Taurine pangenome uncovers a segmental duplication upstream of *KIT* associated with depigmentation in white-headed cattle

Sotiria Milia[1,*], Alexander S. Leonard[1,*], Xena Marie Mapel[1], Sandra Milena Bernal Ulloa[2], Cord Drögemüller[3], Hubert Pausch[1,#]

[1]Animal Genomics, ETH Zurich, Zurich, Switzerland
[2]Animal Physiology, ETH Zurich, Zurich, Switzerland
[3]Institute of Genetics, Vetsuisse Faculty, University of Bern, Bern, Switzerland

* joint first authorship
# corresponding author: hubert.pausch@usys.ethz.ch

Abstract

Cattle have been selectively bred for coat color, spotting, and depigmentation patterns. The assumed autosomal dominant inherited genetic variants underlying the characteristic white head of Fleckvieh, Simmental, and Hereford cattle have not been identified yet, although the contribution of structural variation upstream the *KIT* gene has been proposed. Here, we construct a graph pangenome from 24 haplotype assemblies representing seven taurine cattle breeds to identify and characterize the white head-associated locus for the first time based on long-read sequencing data. Through examining assembly path similarities within the graph, we reveal an association between two most likely serial alleles of a complex structural variant 66 kb upstream *KIT* and facial depigmentation. The complex structural variant contains a variable number of tandemly duplicated 14.3 kb repeats, consisting of LTRs, LINEs, and other repetitive elements, leading to misleading alignments of short and long reads when using a linear reference. We align 250 short-read sequencing samples spanning 15 cattle breeds to the pangenome graph, further validating that the alleles of the structural variant segregate with head depigmentation. We estimate an increased count of repeats in Hereford relative to Simmental and other white-headed cattle breeds from the graph alignment coverage, suggesting a large under-assembly in the current Hereford-based cattle reference genome which had fewer copies. We show that exploiting assembly path similarities within graph pangenomes can reveal trait-associated complex structural variants.

Introduction

Natural selection, domestication, and selective breeding have shaped the genetic and phenotypic variation of modern cattle (*Bos taurus*). Hundreds of cattle breeds are recognized worldwide belonging to the indicine (*Bos taurus indicus*) or taurine (*Bos taurus taurus*) subspecies (Loftus et al., 1994). Coat color, spotting, and piebald patterns differentiate the breeds as these traits have frequently been used for breed formation (Cieslak et al., 2011). Depigmentation phenotypes can vary from tiny white spots, white head and leg markings, large irregular white patches, symmetrical patterns such as a white band around the

midsection or white stripes along the dorsal and ventral midlines, to almost complete depigmentation in white born animals (Olson, 1981). The genetic basis of coat color variation has been studied extensively in many domesticated species including cattle, and several large-effect loci, sometimes associated with pleiotropic effects, have been identified (Bannasch et al., 2021; Durkin et al., 2012; Haase et al., 2009; Henkel et al., 2019; Joerg et al., 1996; Rubin et al., 2012; Trigo et al., 2021). Variation in coat colour is mostly the result of spontaneous mutations, and the derived trait-associated alleles are often propagated through selective breeding because the mutant animals' striking appearance and special value to their owners result in iconic, breed-defining phenotypes. In particular, allelic variation nearby *KIT* encoding KIT proto-oncogene, receptor tyrosine kinase is responsible for several breed-defining coat color patterns (Durkin et al., 2012; Grosz and MacNeil, 1999; Hayes et al., 2010; Küttel et al., 2019; Liu et al., 2009; Qanbari et al., 2014).

Structural variants (SVs) affecting the expression of *KIT* contribute to depigmentation phenotypes in cattle and other species (Artesi et al., 2020; Dürig et al., 2017; Durkin et al., 2012; Giuffra et al., 2002; Küttel et al., 2019; Nagle et al., 1995; Venhoranta et al., 2013). A series of three *KIT*-related alleles, all complex structural rearrangements expected to reflect gain of function variants resulting in dysregulated *KIT* expression, cause the color-sided pattern in cattle and yak (Artesi et al., 2020; Durkin et al., 2012; Küttel et al., 2019). A sequence-based genome-wide association study mapped the white-headed phenotype in Fleckvieh cattle also to a region encompassing *KIT* (Qanbari et al., 2014), but the causal variant and functional mechanism remain unclear. Increased sequencing coverage upstream of *KIT* has been detected in short-read sequenced Hereford and Simmental cattle, both of which have a characteristic white head (Whitacre, 2014), suggesting that a copy number variant might contribute to this enigmatic phenotype. Due to the difficulty of resolving SVs using short reads and alignment-based methods (Bickhart and Liu, 2014), further characterization and validation of this region has not been attempted to date.

Long-read sequencing and genome assembly methods enable detecting and resolving SVs that were previously difficult to identify using short-read sequencing data. Pangenomes, which incorporate a series of assemblies rather than a single reference genome, can fairly represent all types of variation present in a population. There are several approaches to constructing pangenomes, using reference-backed (Li et al., 2020), phylogeny-guided (Armstrong et al., 2020), or all-to-all (Garrison et al., 2023) alignments. These graph pangenomes reduce reference bias because the read aligners can “see” all alleles simultaneously, improving variant calling accuracy and enabling SV genotyping (Liao et al., 2023). Some graph aligners, like vg giraffe (Sirén et al., 2021) and experimental updates to GraphAligner (Rautiainen and Marschall, 2020), can exploit haplotype information stored in the graph (which input assemblies took which variant paths), further improving genotyping speed and accuracy.

Here, we combine 14 newly assembled HiFi-based haplotypes with ten previously published assemblies to create a 24-assembly taurine pangenome graph. We identify candidate regions for the white-headed phenotype directly from the graph structure, and then align 250 short-read samples from 15 breeds to confirm the association between serial alleles of a complex structural variant upstream of *KIT* and the phenotype.

Results

Head pigmentation segregates among 24 taurine assemblies

We analyzed 24 long read sequencing-based assemblies representing seven taurine cattle breeds to investigate the genetic underpinnings of head pigmentation. These assemblies included 14 newly generated HiFi-based haplotypes from Simmental, Evolèner, Rätisches Grauvieh, Brown Swiss, Original Braunvieh, and Braunvieh cattle that were assembled through trio-binning (N=6) or a dual-assembly approach (N=8) when parental data were not available (Table 1). The newly generated assemblies were highly contiguous with a mean contig N50 of 78.2 Mb, as well as highly accurate with a mean quality value (QV) of 48, or approximately 1 error every 62 kb. We also validated the gene completeness with compleasm, identifying on average 98% of expected conserved genes. A Simmental haplotype (SIM_2) produced through trio-binning required some manual curation to correct an under assembled region of interest introduced later (see Methods). We further collected nine publicly available long read-based haplotype assemblies including a Simmental (Heaton et al., 2021) and a Highland (Rice et al., 2020) haplotype assembled with ONT reads, as well as seven HiFi-assembled Brown Swiss and Original Braunvieh haplotypes (Leonard et al., 2023a, 2023b, 2022) (Supplementary Table 1). We also included the cattle reference genome (ARS-UCD1.2) which was assembled from PacBio Continuous Long Read data collected from a Hereford cow (Rosen et al., 2020).

Table 1. 14 newly HiFi-assembled taurine haplotypes from Evolèner (EVO), Simmental (SIM), Rätisches Grauvieh (RGV), Original Braunvieh (OBV), or Brown Swiss (BSW) ancestry that were created with trio-binning (Trio) or a dual-assembly (Dual) approach. Braunvieh (BV) indicates samples with both Brown Swiss and Original Braunvieh ancestry. HiFi coverage is given in gigabases per animal, and so is shared when both haplotypes are given. N50 is calculated with respect to the ARS-UCD1.2 size. The quality value (QV) score is estimated with merqury from short reads. Gene completeness is calculated by compleasm on the cetartiodactyl ODB10 gene set.

