternal + maternal)/total) per window across all autosomes. Proportions close to 1 indicate unambiguous phasing. (**f**) Histogram of paternal assigned read ratio (paternal/maternal if paternal > maternal else maternal/paternal) per window across all autosomes. Values close to 1 indicate balanced phasing, while higher values indicate either paternal or maternal reads are disproportionately assigned. Colours are taken from (**e**).

Regions with high proportions of unassigned reads can safely be treated as homozygous sequences present in both offspring haplotypes. However, regions with moderate levels of unassigned reads and high levels of reads assigned to a single haplotype can introduce an assembly gap in the unrepresented haplotype and likely lead to assembly errors - due to mixing of haplotypes - in the overrepresented haplotype. There was a dam/sire-assigned read bias greater than 5-fold for 213 100-Kb bins in the F1

sample, over three times more than observed in a previously analyzed *Bos taurus taurus* intra-breed (Braunvieh x Braunvieh) F1 sample with only 70 bins with large imbalances (**Fig. 1f**).

Given the low binnability and biased assignment of parental reads in the F1 sample, we used hifiasm to generate both a primary (collapsed) assembly and two haplotype-resolved assemblies, to examine the impact of assembler phasing assumptions. We calculated contiguity, correctness, and completeness for all three new assemblies, as well as the existing short-read based draft wisent reference genome[13]. Most metrics demonstrate the outstanding quality of our assemblies, with the primary assembly improving contiguity a thousand-fold over the existing draft wisent assembly (**Table 1**). As a result, both haplotypes and the primary assembly captured a higher fraction of unassembled genome, for an additional 0.3 Gb and 0.5 Gb, respectively, while reducing the number of missing (N) bases from 125 Mb in the draft to 0.05 Mb in the primary assembly**.** The overwhelming majority (93%; 2.86 Gb) of the primary assembly sequence was in scaffolds that aligned to the 29 autosomes and two sex chromosomes of the *Bos taurus taurus* reference sequence.

Table 1. Assembly statistics of the wisent draft assembly[13], the two haplotype-phased assemblies, and the primary assembly generated in this study.

		Draft	**Haplotype 1 (paternal)**	**Haplotype 2 (maternal)**	**Primary**
Assembly accession		(Wang et al. 2017)[13]	This study	This study	GCA_963879515.1 (this study)
Assembled genome size (Gb)		2.57	2.83	2.86	3.07
Number of contigs		243,242	486	480	248
Contig N50 (Mb)		0.02	74	59	91
Number of scaffolds		29,074	402	338	210
Scaffold N50 (Mb)		4.0	95	105	105
Quality Value		-	55.6	55.7	56.4
Largest scaffold (Mb)		31.6	164.4	164.1	163.9
Compleasm (%)	Single copy	98.4	93.4	88.4	98.5
	Duplicated	0.6	1.1	1.1	1.2
	Fragmented	0.7	0.3	0.3	0.2
	Missing	0.3	5.3	10.2	0.1

Assembly completeness was estimated using compleasm and the set of 9,226 highly conserved mammalian genes (hereafter called BUSCO genes). The draft and primary assembly had high completeness scores (99.0% vs 99.7%). While the contiguity and correctness of the haplotype-resolved assemblies were comparable to the primary assembly, the gene completeness was noticeably lower, especially for the maternal haplotype at 89.5%. Many of the BUSCO genes missing from one haplotype were found in either the other haplotype or the primary assembly, suggesting that the diploid sequence was incorrectly assigned to only a single haplotype. We confirmed the sporadically missing BUSCO genes should have been in large regions (> Mb) of missing syntenic sequence (**Supplementary Fig. 2**), rather than local mis-assemblies disrupting BUSCO identification.

Extended runs of homozygosity complicate haplotype phasing

The differences in sequence and gene content observed between the haplotype assemblies and the primary assembly were surprising, as we did not encounter such a pattern in other bovine assemblies constructed earlier through trio binning[19] and so prompted a detailed investigation. We looked at the distribution and location of regions with lower-than-expected heterozygosity (runs of homozygosity – ROH) by aligning short-reads of the F1 to the primary wisent assembly to examine whether excessive ROH had an impact on the assembly construction. Variants called from these alignments revealed that ROH covered a total of 1.38 Gb in the F1 genome corresponding to a genomic inbreeding coefficient (F_{ROH}) of 0.52 **(Fig. 2a; Supplementary Table 4).** Of these, 139 were longer than 2 Mb and tended to correspond with large regions of missing sequence in the haplotype assemblies, which appeared to derive from the hifiasm unitig graph, where unbalanced assignment of paternal phases lead to incorrect haplotype separation (**Supplementary Fig. 2**). Although many of the ROH were correctly assembled, it was striking to observe such a correspondence between regions of lower-than-expected heterozygosity and regions that are difficult to resolve in the assembly using a trio binning approach. Given the high binning variability of the wisent F1, using a bin-first-then-assemble method[18] encounters similar haplotype-resolved assembly issues (**Supplementary Table 5**), partially improving BUSCO score (although some previously missing loci are still missing, including the region from **Supplementary Fig. 2**), but considerably worsening the haplotype switch rate and contiguity.

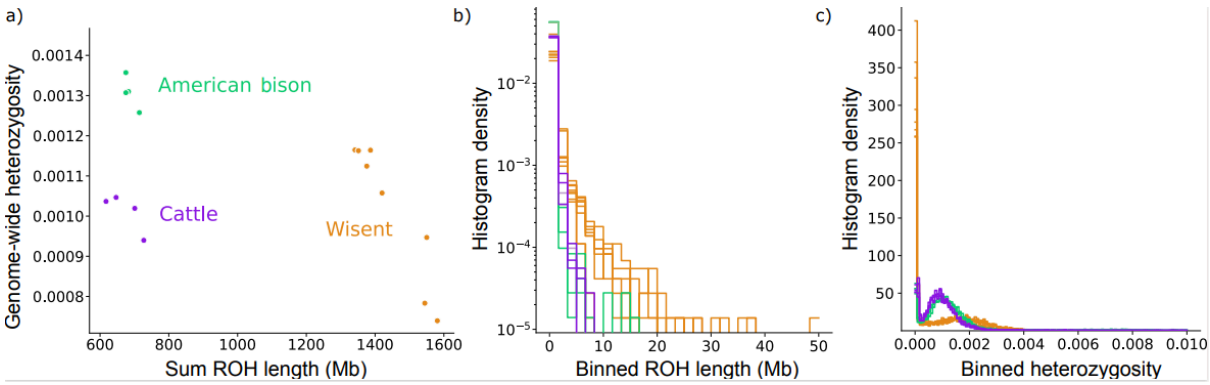

Figure 2. Genome-wide heterozygosity and runs of homozygosity. (a) Total length of runs of homozygosity (ROH) versus genome-wide heterozygosity in each wisent (*n* = 8), American bison (*n* = 4), and taurine cattle (*n* = 4) sample. **(b)** Histogram of ROH lengths derived from 10 Kb bins for the same samples from **(a)**, with substantially longer ROH present in wisent. **(c)** Histogram of heterozygosity in 1 Mb windows for the same samples from **(a)**, showing wisent have both many more low heterozygosity bins but also an increase in heterozygosity in some regions.

The genomic inbreeding of the F1 was exceptionally high compared to what is commonly observed in other bovines but was consistent with what we observed in other wisent samples. ROH covered a large fraction of the genome also in the seven (5 captive, 2 wild) additional wisents we investigated (**Fig. 2a; Supplementary Table 4**). On average, 1.44 Gb of their genomes were in ROH resulting in a F_{ROH} of 0.56. Between 121 and 168 ROH longer than 2 Mb were observed in the eight wisents studied. These long homozygous regions likely reflect founder effects resulting from the genetic bottleneck and expansion of the wisent population at the beginning of the 20th century[9]. Compared to the wisent

samples, we found fewer ROH in American bisons and taurine cattle covering only a quarter of their genomes (American bison: 662 Mb; taurine cattle: 665 Mb) (**Fig. 2a)**. ROH longer than 2 Mb were eight- and two-fold less abundant in American bison and cattle, respectively (**Fig. 2b)**. A relatively low F_{ROH} value in American bison is likely the result of hybridization with various domestic cattle breeds, which has been encouraged since the late 1800's, when most of the surviving bison individuals were maintained by cattle ranchers in private herds[25].

Genome-wide heterozygosity in the wisent ($1.04 \pm 0.16 \times 10^{-3}$) was similar to that of cattle ($1.04 \pm 0.22 \times 10^{-3}$) but lower than in American bison ($1.35 \pm 0.19 \times 10^{-3}$), which is in line with previous studies[8,9]. Heterozygosity varied strongly along the genome in wisent as evidenced by an excess of regions with almost no heterozygous sites (**Fig. 2c**). However, some genomic regions had higher heterozygosity in wisent than in cattle or American bison, reflecting their larger ancient effective population size[9]. These findings corroborate that genome-wide heterozygosity levels are unable to reflect demographic effects that can lead to extended segments of homozygosity and are therefore of limited utility to assess variability of populations[26]. Moreover, our findings emphasize that the average genome-wide heterozygosity can be a misleading metric to consider when conducting haplotype-resolved analyses, as long stretches of the genome may be homozygous and thus unable to be assigned into haplotypes, even for relatively "normal" genome-wide heterozygosity levels.

Repeat and gene content of the wisent assembly

The high accuracy of HiFi reads can benefit the assembly of repetitive sequence[27], and so we investigated the repeat content in the new wisent assembly. Since all examined metrics provide confidence that the primary wisent assembly is highly contiguous and near complete, we used it for all downstream analyses. Repetitive elements were identified and classified using a wisent-specific *de novo* repeat library constructed with RepeatModeler. This approach showed that 48.90% of the wisent genome are repetitive sequence, which was similar to that of the draft assembly (47.30%), American bison (43.52%), and domestic taurine cattle (41.73%) when equally using species-specific *de novo* repeat libraries (**Supplementary Table 6**). Within the first class of transposable elements (TEs), also called retrotransposons, LINEs (27.76%) were the most abundant type, which is also in agreement with their high prevalence in the bovine genome[28], followed by LTRs (4.44%) and SINEs (3.82%). We found that substantially more satellite DNA was in the HiFi-based wisent (9.69%) than in the American bison (2.04%) and cattle (0.07%) assemblies. Unclassified repeats made up just 0.85% of the wisent repeat content. While this value was similar to that of the American bison (0.88%), it was 11-fold larger than in domestic cattle (0.08%). We ran another round of repeat masking on the American bison and cattle genome using the wisent-specific repeat library generated with RepeatModeler to test the impact of the repeat library on the identification and classification of repetitive sequence. We found that the overall repeat content did not change substantially when using either a species-specific (American bison: 43.52%; cattle: 41.73%) or a wisent-specific repeat library (American bison: 43.27%; cattle: 41.90%), suggesting that *de novo* repeat libraries are not sensitive enough to identify *de novo* repeats in highly

similar genomes (1% divergence between wisent and bison and wisent and cattle) due to the levels of noise **(Supplementary Table 6**).

We performed a Kimura distance-based copy divergence analysis of TEs in the wisent assembly to estimate the age of TEs. We observed a predominance of young LINEs and LTRs, as evident from their clustering on the left side of the graph, which indicates minimal deviation from the consensus sequence (**Fig. 3a)**. Additionally, the wisent assembly contained unidentified young repeat copies, classified as "unknown" in the graph. LINEs and LTRs were also the most abundant type of ancient or degenerated TEs, as indicated by their clustering on the right side of the graph.

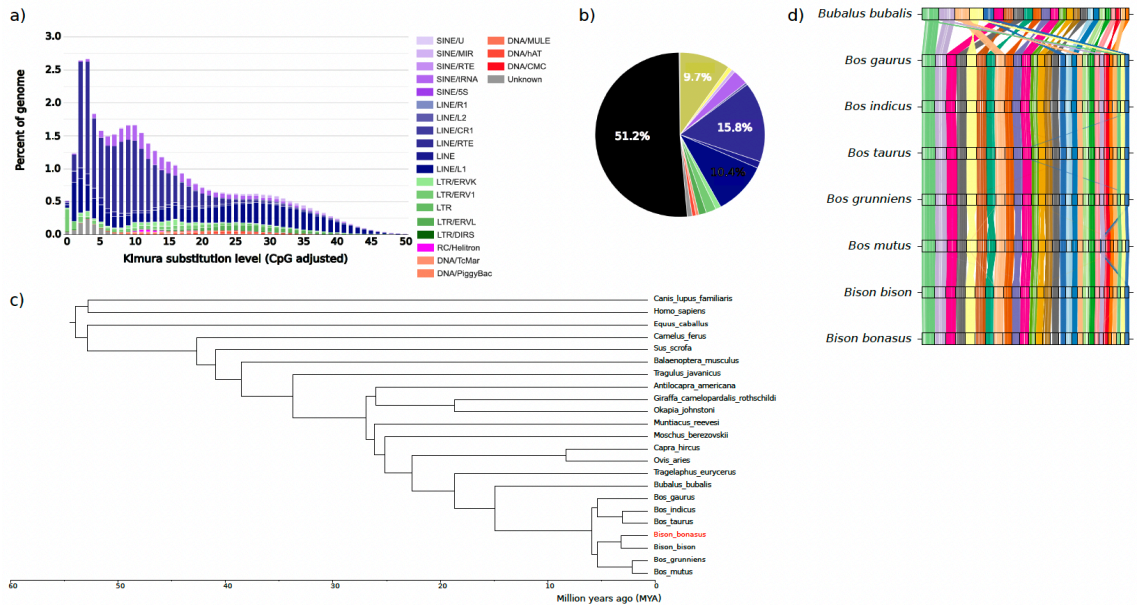

Figure 3. Primary assembly of the wisent (*Bison bonasus*). (a) Kimura substitution levels between the repeat consensus and its copies. The histogram plot shows the age distribution of transposable elements (TEs). The total amount of DNA in each TE class was split into bins of 5% Kimura divergence. (**b)** Pie chart representing the percentage of the genome in different repeats. Repeats are colored following the legend in (a), whereas the % of unmasked genome is in black. **(c)** Phylogenetic tree constructed from single copy BUSCO genes identified in compleasm. The axis shows the divergence time in million years ago (MYA). All nodes were supported in more than 94% of bootstrapped iterations. **(d)** Synteny plot showing the conservation of large-scale gene linkage and gene order across 8 species. The conserved unique single copy BUSCO genes are connected by lines according to their chromosomal location. Chromosomes are ordered by total size, from the largest to the smallest. Sex chromosomes and the mitochondrial genome are excluded from this analysis.