Sample / Acronym	HiFi Coverage (Gb)	Haplotype	Size	contigs	N50	QV	Gene completeness
GIRxSIM / SIM_2	91.4	Trio	3.16	886	52.5	53.1	99.6
RGVxSIM / SIM_3	140.1	Trio	3.21	682	89.9	58.8	99.5
RGVxSIM / RGV			3.09	816	91.9	58.3	97.2
DWZxEVO / EVO	139.1	Trio	3.19	527	92.1	57.5	99.7
Brown Swiss / BSW_3	111.8	Trio	3.11	1031	80.6	46.7	98.4
Brown Swiss / BSW_4			3.11	857	71.2	47.1	98.3
Original Braunvieh / OBV_3	135.8	Dual	2.95	2250	72.6	44.5	97.0
Original Braunvieh / OBV_4			2.99	1550	71.5	44.9	97.9
Braunvieh / BV_1	122.9	Dual	2.95	2336	74.9	47.4	98.2
Braunvieh / BV_2			2.99	1755	72.4	47.6	96.6
Braunvieh / BV_3	145.0	Dual	3.02	1701	83.2	43.1	95.1
Braunvieh / BV_4			2.85	2099	82.6	44.2	96.6
Braunvieh / BV_5	131.0	Dual	3.05	1373	85.0	38.2	98.6
Braunvieh / BV_6			2.91	1681	74.6	39.1	96.5

A white head (Figure 1A) is characteristic for Simmental and Hereford cattle while cattle from the other five breeds included in the pangenome have almost fully pigmented heads. The head pigmentation of all individuals (or their parents for the trio-binned assemblies) from

which the haplotypes were assembled matched these expectations. An approximate tree constructed with mash from the assemblies (Figure 1B) reveals that Simmental and Hereford belong to different clades despite sharing the white-headed phenotype, while the overall tree is compatible with previously reported breed phylogenies (Decker et al., 2014).

Pangenome graph construction and association testing identifies a trait-associated complex structural variant

The 24 haplotype assemblies were incorporated into a graph pangenome to make complex SVs amenable to association testing. We constructed graph pangenomes for each autosome separately for the 24 taurine assemblies using reference-free base-level alignment with PGGB (Garrison et al., 2023). Graph construction required approximately 484 CPU hours in total across the autosomes with a peak memory usage of 24.5 Gb. Combined, the graphs contained 3,701,244,547 nucleotide bases across 57,964,174 nodes connected by 79,388,152 edges. The reference genome autosomes only cover 2,489,368,269 bases, leading to approximately 1,212 Mb of non-reference sequence. Much of this non-reference sequence originates from centromeres assembled on the HiFi assemblies. The graphs contained 1,145 nodes larger than 100 kb, totaling 816 Mb of sequence, of which 695 nodes contained 588 Mb of centromeric satellites. There were also several large, non-repetitive nodes that were erroneously not merged into syntenic regions in the graphs, inflating the amount of non-reference sequence. The largest instance was a single 2.6 Mb node on BTA10:22720789-25310574, where the actual coordinates with poor alignments are between approximately 23-24 Mb (Supplementary Figure 1). However, this specific region in ARS-UCD1.2 has previously been reported for poor alignment and imputation quality (Pausch et al., 2017), and so this issue is not unique to graph building. These artefactually large nodes are overall rare, and do not substantially impact graph-wide analyses.

We searched for regions in the graph pangenome where the haplotype paths were classified into distinct groups (white-headed Simmental and Hereford vs. all others with colored heads) to identify variants associated with a white head. To do so, we assessed the pairwise Jaccard similarity over the paths taken by each assembly through 1 kb windows and took the ratio of average similarities between similar phenotypes and opposite phenotypes. A large ratio indicates regions of the graph where e.g., a Simmental and Hereford assembly are similar, a Brown Swiss and Highland are similar, but the Simmental and Brown Swiss are distinct. This “phenotype signature”-approach identified a single prominent peak at BTA6:70.09-70.12 Mb, with a second, less strongly associated cluster very close by at BTA6:70.05 Mb (Figure 1C). Several smaller, single window peaks (e.g., BTA6:54.7 Mb and BTA21:47.3 Mb) were primarily driven by most likely breed-specific SVs of Simmental cattle (Supplementary Figure 2), and so were not considered for subsequent analysis.

The most prominent association signal was from a large, complex bubble spanning BTA6:70099582-70123136. All 20 assemblies from color-headed animals effectively contained a 20.6 kb deletion with respect to the reference genome, while the three white-headed Simmental assemblies contained more sequence than the Hereford-based ARS-UCD1.2 assembly. A less prominently associated second cluster was a 6.8 kb insertion present approximately 47 kb further upstream in color-headed animals of both breeds (SIM and HER), in addition to these assemblies traversing the ~1 kb sequence contained in

BTA6:70052702-70053557 twice. The white-headed animals only traversed those reference coordinates once. These two bubbles were respectively 66 and 113 kb upstream the translation start site of *KIT* (NM_001166484.1) at 70,166,794 bp (Figure 1D), a previously reported positional and functional candidate gene for head depigmentation in cattle (Fontanesi et al., 2010; Qanbari et al., 2014; Whitacre, 2014).

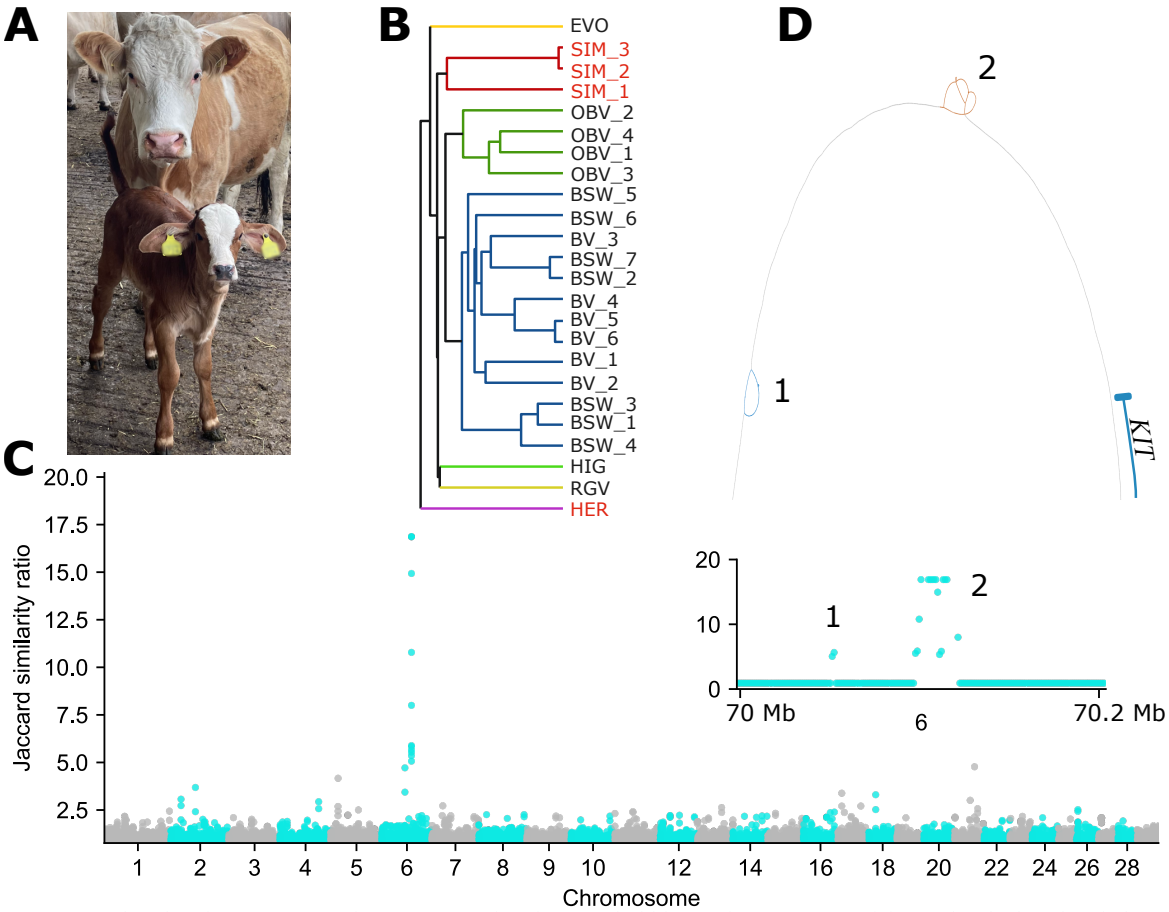

Figure 1. White head in taurine cattle is associated with an SV bubble on chromosome 6. (A) Purebred white-headed Simmental cow with its daughter showing a large white blaze. The daughter originates from a cross with a Gir sire. (B) Mash-based phylogenetic tree of the 24 input taurine assemblies across chromosome 1. Sample names from breeds with a white/color-headed phenotype are red/black respectively, while each breed cluster on the tree has its own color. (C) Jaccard similarity ratio within 1 kb bins (with respect to the ARS-UCD1.2 reference genome) of the pangenome graph between white-headed and color-headed groups. There were two separate regions of interest, both on chromosome 6 upstream of KIT *(indicated with 1 and 2 in the inset). (D) An approximately 150 kb region containing the two regions of interest and the first coding exon of the* KIT *gene.*

A graph bubble containing repetitive sequence is associated with coat color of the bovine head

We characterized the sequence content within and adjacent to the most strongly trait-associated graph bubble to reveal a possible functional explanation for the depigmented heads of Simmental and Hereford cattle. The bubble contains a total number of 276 nodes which correspond to a combined length of ~43 kb for each of the three Simmental haplotypes and ~26 kb for the Hereford-based ARS-UCD1.2 assembly. The three Simmental assemblies contain three copies of a 14.3 kb sequence, which form the bubble. The ARS-UCD1.2

reference genome contains two copies of this segment but misses 1.8 kb and 800 bp of sequence from each copy respectively relative to the full 14.3 kb sequence observed in the Simmental assemblies. The assemblies of the color-headed cattle appear to have only part (~5.5 kb) of one copy of the 14.3 kb sequence, corresponding to a deletion of 20.6 kb relative to the ARS-UCD1.2 reference genome. The summed length of the nodes within the graph bubble is approximately 29 kb, as some nodes are traversed multiple times across the tandemly duplicated sequences, demonstrating there is sufficient sequence similarity between and across haplotypes to share and reuse nodes (Figure 2A).

The full sequence of the segmental duplication identified in the Simmental assemblies contains transposable elements consisting mostly of long terminal repeats (LTRs). The structure of each copy of the segmental duplication starts with a short interspersed nuclear element (SINE) and ends with a long interspersed nuclear element (LINE) (Figure 2B). This repetitive nature of the duplication contributes to the complexity and variability of the region, thus leading to misleading alignments of both short and long reads when using a linear reference (Supplementary Figure 3). The color-headed animals share the sequence immediately preceding and following the segmental duplication but have a truncated copy of the tandemly duplicated sequence that misses most of the LTR elements.

A detailed functional investigation and identification of putative *cis*-regulatory sequences within or nearby the segmental duplication is challenging due to the lack of a comprehensive functional annotation and three-dimensional structure of the bovine genome. Therefore, we characterized the sequence content of the trait-associated region based on human regulatory elements for which we lifted the coordinates onto the ARS-UCD1.2 cattle reference genome. We mapped human promoters, enhancers, and transcription factor binding sites to the orthologous region of bovine ARS-UCD12 identifying various regulatory elements nearby the segmental duplication including distant enhancer like sequences (Supplementary Figure 3).

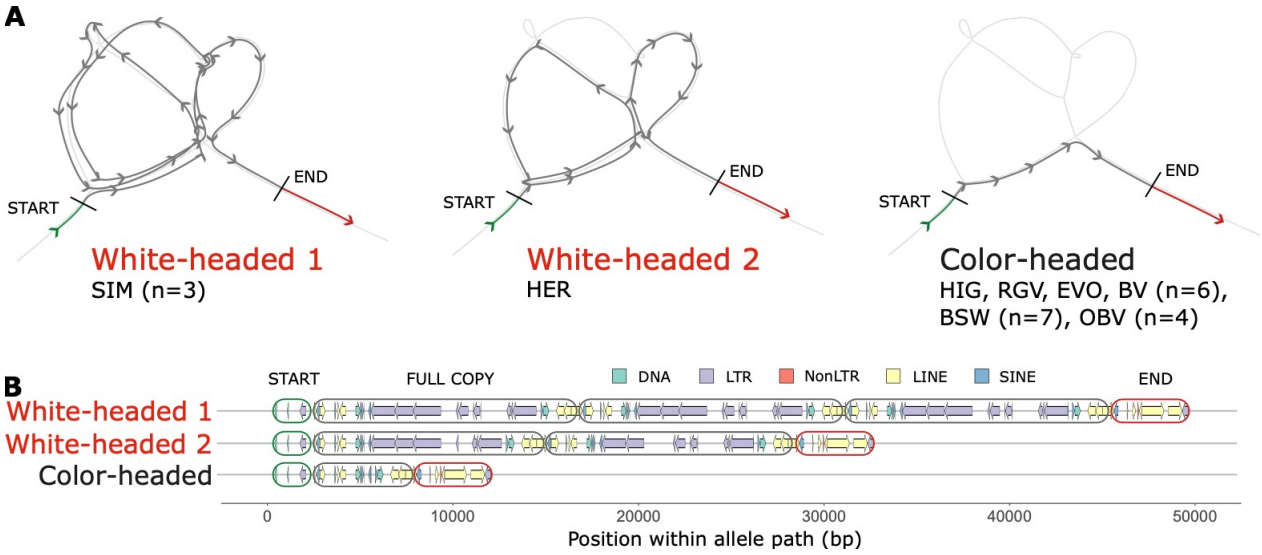

Figure 2. Topology of a trait-associated bubble upstream the KIT *gene. (A) Bandage plots showing the paths traversed for the three observed alleles (two for white-headed breeds and one for color-headed) through the pangenome. Unless stated otherwise, n=1 for the breed. The grey arrow paths indicate the segmental duplication regions, while the green and red arrow paths respectively indicate the sequence preceding and following the segmental duplication. (B) Repeat structure within the candidate SV region, for 5 classes of genomic elements. The start and end sequences are taken from (A), while the*

full copy refers to the 14.3 kb segmental duplication copy found in all three Simmental assemblies. The arrow direction denotes the strand of the repeat element (+/forward and -/reverse).

Pangenome-based genotyping in a diversity panel of short-read sequenced cattle supports that a complex SVs underlies depigmentation in white-headed cattle

To further validate an association between the segmental duplication and the white-headed phenotype, we collected 250 publicly available whole genome short read sequencing samples spanning 15 taurine breeds (Supplementary Table 2). The average read depth of these samples estimated from alignments against ARS-UCD1.2 was 12.3 (± 4.4)-fold. Cattle of the Simmental, Hereford, Groningen White Headed, Kazakh White Headed, Yaroslavl, Montbéliarde, and Normande breeds show different types of the white-head phenotype, while the other breeds do not (although some are completely white colored or have minor white head markings). We validated the breed-identification associated with each sample through a principal component analysis of sequence variant genotypes called from reference alignments (Supplementary Figure 4), with breeds clustering largely as expected. Cattle from white-headed and color-headed breeds were sometimes more closely clustered than two cattle from white-headed breeds (as observed earlier in Figure 1B), indicating that the allele is likely not private to a particular ancestral taurine lineage but is segregating in more distantly related breeds, possibly indicating admixture. Otherwise, this could also indicate that the causal mutation event occurred before the formation of modern cattle breeds, or allelic heterogeneity.

We then aligned all short-read sequencing samples with vg giraffe to a pangenome subgraph spanning BTA6:69099582-71123136 (1 Mb up- and downstream of the most significantly associated bubble), keeping the alignments in the graph node coordinate system, and assessed normalized coverage over this region. We used short-read samples of six Brown Swiss cattle from which twelve haplotypes were assembled to estimate the accuracy of the pangenome alignments and coverage counting. All twelve haplotype-resolved assemblies had the deletion allele, retaining approximately 5.5 kb of the segmental duplication sequence. The mean short read coverage across the SV region predicts a sequence length closer to 8000 ± 620 bp, suggesting the coverage estimates are inflated, likely due to the abundant repetitive elements or bias from extracting reads based on alignments to the linear reference. The alignment mapping qualities were also reduced in this region, with many reads with a mapping quality of 0. Filtering out these reads removed almost all reads across this region, and so we retained these reads but ensured a read could only map once.

Normalized coverage ± 1 Mb from the candidate SVs (excluding the SVs themselves) was uniform across all breeds regardless of head coloration (Figure 3A), corroborating our graph alignment and normalization approach works as intended. In contrast, the normalized coverage within the first SV region was consistently missing in white-headed breeds but inconsistently present in color-headed breeds (Figure 3B, Supplementary Figure 5), indicating poor association with the phenotype when considering additional breeds not represented in the graph. The coverage in the second, more strongly associated SV region was consistently normal in color-headed breeds and consistently elevated in white-headed

breeds (Figure 3C), apart from Montbéliarde, and Normande, confirming the duplication is widely present in and likely unique to white-headed breeds.