We used BUSCO genes shared between the wisent and 22 other species to build a phylogenetic tree, which was then used to estimate divergence times (**Fig. 3c**). The topology and divergence estimates of the *Bovini* subset of the tree are in line with a phylogenetic reconstruction from nuclear whole-genome sequences[29]. The wisent formed a clade with the American bison after diverging from this species approximately 2 MYA. This clade was sister to the domestic (*Bos grunniens*) and wild yak (*Bos mutus*) clade, from which it diverged approximately 5 MYA. The domestic taurine (*Bos taurus taurus*) and indicine (*Bos taurus indicus*) cattle formed a clade of their own, with the gaur (*Bos gaurus*) acting as an outgroup, corroborating an earlier phylogenic reconstruction of the Bovinae subfamily[30]. As expected,

the water buffalo (*Bubalus bubalis*) was the most distantly related species within the subfamily, diverging from the other *Bos* and *Bison* species approximately 15 MYA. BUSCO genes were also used to perform a synteny analysis to investigate how gene order has changed during the evolution of species within the *Bovinae* subfamily. We found gene order to be conserved between wisent, bison and various *Bos* species but, as expected, gene order is less conserved when compared to distantly related species, such as the water buffalo **(Fig. 3d).**

Pangenome analysis reveals a *Bison*-specific deletion inactivating *THRSP*

We then built per-autosome super-pangenomes with the five *Bos* and two *Bison* assemblies, excluding *Bubalus bubalis* due to the different assembled karyotype. Although *Bos gaurus* has a Robertsonian translocation between chromosomes 1 and 29[31], this genome was not assembled through the centromere fusion, leaving 29 separate assembled autosomes and so was included. We assessed the structural variant (SV) diversity (**Fig. 4a**), finding the wisent sample contains 74,770 SVs (insertions: 37,814, deletions: 36,956) relative to *Bos taurus taurus*, matching previous findings for other distantly related bovids[32,33]. We find many SVs private to either the American bison or wisent, contrasting to wild or domestic yak which have fewer private SVs, as expected given the more recent split of yak. We find a pronounced increase of SVs private to wisent on chromosome 7 between 10 and 10.6 Mb (27 times the genome-wide rate of private wisent SVs). Many of the private SVs were near or overlapping genes in the olfactory receptor 7 subfamily A (*OR7A*), with a total of 12 annotated protein coding genes in this region (**Supplementary Figure 3**), suggesting wisent may have unique variants mediating olfaction compared to the other bovids considered here. Using pairwise overlap of SVs, we can infer a relationship tree (**Fig. 4b**), closely matching the more rigorously constructed BUSCO gene-based phylogeny discussed previously (**Fig. 3c**), demonstrating that SVs also reflect evolutionary histories and are a rich source of variation to analyze.

We find 12,217 SVs uniquely common to *Bison* (i.e., American bison and wisent), including 96 that are predicted to have a high impact on proteins (**Supplementary File 1**). Among them, a 1,580 bp deletion which entirely overlaps the coding sequence of *THRSP* encoding thyroid hormone-responsive protein. *THRSP* has two exons, of which the first contains protein-coding sequence and the second is non-coding. The deletion is predicted to remove the protein-coding exon of *THRSP* in the American bison and wisent haplotypes that were integrated into the super-pangenome (**Fig. 4c**, **Supplementary Fig 4a**). This deletion occurs in the homozygous state in all short read-sequenced bisons (n = 19) and wisents (n = 20) we investigated, indicating it is likely fixed in both *Bison* species (**Fig. 4d**, **Supplementary Fig 4b**). The 1,580 bp deletion was neither detected in the *Bos taurus*, *Bos gaurus*, *Bos mutus* and *Bos grunniens* assemblies that were part of the super-pangenome, nor in any of the short-read sequenced *Bos taurus taurus* and *Bos taurus indicus* (n = 674) and *Bubalus bubalis* (n = 12) samples. As such, the deletion likely occurred in a common ancestor of bison and wisent, after divergence from the other species of the *Bovinae* subfamily. Genetic drift or selective advantages may have led to the fixation of the deletion in wisent and bison.