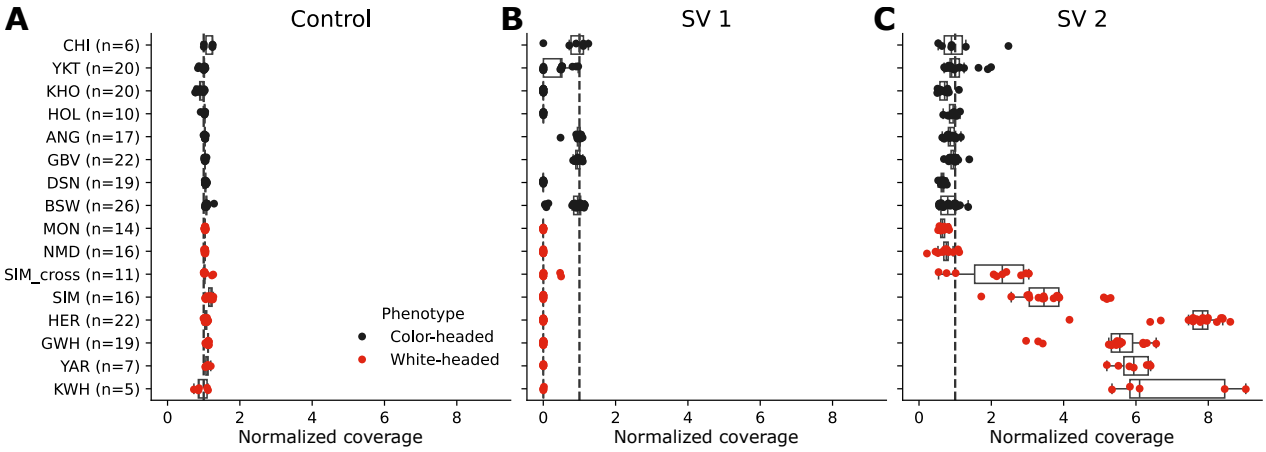

Figure 3. Sequencing depth in short read samples from 15 taurine breeds. (A) Normalized coverage (normalized over both sequencing depth and length of the region) per breed 1 Mb up- and downstream of the SV region. The dashed line indicates the expected normalized coverage of 1. (B) Similar to (A), but for the smaller SV 1 region. (C) Similar to (A), but for the larger SV 2 region.

Within the primary SV candidate region, Hereford animals had the highest average short-read coverage, corresponding to approximately 7-9 copies of the segmental duplication unit. Two additional Hereford samples in previously reported optical maps (Talenti et al., 2022) show an 88.5 kb insertion, corresponding to only about 6 additional copies of the 14.3 kb segmental duplication unit compared to ARS-UCD1.2, again suggesting that coverage estimates from short-read alignments may be slightly inflated. We also corroborated from older PacBio Continuous Long Read alignments that Hereford genomes most likely contain more copies of the segmental duplication than Simmental genomes (Supplementary Figure 6). The 16 Simmental samples had roughly between 3-5 copies, in agreement with the 3 copies derived from the haplotype-resolved assemblies. Samples with lower coverage could possibly indicate heterozygous carriers, as would be expected given that a small number of color-headed animals which are expected to carry two copies of a haplotype without duplication exist in the Simmental breed (Qanbari et al., 2014). Two F1s (RGVxSIM and GIRxSIM) had an intermediate increase in coverage, consistent with one haplotype (Simmental) contributing multiple copies of the repeat unit and the other haplotype (color headed) contributing only part of a copy. This was also broadly true for 7 Holstein x Simmental crosses, although some appeared to inherit the color-headed allele from Simmental parents with fewer copies.

Given the variability within the group of white-headed cattle breeds studied, including the decreased coverage in Montbéliarde and Normande, we examined the specific coverage patterns in node space (not reference coordinate space) per breed (Supplementary Figure 7). As expected, we find color-headed breeds have roughly uniform normalized coverage of 1 across the expected allele path, with little coverage outside (Figure 4A). However, Montbéliarde and Normande show elevated coverage in three parts of the white-headed allele path (Figure 4B), suggesting they may contain multiple partial copies of the segmental duplication, or a different allele structure not captured by the assemblies contained in the

pangenome graph. Again as expected, the white-headed breeds with overall elevated coverage had fairly uniform coverage across different nodes (Figure 4C), primarily differing in the suspected number of copies rather than suggesting different allele structures. The Hereford samples also contained even coverage at the 1.8 kb and 800 bp sequences not observed in the ARS-UCD1.2 reference genome, suggesting that white-headed Hereford cattle contain copies of the segmental duplication matching those observed in the Simmental assemblies, but the reference genome is misassembled at this locus in terms of both copy content and number. The nodes with elevated coverage in both Montbéliarde and Normande and the other white-headed breeds correspond to regions rich with L1MAB (non-LTR retrotransposon) and MER45B (DNA transposon) repeats, possibly indicating that these elements are functionally relevant sequence for the phenotype.

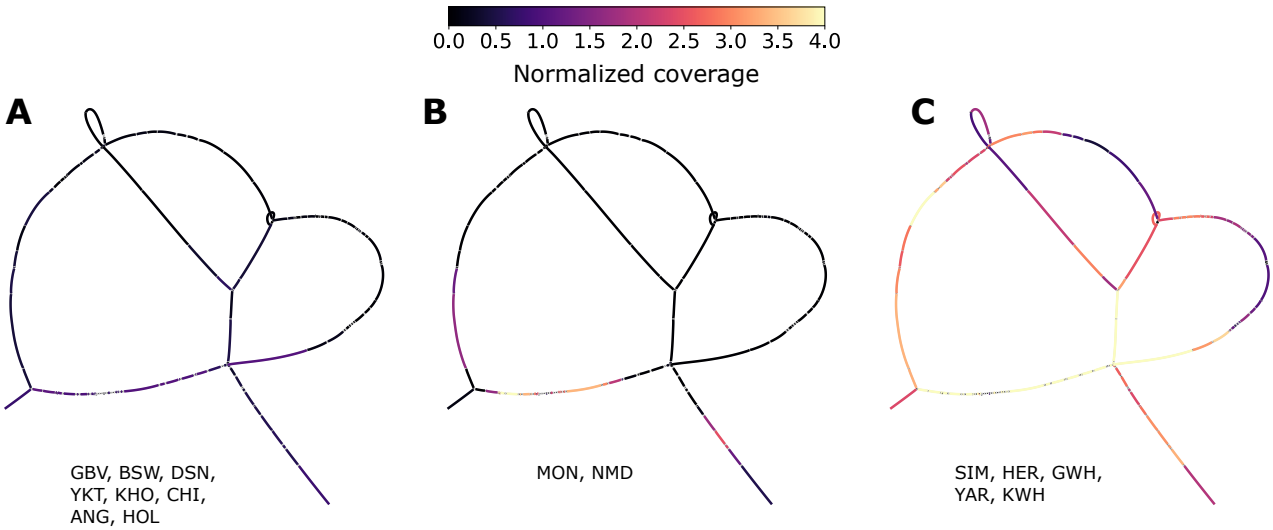

Figure 4. Normalized graph coverage for the three observed coverage patterns. Coverage is averaged across all samples for each node for (A) GBV, BSW, DSN, YKT, KHO, CHI, ANG, and HOL, (B) MON and NMD, and (C) SIM, HER, GWH, YAR, and KWH. Lighter node colors indicate more aligned coverage to that node.

Discussion

Alleles of a segmental duplication upstream of *KIT* are associated with the characteristic white heads of Simmental and Hereford cattle, two globally occurring breeds of cattle known for dominant inherited white spotting patterns. Two F1s (RGVxSIM, GIRxSIM) originating from the crossing between white-headed Simmental cows and color-headed sires from the Gir and Rätisches Grauvieh breeds both have white heads (Figure 1A, Supplementary Figure 8) and carry one copy of the haplotype containing the segmental duplication (SIM_2, SIM_3) which confirms the dominant inheritance of the white-headed phenotype. The white head is a breed-defining pattern in Simmental cattle where most of or all the head is depigmented against a pigmented body with white legs and underside, in addition to varying amounts and extent of white body spots as seen in other breeds such as Holstein and Montbéliarde. A typical Hereford animal is characterized by a white head, a white stripe running from the back of the head to the withers, and white legs and underparts (Supplementary Figure 9). Evidence for a series of multiple alleles at the so-called spotting (S) locus was reported already more than 90 years ago, explaining the different degree of white pattern in Simmental and Hereford cattle (Ibsen, 1933). A potential negative side-effect of white-headed cattle,

similar to gonadal hypoplasia in *KIT*-related color-sided cattle (Venhoranta et al., 2013), is an increased susceptibility to infectious bovine keratoconjunctivitis, eye cancer, or squamous cell carcinoma under UV light exposure (Anderson, 1991). However, a peculiar pigmentation around the eyes, which is a moderately to highly heritable trait associated with reduced susceptibility to eye disease, occurs in some white-headed cattle breeds (e.g., Fleckvieh, Simmental, Hereford, Groningen White Headed) and *KIT* alleles possibly contribute to this pigmentation pattern (Gonzalez-Prendes et al., 2022; Jara et al., 2022; Pausch et al., 2012).

Our analyses suggest that a segmental duplication upstream the *KIT* gene impacts head pigmentation through a regulatory mechanism as the tandemly duplicated sequences do not overlap coding sequence. The repeat elements identified within the duplication may have regulatory function themselves or the breakpoints of the inserted sequence may overlap regulatory elements, thereby impacting the expression of *KIT* which leads to depigmented heads. The visible white markings are caused by a lack of melanocytes in the hair follicles and skin. We suspect that the dominance of the different white-head associated alleles reflects a gain of function resulting from dysregulated expression of the affected *KIT* gene. We thus provide another example of how iconic coat color phenotypes, which are common in domestic animal breeds, may serve as an excellent system for identifying mutations that affect fundamental processes of melanocyte development, migration, survival, and proliferation. Such mechanisms have frequently been described for *KIT* and *ASIP* mediated depigmentation patterns observed in several species (Artesi et al., 2020; Bannasch et al., 2021; Brooks et al., 2008; David et al., 2014; Henkel et al., 2019). Lifting over human genome annotations to the cattle reference localized several putative enhancer sequences nearby the segmental duplication which may be involved in regulating *KIT* expression. However, this approach neglects bovine-specific regulatory elements and is unable to correctly resolve the coordinates of diverged regulatory sequences, requiring further validation that the predicted enhancer elements exist and are relevant in the bovine genome. It could be speculated that the insertion of a largely repetitive sequence itself has no immediate effect but moves a distant enhancer of *KIT* even further upstream. Such a positional separation between regulatory and coding sequences could potentially affect the spatiotemporal expression of *KIT* during development (Berrozpe et al., 2013, 2006, 1999), thereby causing depigmentation of the head in animals carrying at least one copy of the duplication. Additional copies of the segmental duplication would further increase the distance between the enhancer sequences and *KIT*, potentially further dysregulating its expression. Such a mechanism could explain the larger area of depigmentation in Hereford cattle than in Simmental cattle, and the partial phenotypes observed in Montbéliarde and Normande, which likely contain multiple partial copies of the duplication.

We also observe that the colored-head associated “deletion” allele (with respect to the Hereford-based ARS-UCD1.2 reference) is extremely common and therefore most likely represents the ancestral bovine *KIT* allele. In addition to the 12 color-headed taurine breeds examined here, we identified the same 20.6 kb deletion in optical mapping data from six Sanga (African taurine x indicine) cattle breeds (Talenti et al., 2022), as well as in pangenomes containing three further taurine breeds (Jang et al., 2023) and two indicine breeds and three non-cattle bovids (Leonard et al., 2023a) (Supplementary Table 3). Thus, it is very likely that the color-headed phenotype is the ancestral state, while the white-headed phenotype has evolved by mutation(s). Although there are no assemblies to confirm the

allelic structure, the Russian Yaroslavl breed also appears to have a similar SV to Hereford and Simmental, despite geographically limited opportunities of introgression (Zinovieva et al., 2020). However, there is limited metadata and no pedigree available to confirm the ancestry of the public Yaroslavl samples. The smaller 6-7 kb SV is also widespread but again is less consistent with the color-headed phenotype, suggesting that it does not play a role in head pigmentation or may affect a different phenotype.

The number of copies of the segmental duplication also varies within breeds. We observed considerable variation in Simmental and Fleckvieh cattle, as expected given the sporadic occurrence of color-headed individuals in these breeds (Qanbari et al., 2014). One can speculate that the degree of depigmentation is dosage-dependent or that the segmental duplication contributes to depigmentation beyond the head. However, such investigations require extensive phenotype observations and precise copy number resolution, neither of which are available for the public short-read samples considered in our study. The presence of other *KIT* alleles associated with head and coat color phenotypes (Hayes et al., 2010; Pausch et al., 2012) may further complicate such an analysis.

The use of pangenomes for association testing has largely relied on presence/absence variation between groups (Brynildsrud et al., 2016; Leonard et al., 2022), decomposed graph genome variation (Chin et al., 2023), or genotypes obtained by mapping of short reads to pangenomes (Cochetel et al., 2023; Sirén et al., 2021). We have developed and applied an approach that compares path similarities and differences between assemblies of samples with divergent phenotypes. While this pangenome “phenotype signature” analysis successfully identified and resolved a promising SV associated with a white-headed phenotype, it has limited sensitivity to small variation or highly multiallelic variation, as both weaken the Jaccard similarity across the windows tested. Furthermore, non-reference bias is only mitigated for the breeds represented in our pangenome; any additional, unrepresented allelic structure (e.g., possibly in Montbéliarde and Normande) is still lost in downstream alignment and analysis. Potential misassemblies, like that likely found in the reference genome, further complicate association testing on graphs. As bovine assemblies become increasingly available (Smith et al., 2023), the approaches outlined here will become increasingly powerful for resolving complex variation associated with binary phenotypes. This will be particularly important in cases such as those presented here where even long read alignments to a linear reference fail to consistently resolve the variation.

Methods

Ethics statement

The sampling of blood was approved by the veterinary office of the Canton of Zurich (animal experimentation permit ZH 200/19).

Animals

Two purebred Simmental (*Bos taurus taurus*) cows were inseminated with semen from Gir (*Bos taurus indicus*) and Rätisches Grauvieh (*Bos taurus taurus*) sires. A purebred Evolèner

cow was inseminated with semen from a dwarf zebu (DWZ). Two female calves (GIR x SIM and DWZ x EVO) and a male calf (RGV x SIM) were delivered at term. DNA was prepared from blood samples of the F1s and dams, and from cryopreserved semen samples of the sires. DNA of a purebred Brown Swiss cow was prepared from a blood sample provided by Braunvieh Schweiz. DNA of three Braunvieh (cross between Brown Swiss and Original Braunvieh) and one Original Braunvieh bulls was prepared from testis tissue sampled at a commercial slaughterhouse from a previous project (Mapel et al., 2024).

Sequencing

PacBio HiFi reads were collected from four F1s used for trio-binning with three SMRT Cell 8M sequenced for each sample on a Sequel IIe, and from another four animals used for dual-assembly with a total of four SMRTs Cell 25M sequenced on a Revio and five SMRT Cells 8M sequenced on a Sequel IIe. Illumina paired-end (2x150 bp) reads were collected from the four F1s and their parents.

Genome assemblies

The Hereford (HER), Simmental (SIM_1), Highland (HIG), Original Braunvieh (OBV_1 and OBV_2), and Brown Swiss (BSW_1, BSW_2, BSW_5, BSW_6, and BSW_7) were downloaded from public sources (Supplementary Table 1). We assembled the new genomes with hifiasm (v0.19.5) (Cheng et al., 2021) with default parameters. Where available, we used parental short reads to create haplotype-resolved assemblies with yak (v0.1). Contigs were then scaffolded to the ARS-UCD1.2+Y reference with RagTag (v2.1.0) (Alonge et al., 2022). We assessed assembly quality with merqury (6b5405) (Rhie et al., 2020) and completeness with compleasm (v0.2.2) (Huang and Li, 2023) using the Cetartiodactyla lineage.

Identification of a missing copy of the segmental duplication in the SIM_2 assembly

The SIM_2 haplotype was initially assembled with only two copies of the segmental duplication. We used the parental short reads to triobin the HiFi reads using meryl (v1.3) and Canu (v2.2) (Koren et al., 2018) before aligning to the reference genome with minimap2 (v2.24) (Li, 2018). We then extracted all HiFi reads corresponding to the Simmental haplotype from the region 6:69099582-71123136 with SAMtools fastq (v1.16.1) (Danecek et al., 2021). We reassembled these reads with hifiasm with the additional parameters -D 10 -N 500 to improve repeat sensitivity. We then manually split the original Simmental assembly between the two repeat copies and used RagTag patch to “gap fill” with the sequence of the newly assembled third copy.

Pangenome construction

The pangenome graph was built with pggb (736c50d) (Garrison et al., 2023), using wfmash (67ab187) (Marco-Sola et al., 2021), seqwish (f44b402) (Garrison and Guarracino, 2023), smoothxg (31d99f2) (https://github.com/pangenome/smoothxg), gfaffix (0.1.5) (https://github.com/marschall-lab/GFAffix), and odgi (de70fcda) (Guarracino et al., 2022). We estimated the pangenome percent-identity with mash (v2.3) (Ondov et al., 2019, 2016) triangle as 97-99% across different autosomes, as well as a segment-length of 75k and a min-match-length of 31. We used the hierarchy linkage clustering with the UPGMA algorithm from scipy (v1.11.4) (Virtanen et al., 2020) to approximate the phylogenetic tree across

chromosome 1 after removing centromeres, which are unevenly present on assemblies. We then extracted the relevant subgraph using odgi extract for the region 6:69099582-71123136 followed by cycle-breaking with odgi sort -p "wc". The subgraph region was visualized with BandageNG (v2022.09) (https://github.com/asl/BandageNG) (Wick et al., 2015).

Calculation of the Jaccard similarity ratio

We used odgi extract to bin every 1 kb of reference sequence into subgraphs, and then used odgi similarity to calculate the Jaccard similarity of all paths across these subgraphs. We calculated the Jaccard ratio as the ratio of intra-group similarities (white-white or colored-colored) to inter-group similarities (white-colored), such that a high value demonstrates both groups are internally consistent but distinct across groups.