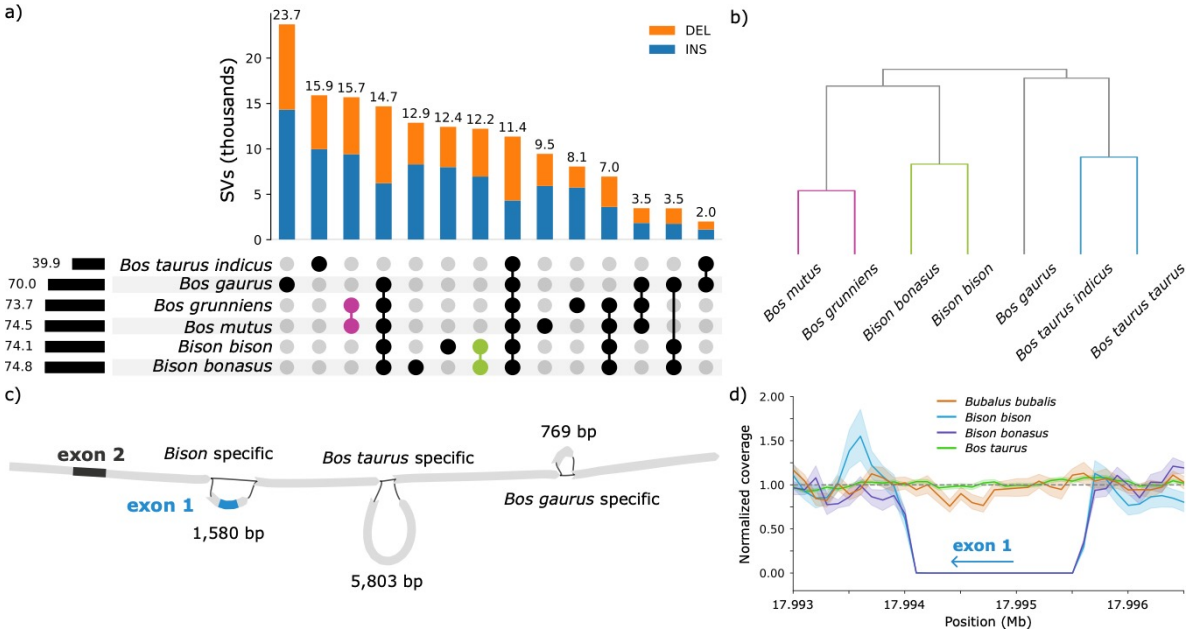

Figure 4. Structural variant analysis from a seven-assembly super-pangenome. **(a)** UpSet plot of SVs called from the 29 autosomes, where the total number of SVs per assembly is shown on the left and the number of intersecting SVs shown above each grouping. The pink and green markers highlight SVs private to both yaks or bison. **(b)** Relationship tree inferred from pairwise overlaps of SVs. Assemblies are correctly grouped into yak (purple), bison (green), and cattle (blue) clades. **(c)** A 1,580 bp deletion that includes the first coding exon of the *THRSP* gene was detected in the wisent and bison assemblies. **(d)** Normalized coverage of short read-sequenced data in wisent (n = 20), bison (n = 19), cattle (n = 674), and water buffalo (n = 12) samples with at least 5-fold coverage around the 1,580 bp deletion. Error bars represent the 95% confidence intervals, and the dashed line indicates the expected normalized coverage of 1.

We lifted the *THRSP* gene annotation from the *Bos taurus taurus* assembly (ARS-UCD1.2) to the haplotype-resolved, primary, and draft assemblies of wisent, finding alignment of the non-coding second exon in the paternal haplotype and primary assembly. We confirm that the partial *THRSP* gene is contained within a previously reported synteny group[34] corroborating this region is evolutionarily highly conserved and likely under similar transcriptional regulation and function across species. As expected, given the deletion uncovered from the pangenome, and in line with a partial alignment of *THRSP* in a highly contiguous American bison assembly[35] as well as the missing *THRSP* gene in the highly fragmented American bison reference assembly (GenBank accession: GCA_000754665.1)[36], the coding exon was missing in the primary wisent assembly. The liftover of the coding and non-coding *THRSP* exons was not successful for the previous draft assembly and the maternal haplotype assembly, demonstrating the utility of our near-complete and highly contiguous primary assembly for genomic investigations.

The amino acid sequence of *THRSP* is evolutionarily highly conserved (**Supplementary Fig. 4d)**. The expression of *THRSP* varies across tissues but it is elevated in tissues that synthesize fatty acids[37]. Comprehensive transcriptomic data (n=8642) from the cattle Genotype-Tissue Expression (GTEx) project[38] confirm that *THRSP* is highly expressed in adipose tissue (5351 TPM), intramuscular fat (66

TPM), liver (34 TPM), muscle (24 TPM) and lactating mammary gland (14 TPM) of cattle (**Supplementary Fig. 4**). We hypothesized that deletion of the first exon including the entire protein-coding sequence in wisent and bison represents a functional knock-out of *THRSP*. Transcriptomic data are not available for wisent but bison RNA-seq data from a three years old cow are publicly available for liver, spleen, lung, skeletal muscle, kidney and supramammary lymph node tissues[36]. We mapped the bison transcriptomes to the *Bos taurus taurus* reference sequence and compared gene expression with age- and sex-matched bovine samples from cattle GTEx for liver and muscle, i.e., two tissues with high *THRSP* expression and a decent number of informative GTEx samples (n_{liver}=14; n_{muscle}=43). The Spearman correlation coefficient estimated for 17,150 genes was 0.876 and 0.878 for liver and muscle tissue, respectively, indicating that overall gene expression levels in these tissues correlate well between bison and cattle. However, as expected given the deletion, we did not detect expression of the coding exon of *THRSP* in any of the bison tissues (**Supplementary Fig. 4f**). Interestingly, the non-coding second exon was also not expressed in any of the bison tissues, although it is not affected by the deletion. We then aligned 73 transcriptomic data from 19 tissues from 4 water buffaloes[39] against the *Bos taurus taurus* reference sequence. Water buffalo is substantially more diverged from cattle than wisent and bison, but the expression profile of both *THRSP* exons was similar to cattle with the highest transcript abundance in adipose tissue (9,776 TPM), mammary gland (1,570 TPM), skin (440 TPM), liver (298 TPM), and skeletal muscle (37 TPM) (**Supplementary Fig. 4f)**. Collectively, these findings show that the non-coding exon doesn't produce mRNA in bison which supports that the deletion of the coding first exon inactivates *THRSP* and that bison and wisent are lacking the thyroid hormone-responsive protein.

Given the crucial contribution of THRSP in lipogenesis and fatty acid synthesis in the mammary gland and other tissues[37,40,41], we suspect that lack of THRSP impacts lipid metabolism in the two *Bison* species. Mice lacking THRSP produce milk with significantly less medium-chain fatty acids resulting from a decreased lipogenesis in the mammary gland[41]. Neither bison nor wisent have been domesticated, and so the composition of their milk has not been investigated. However, bison meat has lower fat than beef[42,43] which agrees with reduced accumulation of fat in adult *THRSP* knockout mice[44]. While *THRSP* mRNA expression in skeletal muscle tissue is correlated with intramuscular fat content in crossbred cattle[45], the precise function of THRSP in the deposition of intramuscular fat remains to be elucidated[46]. Bison and wisent appear as intriguing model organisms to study the impact of missing THRSP on transcriptional changes in lipid metabolism pathways and to investigate a possible causal relationship between *THRSP* expression and fat accumulation. Hybridization between *Bison* and domestic cattle which have a functional *THRSP* gene is relatively common[47], and so the phenotypic and genetic diversity of the offspring can be exploited to investigate functional consequences arising from lack of THRSP in natural knockouts.

Materials and Methods

Ethics statement

No animals were sampled for this study. No ethics approval was required for this study.

Sample selection

Blood samples of six captive wisents were provided by the Bern animal park and Langenberg animal park in Switzerland, respectively **(Supplementary Table 1).** Blood samples of five wisents were collected prior to our study for the purpose of establishing a biobank at the animal park. Another blood sample was collected from one wisent after it was killed. The decision to kill the animal was independent from our study. High-molecular weight DNA was extracted from blood using the Qiagen MagAttract HMW DNA Kit.

Long-read (Pacific Biosciences) and short-read (Illumina) sequencing

A DNA sample from a male wisent (F1) was used for PacBio HiFi sequencing. DNA fragment length and quality were assessed with the Femto Pulse system (Agilent). A HiFi library was prepared and sequenced on three SMRT cells 8M with a Sequel IIe. We also used an Illumina NovaSeq 6000 machine to generate paired end (2 x 150 bp) reads from DNA extracted from all wisents including the F1 and its parents.

Genome assembly

The wisent F1 HiFi reads were assembled with hifiasm v0.19.5[17] using default parameters and "-a 5 -n 5 --primary" to produce a primary assembly. We reran hifiasm without the primary flag and instead provided parental *k*-mer databases that were created with yak v0.1 (https://github.com/lh3/yak) from parental short reads to produce two haplotype-resolved assemblies. The contig outputs were then scaffolded to the ARS-UCD2.0 *Bos taurus taurus* reference genome (GenBank assembly accession: GCA_002263795.4) with RagTag v2.1.0[48] primarily to orient and assign chromosome identifiers.

Assembly validation and completeness

We assessed assembly base-level quality and phasing blocks with merqury (6b5405)[49] using meryl (https://github.com/marbl/meryl) *k*-mer databases created from parental and offspring short read sequencing. We used calN50.js (https://github.com/lh3/calN50) to evaluate the contig contiguity. We ran compleasm v0.2[50] using the *mammalia_odb10* database, which contains 9,226 highly conserved genes from 24 species. The hifiasm-processed unitig graph was visualized with bandageNG v2022.9[51].

Analysis of haplotype specific F1 reads

We used canu v2.2[52] to assign paternal/maternal/unknown haplotypes to each read, using parental *k*-mer databases created from Illumina reads with meryl v1.3. Tagged reads were then aligned with minimap2 v2.26[53] using the map-hifi preset to the ARS-UCD2.0 *Bos taurus taurus* reference genome. The alignments were sorted with SAMtools v1.19.2[54] and the starting coordinate for every primary alignment was recorded for each haplotype tag. We used a 100 Kb window to bin the read counts for each haplotype across the autosomes. We estimated *k*-mer distributions by creating *k*-mer databases

and histograms with meryl v1.3 from each individual's HiFi reads. Sample information is given in **Supplementary Table 2**.

Repetitive sequence analysis

We generated a wisent-specific *de novo* repeat library for the primary assembly using RepeatModeler v2.0.4[55] to identify, classify, and mask repetitive elements. RepeatModeler was run in combination with RECON v1.08[56], RepeatScout v1.0.6[57], and Tandem Repeats Finder v4.09.1[58]. The complete Dfam v3.7[59] and RepBase (final version 10/26/2018)[60] libraries were used to classify repetitive elements based on homology in different repeat families. Consensus sequences obtained from RepeatModeler were used to softmask the genome with RepeatMasker v4.1.5 (https://www.repeatmasker.org/). We ran RepeatMasker by specifying the *rmblast* search engine and the slow search mode. We used the *calcDivergenceFromAlign.pl* script provided by RepeatMasker to summarize the Kimura substitution levels between the repeat consensus and its copies. Repeat landscape plots were produced with the *createRepeatLandscape.pl* script bundled in RepeatMasker. We applied this approach also to the American bison (*Bison bison*) (GenBank assembly accession: GCA_030254855.1) and taurine cattle (*Bos taurus taurus*) (GenBank assembly accession: GCA_002263795.3) assemblies for comparison purposes.

Phylogenetic tree construction

We constructed a phylogenetic tree from complete single copy orthologs identified by compleasm v0.2[61]. We ran compleasm using the *mammalia_odb10* database on an additional set of 22 species (**Supplementary Table 3**). We aligned the protein sequence of the 8,392 complete single copy genes shared between wisent and 22 other species using MAFFT v7.490[62]. Protein sequence alignments were trimmed using trimAl v1.4[63], specifying a gap threshold of 0.8 and a minimum average similarity of 0.001. Trimmed protein sequences were concatenated to form a supermatrix, which was provided to RaxML v8.2.12[64] to reconstruct a maximum likelihood phylogeny. RaxML was run with the PROTGAMMAJTT model and 1,000 bootstrap replicates. Divergence times were estimated in r8s[65] (https://github.com/iTaxoTools/pyr8s) using fossil records previously reported in[1]. The resulting phylogenetic tree was visualised in FigTree v1.4.4 (https://github.com/rambaut/figtree).

Synteny

Synteny was identified using the Chromosomal Orthologous Link analysis approach (https://github.com/chulbioinfo/chrorthlink). We used the set of complete single copy mammalian genes identified with compleasm v0.2 that were used in the phylogenetic tree analysis to assess the conservation of large-scale gene linkage and gene order compared to that of seven other members of the Bovidae family (American bison, wild yak, domestic yak, taurine cattle, indicine cattle, gaur, and water buffalo). Synteny plots were generated in R v4.2.2 using the genoPlotR library[66].

Annotation of the wisent assemblies

We used liftoff v1.6.3[67] to map the annotation (in GFF) of taurine cattle onto the F1 maternal and paternal haplotypes, the F1 primary assembly, and the existing draft assembly. Liftoff was run using "-copies" to look for extra gene copies in the target genome and "-sc 0.95" to specify a minimum sequence identity in exons/CDS of 95% to consider a gene a copy.

Structural variants

We constructed per-chromosome pangenomes with minigraph v0.20[68] using "-cxggs -j 0.2" from the five *Bos* and two *Bison* species from the synteny analysis, using the *Bos taurus taurus* reference sequence as backbone and adding assemblies in order of their mash v2.3 divergence[69]. Graph paths (P-lines) were reconstructed using minigraph call, allowing vg v1.55.0 deconstruct[70] to call structural variants (SV) for each assembly using taurine cattle as reference. We used BCFtools query v1.19[71] to print genotypes for each SV, which were then plotted with upsetplot v0.9 (https://github.com/jnothman/UpSetPlot). We estimated the SV-tree considering the reciprocal of number of SVs between each pair of assemblies, followed by applying an UPGMA clustering with SciPy v1.12[72]. The functional impact of SVs was predicted with the Variant Effect Predictor (VEP) tool[73].

Alignment of short-read DNA samples

We supplemented our short-read dataset with previously generated short-read sequencing data of two male wild wisents (BBO_3569 and BBO_3574) (BioProject: PRJNA312492)[9]. We also included short-read sequencing data of four American bisons (BioProject: PRJNA343262), and four cattle from European taurine breeds (BioProject: PRJNA176557)[74]. For more information about these samples, refer to **Supplementary Table 1**. The quality of the raw sequencing data was assessed using the FastQC v0.11.9 software (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were aligned to our wisent assembly using the MEM algorithm of the Burrows-Wheeler Alignment (BWA) software v0.7.17-r1188[75] with "-T 20" to output only alignments with mapping quality >20. Duplicate reads were marked with samblaster v0.1.24[76], and SAMtools v1.19.2 was used to convert the SAM file into a binary BAM format. Sambamba v0.8.1[77] was used for coordinate-sorting, and bamtools v2.5.1[78] and Qualimap v2.3[79] were used to assess the quality of the alignments. Reads were also aligned to the ARS-UCD1.2 assembly.

Variant calling, postfiltering, and statistics

Variants were called using Freebayes v0.9.21[80] specifying a minimum base quality of 20, a minimum alternate fraction of 0.20, a minimum alternate count of 2, a haplotype length of 0, and a ploidy level of 2. We used a custom python script by setting to missing individual variants whose depth was <1/3 or >2.5 the average genome coverage, as estimated by Qualimap. BCFtools v1.19 was used to further discard SNPs closer than 5 bp to insertions/deletions (InDels), InDels closer than 5 bp to other InDels, variants with a PHRED-quality score <30, and variants with an allele count <2. Finally, BCFtools stats was used to obtain statistics on the final set of called variants. Only autosomal bi-allelic SNPs (InDels excluded) were used in the downstream analyses.