Identification of repetitive elements in the graph bubble

We used RepeatMasker (https://www.repeatmasker.org/, v4.1.5) with the repbase repeat library to identify repetitive elements in the SIM_3 assembly haplotype. We used odgi procbed to adjust the repeat coordinates into the subregion spanned by the graph bubble, before using odgi inject to annotate the graph with repeat paths. Finally, we used odgi untangle with "-g" to produce a sequence arrow map, which was then visualized in R with the packages gggenes (v.0.5.1) and ggplot2 (v3.4.4).

Short read data collection and mapping to the graph

We downloaded publicly available short read sequencing data from study accessions PRJNA814817, PRJEB9343, PRJNA176557, PRJEB45822, PRJEB56301, PRJNA494431, PRJNA762180, PRJNA642008 and PRJEB18113 (Supplementary Table 1) using fastq-dl (v2.0.4) with "--group-by-sample". All fastq sequences were trimmed with fastp (v0.23.4) (Chen et al., 2018) using default parameters. Reads were then aligned to ARS-UCD1.2+Y using strobealign (6bbc5b7) (Sahlin, 2022), followed by read group sorting, duplicate marking, and coordinate sorting with SAMtools. We then extracted reads from the region of interest with SAMtools fastq, and then aligned these reads to the pangenome using vg giraffe (v1.51.0) (Sirén et al., 2021) with "--named-coordinates --max-multimaps 1 -o gaf". We assessed per-node coverage with gafpack (v0.1.0).

PCA breed validation

Variants were called from the aligned BAM files using DeepVariant (v1.6) (Poplin et al., 2018), using the WGS model on chromosomes 1-3. We conducted a PCA using plink2 (v2.00a4LM) (Chang et al., 2015), first thinning by linkage above r>0.8 every 100 kb, excluding variants below minor allele frequency of 10%, and only retaining biallelic SNPs. Samples were then colored by assigned breed, confirming the accuracy of the metadata associated with each SRA sample.

Validation of the SV in other breeds and species

We used the merged VCF file of structural variants called from optical mapping data from (Talenti et al., 2022) to investigate the presence of the candidate SV in taurine and indicine cattle. We also examined the 14-breed pangenome from (Jang et al., 2023), first realigning the constituent assemblies to the pangenome using minigraph --call (v0.20) (Li et al., 2020),

followed by examining the paths taken through bubble variation as described in (Leonard et al., 2023a). We also aligned PacBio Continuous Long Read data from the Simmental, Holstein, Hereford, Original Braunvieh and Evolèner breeds available at the study accession PRJEB72196 to the reference genome using minimap2 with the "map-pb" preset. We then extracted all aligned reads over the region BTA6:69099582-71123136 using SAMtools fastq, and then used minimap2 to align the 14.3 kb segmental duplication to these reads, counting the number of hits > 10 kb to estimate the copy number per sample.

Prediction of regulatory elements overlapping the segmental duplication

We downloaded coordinates from publicly available human regulatory elements curated by the FANTOM5 (Andersson et al., 2014) and ENCODE (Moore et al., 2020) consortia from https://fantom.gsc.riken.jp/5/datafiles/latest/extra/Enhancers/human_permissive_enhancers_phase_1_and_2.bed.gz and https://downloads.wenglab.org/V3/GRCh38-cCREs.bed, respectively. Genome coordinates of human regulatory elements were converted from the GRCh37/hg19 and GRCh38/hg38 assemblies to the ARS-UCD1.2 assembly using the LiftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver) from the UCSC Genome Browser.

Data Access

HiFi reads used for genome assembly are available in the European Nucleotide Archive (ENA) (https://www.ebi.ac.uk/ena/browser/home) at the study accession PRJEB42335 under sample accessions SAMEA115129361, SAMEA113612082, SAMEA113612088, SAMEA113612091, SAMEA115129362, SAMEA115129363, SAMEA115129364, and SAMEA115129365. Parental and F1 short reads are available in the ENA at the study accession PRJEB28191 under sample accessions SAMEA115121766 and SAMEA115121767 for the RGVxSIM (SAMEA115129365), SAMEA115121768 and SAMEA115121769 for the DWZxEVO (SAMEA115129364), SAMEA115121770 and SAMEA115121771 for the GIRxSIM (SAMEA115129363), and SAMEA115121772 and SAMEA115121773 for the BSWxBSW (SAMEA115129361). PacBio CLR data are available at the study accession PRJEB72196 under sample accessions SAMEA115160498, SAMEA115160497, SAMEA115160496, SAMEA115160495, SAMEA115160494, and SAMEA115160493. All scripts and workflows are available online (https://github.com/AnimalGenomicsETH/pangenome_KIT) as well as in the Supplementary Code.

Author contributions

SM analyzed pangenome variation, interpreted results, and wrote the manuscript; ASL assembled and aligned long reads, developed the phenotype signature approach, analyzed pangenome variation, analyzed short-read data, interpreted results, and wrote the manuscript; XMM extracted high molecular weight DNA and revised the manuscript; SMBU was responsible for animal experimentation and the sampling of blood; CD contributed data, interpreted results and contributed to the writing of the manuscript; HP conceived the study,

interpreted results, and wrote the manuscript; all authors approved the final version of the manuscript.

Acknowledgements

We are thankful for the technical support provided by Dr. Anna Bratus-Neuenschwander from the ETH Zürich technology platform Functional Genomics Center Zurich (https://fgcz.ch) and the Next Generation Sequencing Platform at the University of Bern for sequencing and DNA fragment analysis. We also thank Eirini Lampraki from Pacific Biosciences for DNA sequencing on a Revio system. We thank Flavio Ferrari (AgroVet-Strickhof) for animal handling.

Funding

This study was supported by grants from the Swiss National Science Foundation (SNSF, grant ID 204654), the Arbeitsgemeinschaft Schweizerischer Rinderzüchter (ASR), Zollikofen, Switzerland and the Federal Office for Agriculture (FOAG), Bern, Switzerland. The funding bodies were neither involved in the design of the study and collection, analysis, and interpretation of data nor in writing the manuscript.

References

Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z.B., Schatz, M.C., Soyk, S., 2022. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biology 23, 258. https://doi.org/10.1186/s13059-022-02823-7

Anderson, D.E., 1991. Genetic study of eye cancer in cattle. J Hered 82, 21–26. https://doi.org/10.1093/jhered/82.1.21

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F.O., Jørgensen, M., Andersen, P.R., Bertin, N., Rackham, O., Burroughs, A.M., Baillie, J.K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C.J., Meehan, T.F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C.O., Heutink, P., Hume, D.A., Jensen, T.H., Suzuki, H., Hayashizaki, Y., Müller, F., Forrest, A.R.R., Carninci, P., Rehli, M., Sandelin, A., 2014. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461. https://doi.org/10.1038/nature12787

Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I.T., Novak, A.M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., Genereux, D., Johnson, J., Marinescu, V.D., Alföldi, J., Harris, R.S., Lindblad-Toh, K., Haussler, D., Karlsson, E., Jarvis, E.D., Zhang, G., Paten, B., 2020. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature 587, 246–251. https://doi.org/10.1038/s41586-020-2871-y

Artesi, M., Tamma, N., Deckers, M., Karim, L., Coppieters, W., Van den Broeke, A., Georges, M., Charlier, C., Durkin, K., 2020. Colour-sidedness in Gloucester cattle is