Genome-wide heterozygosity

Heterozygosity was calculated in 1 Mb sliding windows as the number of heterozygous bi-allelic SNPs divided by the total number of bases that had >1/3 and <2.5-times the average genome coverage[81,82]. Heterozygosity was corrected for the number of sites that were excluded because of coverage. Windows with less than 60% of bases within a normal coverage range were excluded.

Detecting runs of homozygosity

We identified runs of homozygosity (ROH) using the approach presented in[81], which uses a corrected measure of heterozygosity estimated in 10 Kb windows[82]. The heterozygosity threshold within a candidate ROH was relaxed to allow peaks of heterozygosity if their inclusion did not inflate the heterozygosity within the final ROH, which had to be below 0.25 the average heterozygosity. This minimized the impact of local assembly or alignment errors.

Realized genomic inbreeding

The realized genomic inbreeding coefficient (F_{ROH}) was estimated from the sum of autosomal ROH longer than 100 Kb divided by the genome length of the first 29 autosomes in the wisent genome (L = 2,682,350,267 bp).

Coverage analysis near the *THRSP* deletion

We assessed coverage near the *THRSP* deletion in the six short read-sequenced wisent samples and 719 publicly available short read samples of wisent, bison, taurine, cattle, indicine cattle, and water buffalo aligned to ARS-UCD1.2 with samtools depth with flags "-aa -r 29:17990000-18000000". Accession numbers of the DNA sequencing data are provided in **Supplementary Table 7**. Coverage was normalized based on the mean sequencing depth across this interval, excluding the deletion region (29:17993500-17996000).

Transcriptome analyses

Publicly available RNA sequencing data from bison (BioProject: PRJNA257088[36]) and water buffalo (BioProject: PRJNA951806[39]) were aligned to the ARS-UCD1.2 assembly and Refseq version 106 annotation with STAR v2.7.9a[83]. Integrative Genomics Viewer v2.14.0[84] was used to visualize the alignments. Accession numbers of the RNA sequencing data are listed in **Supplementary Table 7**. Transcript abundance was quantified using the kallisto v0.46.1 software[85]. Gene expression from *Bos taurus taurus* transcriptomes was obtained from a publicly available TPM matrix (https://zenodo.org/records/7560235) built by the cattle Genotype-Tissue Expression (GTEx) project[38]. Data from the cattle GTEx were filtered by tissue and only samples obtained from females older than 8 months were retained. Gene expression was averaged over these samples and compared to gene expression in the bison sample using Spearman correlation coefficients.

Data and resource availability

The primary assembly of the wisent is publicly available in the European Nucleotide Archive (ENA) under accession GCA_963879515.1 (https://www.ebi.ac.uk/ena/browser/view/GCA_963879515.1). The annotation of the primary assembly is currently underway at Ensembl. HiFi reads of the F1 are available in the ENA at the study accession PRJEB71066 under sample accession SAMEA114863253. Illumina paired-end reads of six captive wisents are available in the ENA at the study accession PRJEB71066 under sample accessions SAMEA115388352, SAMEA115388353, SAMEA115388354 (F1), SAMEA115388355 (dam), SAMEA115388356 (sire), SAMEA115388357. Codes used in this study are available in GitHub: https://github.com/cbortoluzzi/WisentGenomeAssembly.

Acknowledgments

We thank Stefan Hoby from the Tierpark Bern and Martin Kilchenmann from the Tierpark Langenberg for providing DNA and tissue samples used in this study. We are thankful for the technical support provided by Dr. Anna Bratus-Neuenschwander from the ETH Zurich technology platform Functional Genomics Center Zurich (https://fgcz.ch) for sequencing and DNA fragment analysis.

Author contributions

CB characterized genomic diversity, identified ROH, analyzed repeats and synteny, built phylogenetic trees, and drafted the paper; XMM prepared DNA for sequencing; SN and FJ sampled tissue; HP conceived the study, examined SV diversity, investigated putative trait-associated SVs, performed transcriptome analyses, and drafted the paper; ASL assembled genomes, investigated the impact of ROH on the assembly process, built and decomposed pangenomes, contributed to the analysis of putative trait-associated SVs, and drafted the paper. All authors read and approved the final manuscript.

Competing Interests

The authors declare that they do not have any conflict of interest to disclose.

Funder Information

This study was supported by the Swiss National Science Foundation (SNSF, grant ID 204654).

References

1. Chen, L., Qiu, Q., Jiang, Y., Wang, K., *et al.* Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science* **364**, eaav6202 (2019).

2. Bont, R. de. Extinct in the Wild: Finding a Place for the European Bison, 1919–1952. in *Spatializing the History of Ecology* (Routledge, 2017).

3. Tokarska, M., Pertoldi, C., Kowalczyk, R. & Perzanowski, K. Genetic status of the European bison Bison bonasus after extinction in the wild and subsequent recovery. *Mammal Review* **41**, 151–162 (2011).

4. Plumb, G. A range-wide conservation action plan for the European bison. *Oryx* **56**, 171–171 (2022).

5. Olech, W. & Perzanowski, K. European Bison Species Strategic Review – perspectives and challenges. *European Bison Conservation Newsletter* **14**, 5–10 (2022).

6. Soubrier, J., Gower, G., Chen, K., Richards, S. M., Llamas, B., Mitchell, K. J., Ho, S. Y. W., Kosintsev, P., Lee, M. S. Y., Baryshnikov, G., Bollongino, R., Bover, P., Burger, J., Chivall, D., Crégut-Bonnoure, E., Decker, J. E., Doronichev, V. B., Douka, K., Fordham, D. A., Fontana, F., Fritz, C., Glimmerveen, J., Golovanova, L. V., Groves, C., Guerreschi, A., Haak, W., Higham, T., Hofman-Kamińska, E., Immel, A., Julien, M.-A., Krause, J., Krotova, O., Langbein, F., Larson, G., Rohrlach, A., Scheu, A., Schnabel, R. D., Taylor, J. F., Tokarska, M., Tosello, G., van der Plicht, J., van Loenen, A., Vigne, J.-D., Wooley, O., Orlando, L., Kowalczyk, R., Shapiro, B. & Cooper, A. Early cave art and ancient DNA record the origin of European bison. *Nat Commun* **7**, 13158 (2016).

7. Tokarska, M., Marshall, T., Kowalczyk, R., Wójcik, J. M., Pertoldi, C., Kristensen, T. N., Loeschcke, V., Gregersen, V. R. & Bendixen, C. Effectiveness of microsatellite and SNP markers for parentage and identity analysis in species with low genetic diversity: the case of European bison. *Heredity (Edinb)* **103**, 326–332 (2009).

8. Druet, T., Oleński, K., Flori, L., Bertrand, A. R., Olech, W., Tokarska, M., Kaminski, S. & Gautier, M. Genomic Footprints of Recovery in the European Bison. *Journal of Heredity* **111**, 194–203 (2020).