associated with a complex structural variant impacting regulatory elements downstream of KIT. Animal Genetics 51, 461–465. https://doi.org/10.1111/age.12932
Bannasch, D.L., Kaelin, C.B., Letko, A., Loechel, R., Hug, P., Jagannathan, V., Henkel, J., Roosje, P., Hytönen, M.K., Lohi, H., Arumilli, M., Minor, K.M., Mickelson, J.R., Drögemüller, C., Barsh, G.S., Leeb, T., 2021. Dog colour patterns explained by modular promoters of ancient canid origin. Nat Ecol Evol 5, 1415–1423. https://doi.org/10.1038/s41559-021-01524-x
Berrozpe, G., Agosti, V., Tucker, C., Blanpain, C., Manova, K., Besmer, P., 2006. A Distant Upstream Locus Control Region Is Critical for Expression of the Kit Receptor Gene in Mast Cells. Molecular and Cellular Biology 26, 5850–5860. https://doi.org/10.1128/MCB.01854-05
Berrozpe, G., Bryant, G.O., Warpinski, K., Ptashne, M., 2013. Regulation of a Mammalian Gene Bearing a CpG Island Promoter and a Distal Enhancer. Cell Reports 4, 445–453. https://doi.org/10.1016/j.celrep.2013.07.001
Berrozpe, G., Timokhina, I., Yukl, S., Tajima, Y., Ono, M., Zelenetz, A.D., Besmer, P., 1999. The W(sh), W(57), and Ph Kit expression mutations define tissue-specific control elements located between -23 and -154 kb upstream of Kit. Blood 94, 2658–2666.
Bickhart, D.M., Liu, G.E., 2014. The challenges and importance of structural variation detection in livestock. Front. Genet. 5. https://doi.org/10.3389/fgene.2014.00037
Brooks, S.A., Lear, T.L., Adelson, D.L., Bailey, E., 2008. A chromosome inversion near the KIT gene and the Tobiano spotting pattern in horses. Cytogenetic and Genome Research 119, 225–230. https://doi.org/10.1159/000112065
Brynildsrud, O., Bohlin, J., Scheffer, L., Eldholm, V., 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biology 17, 238. https://doi.org/10.1186/s13059-016-1108-8
Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7. https://doi.org/10.1186/s13742-015-0047-8
Chen, S., Zhou, Y., Chen, Y., Gu, J., 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890. https://doi.org/10.1093/bioinformatics/bty560
Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li, H., 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods 18, 170–175. https://doi.org/10.1038/s41592-020-01056-5
Chin, C.-S., Behera, S., Khalak, A., Sedlazeck, F.J., Sudmant, P.H., Wagner, J., Zook, J.M., 2023. Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes. Nat Methods 20, 1213–1221. https://doi.org/10.1038/s41592-023-01914-y
Cieslak, M., Reissmann, M., Hofreiter, M., Ludwig, A., 2011. Colours of domestication. Biological Reviews 86, 885–899. https://doi.org/10.1111/j.1469-185X.2011.00177.x
Cochetel, N., Minio, A., Guarracino, A., Garcia, J.F., Figueroa-Balderas, R., Massonnet, M., Kasuga, T., Londo, J.P., Garrison, E., Gaut, B.S., Cantu, D., 2023. A super-pangenome of the North American wild grape species. Genome Biology 24, 290. https://doi.org/10.1186/s13059-023-03133-2
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., 2021. Twelve years of SAMtools and BCFtools. GigaScience 10, giab008. https://doi.org/10.1093/gigascience/giab008
David, V.A., Menotti-Raymond, M., Wallace, A.C., Roelke, M., Kehler, J., Leighty, R., Eizirik, E., Hannah, S.S., Nelson, G., Schäffer, A.A., Connelly, C.J., O'Brien, S.J., Ryugo, D.K., 2014. Endogenous Retrovirus Insertion in the KIT Oncogene

Determines White and White spotting in Domestic Cats. G3 Genes|Genomes|Genetics 4, 1881–1891. https://doi.org/10.1534/g3.114.013425
Decker, J.E., McKay, S.D., Rolf, M.M., Kim, J., Alcalá, A.M., Sonstegard, T.S., Hanotte, O., Götherström, A., Seabury, C.M., Praharani, L., Babar, M.E., Regitano, L.C. de A., Yildiz, M.A., Heaton, M.P., Liu, W.-S., Lei, C.-Z., Reecy, J.M., Saif-Ur-Rehman, M., Schnabel, R.D., Taylor, J.F., 2014. Worldwide Patterns of Ancestry, Divergence, and Admixture in Domesticated Cattle. PLOS Genetics 10, e1004254. https://doi.org/10.1371/journal.pgen.1004254
Dürig, N., Jude, R., Holl, H., Brooks, S.A., Lafayette, C., Jagannathan, V., Leeb, T., 2017. Whole genome sequencing reveals a novel deletion variant in the KIT gene in horses with white spotted coat colour phenotypes. Animal Genetics 48, 483–485. https://doi.org/10.1111/age.12556
Durkin, K., Coppieters, W., Drögemüller, C., Ahariz, N., Cambisano, N., Druet, T., Fasquelle, C., Haile, A., Horin, P., Huang, L., Kamatani, Y., Karim, L., Lathrop, M., Moser, S., Oldenbroek, K., Rieder, S., Sartelet, A., Sölkner, J., Stålhammar, H., Zelenika, D., Zhang, Z., Leeb, T., Georges, M., Charlier, C., 2012. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. Nature 482, 81–84. https://doi.org/10.1038/nature10757
Fontanesi, L., Tazzoli, M., Russo, V., Beever, J., 2010. Genetic heterogeneity at the bovine KIT gene in cattle breeds carrying different putative alleles at the spotting locus. Animal Genetics 41, 295–303. https://doi.org/10.1111/j.1365-2052.2009.02007.x
Garrison, E., Guarracino, A., 2023. Unbiased pangenome graphs. Bioinformatics 39, btac743. https://doi.org/10.1093/bioinformatics/btac743
Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., Hagmann, J., Vorbrugg, S., Marco-Sola, S., Kubica, C., Ashbrook, D.G., Thorell, K., Rusholme-Pilcher, R.L., Liti, G., Rudbeck, E., Nahnsen, S., Yang, Z., Moses, M.N., Nobrega, F.L., Wu, Y., Chen, H., Ligt, J. de, Sudmant, P.H., Soranzo, N., Colonna, V., Williams, R.W., Prins, P., 2023. Building pangenome graphs. https://doi.org/10.1101/2023.04.05.535718
Giuffra, E., Törnsten, A., Marklund, S., Bongcam-Rudloff, E., Chardon, P., Kijas, J.M.H., Anderson, S.I., Archibald, A.L., Andersson, L., 2002. A large duplication associated with dominant white color in pigs originated by homologous recombination between LINE elements flanking KIT. Mammalian Genome 13, 569–577. https://doi.org/10.1007/s00335-002-2184-5
Gonzalez-Prendes, R., Ginja, C., Kantanen, J., Ghanem, N., Kugonza, D.R., Makgahlela, M.L., Groenen, M.A.M., Crooijmans, R.P.M.A., 2022. Integrative QTL mapping and selection signatures in Groningen White Headed cattle inferred from whole-genome sequences. PLOS ONE 17, e0276309. https://doi.org/10.1371/journal.pone.0276309
Grosz, M., MacNeil, M., 1999. Brief communication. The “spotted” locus maps to bovine chromosome 6 in Hereford-cross population. Journal of Heredity 90, 233–236. https://doi.org/10.1093/jhered/90.1.233
Guarracino, A., Heumos, S., Nahnsen, S., Prins, P., Garrison, E., 2022. ODGI: understanding pangenome graphs. Bioinformatics 38, 3319–3326. https://doi.org/10.1093/bioinformatics/btac308
Haase, B., Brooks, S.A., Tozaki, T., Burger, D., Poncet, P.-A., Rieder, S., Hasegawa, T., Penedo, C., Leeb, T., 2009. Seven novel KIT mutations in horses with white coat colour phenotypes. Animal Genetics 40, 623–629. https://doi.org/10.1111/j.1365-2052.2009.01893.x
Hayes, B.J., Pryce, J., Chamberlain, A.J., Bowman, P.J., Goddard, M.E., 2010. Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour,

Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. PLOS Genetics 6, e1001139. https://doi.org/10.1371/journal.pgen.1001139
Heaton, M.P., Smith, T.P.L., Bickhart, D.M., Vander Ley, B.L., Kuehn, L.A., Oppenheimer, J., Shafer, W.R., Schuetze, F.T., Stroud, B., McClure, J.C., Barfield, J.P., Blackburn, H.D., Kalbfleisch, T.S., Davenport, K.M., Kuhn, K.L., Green, R.E., Shapiro, B., Rosen, B.D., 2021. A Reference Genome Assembly of Simmental Cattle, Bos taurus taurus. Journal of Heredity 112, 184–191. https://doi.org/10.1093/jhered/esab002
Henkel, J., Saif, R., Jagannathan, V., Schmocker, C., Zeindler, F., Bangerter, E., Herren, U., Posantzis, D., Bulut, Z., Ammann, P., Drögemüller, C., Flury, C., Leeb, T., 2019. Selection signatures in goats reveal copy number variants underlying breed-defining coat color phenotypes. PLOS Genetics 15, e1008536. https://doi.org/10.1371/journal.pgen.1008536
Huang, N., Li, H., 2023. compleasm: a faster and more accurate reimplementation of BUSCO. Bioinformatics 39, btad595. https://doi.org/10.1093/bioinformatics/btad595
Ibsen, H.L., 1933. Cattle Inheritance. I. Color. Genetics 18, 441–480. https://doi.org/10.1093/genetics/18.5.441
Jang, J., Jung, J., Lee, Y.H., Lee, S., Baik, M., Kim, H., 2023. Chromosome-level genome assembly of Korean native cattle and pangenome graph of 14 Bos taurus assemblies. Sci Data 10, 560. https://doi.org/10.1038/s41597-023-02453-z
Jara, E., Peñagaricano, F., Armstrong, E., Ciappesoni, G., Iriarte, A., Navajas, E.A., 2022. Revealing the genetic basis of eyelid pigmentation in Hereford cattle. J Anim Sci 100, skac110. https://doi.org/10.1093/jas/skac110
Joerg, H., Fries, H.R., Meijerink, E., Stranzinger, G.F., 1996. Red coat color in Holstein cattle is associated with a deletion in the MSHR gene. Mammalian Genome 7, 317–318. https://doi.org/10.1007/s003359900090
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., 2018. *De novo* assembly of haplotype-resolved genomes with trio binning. Nature Biotechnology 36, 1174–1182. https://doi.org/10.1038/nbt.4277
Küttel, L., Letko, A., Häfliger, I.M., Signer-Hasler, H., Joller, S., Hirsbrunner, G., Mészáros, G., Sölkner, J., Flury, C., Leeb, T., Drögemüller, C., 2019. A complex structural variant at the KIT locus in cattle with the Pinzgauer spotting pattern. Animal Genetics 50, 423–429. https://doi.org/10.1111/age.12821
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., 2022. Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. Nat Commun 13, 3012. https://doi.org/10.1038/s41467-022-30680-2
Leonard, A.S., Crysnanto, D., Mapel, X.M., Bhati, M., Pausch, H., 2023a. Graph construction method impacts variation representation and analyses in a bovine super-pangenome. Genome Biol 24, 124. https://doi.org/10.1186/s13059-023-02969-y
Leonard, A.S., Mapel, X.M., Pausch, H., 2023b. Pangenome genotyped structural variation improves molecular phenotype mapping in cattle. https://doi.org/10.1101/2023.06.21.545879
Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191
Li, H., Feng, X., Chu, C., 2020. The design and construction of reference pangenome graphs with minigraph. Genome Biology 21, 265. https://doi.org/10.1186/s13059-020-02168-z