9. Gautier, M., Moazami-Goudarzi, K., Levéziel, H., Parinello, H., Grohs, C., Rialle, S., Kowalczyk, R. & Flori, L. Deciphering the Wisent Demographic and Adaptive Histories from Individual Whole-Genome Sequences. *Molecular Biology and Evolution* **33**, 2801–2814 (2016).

10. Wojciechowska, M., Puchała, K., Nowak-Życzyńska, Z., Perlińska-Teresiak, M., Kloch, M., Drobik-Czwarno, W. & Olech, W. From Wisent to the Lab and Back Again—A Complex SNP Set for Population Management as an Effective Tool in European Bison Conservation. *Diversity* **15**, 116 (2023).

11. Chen, N.-C., Solomon, B., Mun, T., Iyer, S. & Langmead, B. Reference flow: reducing reference bias using multiple population genomes. *Genome Biology* **22**, 8 (2021).

12. Günther, T. & Nettelblad, C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLOS Genetics* **15**, e1008302 (2019).

13. Wang, K., Wang, L., Lenstra, J. A., Jian, J., Yang, Y., Hu, Q., Lai, D., Qiu, Q., Ma, T., Du, Z., Abbott, R. & Liu, J. The genome sequence of the wisent (Bison bonasus). *Gigascience* **6**, 1–5 (2017).

14. Thomma, B. P. H. J., Seidl, M. F., Shi-Kunne, X., Cook, D. E., Bolton, M. D., van Kan, J. A. L. & Faino, L. Mind the gap; seven reasons to close fragmented genome assemblies. *Fungal Genet Biol* **90**, 24–30 (2016).

15. Smith, T. P. L., Bickhart, D. M., Boichard, D., Chamberlain, A. J., Djikeng, A., Jiang, Y., Low, W. Y., Pausch, H., Demyda-Peyrás, S., Prendergast, J., Schnabel, R. D., Rosen, B. D., & Bovine Pangenome Consortium. The Bovine Pangenome Consortium: democratizing production and accessibility of genome assemblies for global cattle breeds and other bovine species. *Genome Biology* **24**, 139 (2023).

16. Jarvis, E. D., Formenti, G., Rhie, A., Guarracino, A., *et al.* Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519–531 (2022).

17. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–175 (2021).

18. Koren, S., Rhie, A., Walenz, B. P., Dilthey, A. T., Bickhart, D. M., Kingan, S. B., Hiendleder, S., Williams, J. L., Smith, T. P. L. & Phillippy, A. M. *De novo* assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology* **36**, 1174–1182 (2018).

19. Leonard, A. S., Crysnanto, D., Fang, Z.-H., Heaton, M. P., Vander Ley, B. L., Herrera, C., Bollwein, H., Bickhart, D. M., Kuhn, K. L., Smith, T. P. L., Rosen, B. D. & Pausch, H. Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. *Nat Commun* **13**, 3012 (2022).

20. Rice, E. S., Koren, S., Rhie, A., Heaton, M. P., Kalbfleisch, T. S., Hardy, T., Hackett, P. H., Bickhart, D. M., Rosen, B. D., Ley, B. V., Maurer, N. W., Green, R. E., Phillippy, A. M., Petersen, J. L. & Smith, T. P. L. Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. *GigaScience* **9**, giaa029 (2020).

21. Stroupe, S., Martone, C., McCann, B., Juras, R., Kjöllerström, H. J., Raudsepp, T., Beard, D., Davis, B. W. & Derr, J. N. Chromosome-level reference genome for North American bison (Bison bison) and variant database aids in identifying albino mutation. *G3 Genes|Genomes|Genetics* **13**, jkad156 (2023).

22. Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., Floto, R. A., Frost, S. D. W., Corander, J., Bentley, S. D. & Parkhill, J. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology* **21**, 180 (2020).

23. Leonard, A. S., Mapel, X. M. & Pausch, H. Pangenome genotyped structural variation improves molecular phenotype mapping in cattle. *Genome Research.* gr.278267.123 (2024) doi:10.1101/gr.278267.123.

24. Milia, S., Leonard, A. S., Mapel, X. M., Ulloa, S. M. B., Drögemüller, C. & Pausch, H. Taurine pangenome uncovers a segmental duplication upstream of KIT associated with depigmentation in white-headed cattle. 2024.02.02.578587 Preprint at https://doi.org/10.1101/2024.02.02.578587 (2024).

25. Stroupe, S., Forgacs, D., Harris, A., Derr, J. N. & Davis, B. W. Genomic evaluation of hybridization in historic and modern North American Bison (Bison bison). *Sci Rep* **12**, 6397 (2022).

26. Bosse, M. & van Loon, S. Challenges in quantifying genome erosion for conservation. *Front Genet* **13**, 960958 (2022).

27. Chu, C., Borges-Monroy, R., Viswanadham, V. V., Lee, S., Li, H., Lee, E. A. & Park, P. J. Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat Commun* **12**, 3836 (2021).

28. Adelson, D. L., Raison, J. M. & Edgar, R. C. Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *Proc Natl Acad Sci U S A* **106**, 12855–12860 (2009).

29. Wang, K., Lenstra, J. A., Liu, L., Hu, Q., Ma, T., Qiu, Q. & Liu, J. Incomplete lineage sorting rather than hybridization explains the inconsistent phylogeny of the wisent. *Commun Biol* **1**, 1–9 (2018).

30. Wu, D.-D., Ding, X.-D., Wang, S., Wójcik, J. M., Zhang, Y., Tokarska, M., Li, Y., Wang, M.-S., Faruque, O., Nielsen, R., Zhang, Q. & Zhang, Y.-P. Pervasive introgression facilitated domestication and adaptation in the Bos species complex. *Nat Ecol Evol* **2**, 1139–1145 (2018).

31. Mastromonaco, G. F., Coppola, G., Crawshaw, G., DiBerardino, D. & King, W. A. Identification of the homologue of the bovine Rob(1;29) in a captive gaur (Bos gaurus). *Chromosome Res* **12**, 725–731 (2004).

32. Crysnanto, D., Leonard, A. S., Fang, Z.-H. & Pausch, H. Novel functional sequences uncovered through a bovine multiassembly graph. *PNAS* **118**, (2021).

33. Leonard, A. S., Crysnanto, D., Mapel, X. M., Bhati, M. & Pausch, H. Graph construction method impacts variation representation and analyses in a bovine super-pangenome. *Genome Biol* **24**, 124 (2023).

34. Wang, X., Carre, W., Zhou, H., Lamont, S. J. & Cogburn, L. A. Duplicated Spot 14 genes in the chicken: characterization and identification of polymorphisms associated with abdominal fat traits. *Gene* **332**, 79–88 (2004).

35. Oppenheimer, J., Rosen, B. D., Heaton, M. P., Vander Ley, B. L., Shafer, W. R., Schuetze, F. T., Stroud, B., Kuehn, L. A., McClure, J. C., Barfield, J. P., Blackburn, H. D., Kalbfleisch, T. S., Bickhart, D. M., Davenport, K. M., Kuhn, K. L., Green, R. E., Shapiro, B. & Smith, T. P. L. A Reference Genome Assembly of American Bison, Bison bison bison. *Journal of Heredity* **112**, 174–183 (2021).

36. Dobson, L. K., Zimin, A., Bayles, D., Fritz-Waters, E., Alt, D., Olsen, S., Blanchong, J., Reecy, J., Smith, T. P. L. & Derr, J. N. De novo assembly and annotation of the North American bison (Bison bison) reference genome and subsequent variant identification. *Anim Genet* **52**, 263–274 (2021).

37. Kuemmerle, N. B. & Kinlaw, W. B. THRSP (thyroid hormone responsive). *Atlas Genet Cytogenet Oncol Haematol* **15**, 480–482 (2011).

38. Liu, S., Gao, Y., Canela-Xandri, O., Wang, S., Yu, Y., Cai, W., Li, B., Xiang, R., Chamberlain, A. J., Pairo-Castineira, E., D'Mellow, K., Rawlik, K., Xia, C., Yao, Y., Navarro, P., Rocha, D., Li, X., Yan, Z., Li, C., Rosen, B. D., Van Tassell, C. P., Vanraden, P. M., Zhang, S., Ma, L., Cole, J. B., Liu, G. E., Tenesa, A. & Fang, L. A multi-tissue atlas of regulatory variants in cattle. *Nat Genet* **54**, 1438–1447 (2022).

39. Si, J., Dai, D., Li, K., Fang, L. & Zhang, Y. A Multi-Tissue Gene Expression Atlas of Water Buffalo (Bubalus bubalis) Reveals Transcriptome Conservation between Buffalo and Cattle. *Genes* **14**, 890 (2023).

40. Colbert, C. L., Kim, C.-W., Moon, Y.-A., Henry, L., Palnitkar, M., McKean, W. B., Fitzgerald, K., Deisenhofer, J., Horton, J. D. & Kwon, H. J. Crystal structure of Spot 14, a modulator of fatty acid synthesis. *Proc Natl Acad Sci U S A* **107**, 18820–18825 (2010).

41. LaFave, L. T., Augustin, L. B. & Mariash, C. N. S14: insights from knockout mice. *Endocrinology* **147**, 4044–4047 (2006).

42. Cordain, L., Watkins, B. A., Florant, G. L., Kelher, M., Rogers, L. & Li, Y. Fatty acid analysis of wild ruminant tissues: evolutionary implications for reducing diet-related chronic disease. *Eur J Clin Nutr* **56**, 181–191 (2002).

43. Rule, D. C., Broughton, K. S., Shellito, S. M. & Maiorano, G. Comparison of muscle fatty acid profiles and cholesterol concentrations of bison, beef cattle, elk, and chicken. *J Anim Sci* **80**, 1202–1211 (2002).

44. Anderson, G. W., Zhu, Q., Metkowski, J., Stack, M. J., Gopinath, S. & Mariash, C. N. The Thrsp Null Mouse (Thrsptm1cnm) and Diet Induced Obesity. *Mol Cell Endocrinol* **302**, 99–107 (2009).

45. Wang, Y. H., Bower, N. I., Reverter, A., Tan, S. H., De Jager, N., Wang, R., McWilliam, S. M., Cafe, L. M., Greenwood, P. L. & Lehnert, S. A. Gene expression patterns during intramuscular fat development in cattle. *J Anim Sci* **87**, 119–130 (2009).

46. Schering, L., Albrecht, E., Komolka, K., Kühn, C. & Maak, S. Increased expression of thyroid hormone responsive protein (THRSP) is the result but not the cause of higher intramuscular fat content in cattle. *Int J Biol Sci* **13**, 532–544 (2017).

47. Koch, R. M., Jung, H. G., Crouse, J. D., Varel, V. H. & Cundiff, L. V. Growth, digestive capability, carcass, and meat characteristics of Bison bison, Bos taurus, and Bos x Bison. *J Anim Sci* **73**, 1271–1281 (1995).

48. Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z. B., Schatz, M. C. & Soyk, S. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology* **23**, 258 (2022).

49. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 245 (2020).

50. Huang, N. & Li, H. compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics* **39**, btad595 (2023).

51. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).

52. Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H. & Phillippy, A. M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).

53. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

54. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

55. Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C. & Smit, A. F. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457 (2020).

56. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* **12**, 1269–1276 (2002).

57. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**, i351-358 (2005).

58. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–580 (1999).

59. Wheeler, T. J., Clements, J., Eddy, S. R., Hubley, R., Jones, T. A., Jurka, J., Smit, A. F. A. & Finn, R. D. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res* **41**, D70–D82 (2013).

60. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11 (2015).

61. Bortoluzzi, C., Wright, C. J., Lee, S., Cousins, T., Genez, T. A. L., Thybert, D., Martin, F. J., Haggerty, L., Consortium, T. D. T. of L. P., Blaxter, M. & Durbin, R. Lepidoptera genomics based on 88 chromosomal reference sequences informs population genetic parameters for conservation. 2023.04.14.536868 Preprint at https://doi.org/10.1101/2023.04.14.536868 (2023).

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

63. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

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

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

66. Guy, L., Kultima, J. R. & Andersson, S. G. E. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334–2335 (2010).

67. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).

68. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biology* **21**, 265 (2020).

69. Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S. & Phillippy, A. M. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology* **17**, 132 (2016).

70. Liao, W.-W., Asri, M., Ebler, J., Doerr, D., *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).

71. Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M. & Li, H. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

72. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F. & van Mulbregt, P. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261–272 (2020).

73. McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P. & Cunningham, F. The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122 (2016).

74. Stothard, P., Liao, X., Arantes, A. S., De Pauw, M., Coros, C., Plastow, G. S., Sargolzaei, M., Crowley, J. J., Basarab, J. A., Schenkel, F., Moore, S. & Miller, S. P. A large and diverse collection of bovine genome sequences from the Canadian Cattle Genome Project. *GigaScience* **4**, 49 (2015).

75. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]* (2013).

76. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).

77. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).

78. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P. & Marth, G. T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691–1692 (2011).

79. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).

80. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at https://doi.org/10.48550/arXiv.1207.3907 (2012).

81. Bortoluzzi, C., Bosse, M., Derks, M. F. L., Crooijmans, R. P. M. A., Groenen, M. A. M. & Megens, H.-J. The type of bottleneck matters: Insights into the deleterious variation landscape of small managed populations. *Evolutionary Applications* **13**, 330–341 (2020).

82. Bosse, M., Megens, H.-J., Madsen, O., Paudel, Y., Frantz, L. A. F., Schook, L. B., Crooijmans, R. P. M. A. & Groenen, M. A. M. Regions of Homozygosity in the Porcine Genome: Consequence of Demography and the Recombination Landscape. *PLOS Genetics* **8**, e1003100 (2012).

83. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

84. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. Integrative Genomics Viewer. *Nat Biotechnol* **29**, 24–26 (2011).

85. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotech* **34**, 525–527 (2016).