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., Buonaiuto, S., Chang, X.H., Cheng, H., Chu, J., Colonna, V., Eizenga, J.M., Feng, X., Fischer, C., Fulton, R.S., Garg, S., Groza, C., Guarracino, A., Harvey, W.T., Heumos, S., Howe, K., Jain, M., Lu, T.-Y., Markello, C., Martin, F.J., Mitchell, M.W., Munson, K.M., Mwaniki, M.N., Novak, A.M., Olsen, H.E., Pesout, T., Porubsky, D., Prins, P., Sibbesen, J.A., Sirén, J., Tomlinson, C., Villani, F., Vollger, M.R., Antonacci-Fulton, L.L., Baid, G., Baker, C.A., Belyaeva, A., Billis, K., Carroll, A., Chang, P.-C., Cody, S., Cook, D.E., Cook-Deegan, R.M., Cornejo, O.E., Diekhans, M., Ebert, P., Fairley, S., Fedrigo, O., Felsenfeld, A.L., Formenti, G., Frankish, A., Gao, Y., Garrison, N.A., Giron, C.G., Green, R.E., Haggerty, L., Hoekzema, K., Hourlier, T., Ji, H.P., Kenny, E.E., Koenig, B.A., Kolesnikov, A., Korbel, J.O., Kordosky, J., Koren, S., Lee, H., Lewis, A.P., Magalhães, H., Marco-Sola, S., Marijon, P., McCartney, A., McDaniel, J., Mountcastle, J., Nattestad, M., Nurk, S., Olson, N.D., Popejoy, A.B., Puiu, D., Rautiainen, M., Regier, A.A., Rhie, A., Sacco, S., Sanders, A.D., Schneider, V.A., Schultz, B.I., Shafin, K., Smith, M.W., Sofia, H.J., Abou Tayoun, A.N., Thibaud-Nissen, F., Tricomi, F.F., Wagner, J., Walenz, B., Wood, J.M.D., Zimin, A.V., Bourque, G., Chaisson, M.J.P., Flicek, P., Phillippy, A.M., Zook, J.M., Eichler, E.E., Haussler, D., Wang, T., Jarvis, E.D., Miga, K.H., Garrison, E., Marschall, T., Hall, I.M., Li, H., Paten, B., 2023. A draft human pangenome reference. Nature 617, 312–324. https://doi.org/10.1038/s41586-023-05896-x

Liu, L., Harris, B., Keehan, M., Zhang, Y., 2009. Genome scan for the degree of white spotting in dairy cattle. Animal Genetics 40, 975–977. https://doi.org/10.1111/j.1365-2052.2009.01936.x

Loftus, R.T., MacHugh, D.E., Bradley, D.G., Sharp, P.M., Cunningham, P., 1994. Evidence for two independent domestications of cattle. Proc Natl Acad Sci U S A 91, 2757–2761.

Mapel, X.M., Kadri, N.K., Leonard, A.S., He, Q., Lloret-Villas, A., Bhati, M., Hiltpold, M., Pausch, H., 2024. Molecular quantitative trait loci in reproductive tissues impact male fertility in cattle. Nat Commun 15, 674. https://doi.org/10.1038/s41467-024-44935-7

Marco-Sola, S., Moure, J.C., Moreto, M., Espinosa, A., 2021. Fast gap-affine pairwise alignment using the wavefront algorithm. Bioinformatics 37, 456–463. https://doi.org/10.1093/bioinformatics/btaa777

Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., Kaul, R., Halow, J., Van Nostrand, E.L., Freese, P., Gorkin, D.U., Shen, Y., He, Y., Mackiewicz, M., Pauli-Behn, F., Williams, B.A., Mortazavi, A., Keller, C.A., Zhang, X.-O., Elhajjajy, S.I., Huey, J., Dickel, D.E., Snetkova, V., Wei, X., Wang, X., Rivera-Mulia, J.C., Rozowsky, J., Zhang, Jing, Chhetri, S.B., Zhang, Jialing, Victorsen, A., White, K.P., Visel, A., Yeo, G.W., Burge, C.B., Lécuyer, E., Gilbert, D.M., Dekker, J., Rinn, J., Mendenhall, E.M., Ecker, J.R., Kellis, M., Klein, R.J., Noble, W.S., Kundaje, A., Guigó, R., Farnham, P.J., Cherry, J.M., Myers, R.M., Ren, B., Graveley, B.R., Gerstein, M.B., Pennacchio, L.A., Snyder, M.P., Bernstein, B.E., Wold, B., Hardison, R.C., Gingeras, T.R., Stamatoyannopoulos, J.A., Weng, Z., 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 583, 699–710. https://doi.org/10.1038/s41586-020-2493-4

Nagle, D.L., Kozak, C.A., Mano, H., Chapman, V.M., Bućan, M., 1995. Physical mapping of the Tec and Gabrb1 loci reveals that the Wsh mutation on mouse chromosome 5 is associated with an inversion. Human Molecular Genetics 4, 2073–2079. https://doi.org/10.1093/hmg/4.11.2073

Olson, T.A., 1981. The genetic basis for piebald patterns in cattle. J Hered 72, 113–116. https://doi.org/10.1093/oxfordjournals.jhered.a109437

Ondov, B.D., Starrett, G.J., Sappington, A., Kostic, A., Koren, S., Buck, C.B., Phillippy, A.M., 2019. Mash Screen: high-throughput sequence containment estimation for genome discovery. Genome Biology 20, 232. https://doi.org/10.1186/s13059-019-1841-x

Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., Phillippy, A.M., 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology 17, 132. https://doi.org/10.1186/s13059-016-0997-x

Pausch, H., MacLeod, I.M., Fries, R., Emmerling, R., Bowman, P.J., Daetwyler, H.D., Goddard, M.E., 2017. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. Genet. Sel. Evol. 49, 24. https://doi.org/10.1186/s12711-017-0301-x

Pausch, H., Wang, X., Jung, S., Krogmeier, D., Edel, C., Emmerling, R., Götz, K.-U., Fries, R., 2012. Identification of QTL for UV-protective eye area pigmentation in cattle by progeny phenotyping and genome-wide association analysis. PLoS ONE 7, e36346. https://doi.org/10.1371/journal.pone.0036346

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P.T., Gross, S.S., Dorfman, L., McLean, C.Y., DePristo, M.A., 2018. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 36, 983–987. https://doi.org/10.1038/nbt.4235

Qanbari, S., Pausch, H., Jansen, S., Somel, M., Strom, T.M., Fries, R., Nielsen, R., Simianer, H., 2014. Classic Selective Sweeps Revealed by Massive Sequencing in Cattle. PLOS Genetics 10, e1004148. https://doi.org/10.1371/journal.pgen.1004148

Rautiainen, M., Marschall, T., 2020. GraphAligner: rapid and versatile sequence-to-graph alignment. Genome Biology 21, 253. https://doi.org/10.1186/s13059-020-02157-2

Rhie, A., Walenz, B.P., Koren, S., Phillippy, A.M., 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biology 21, 245. https://doi.org/10.1186/s13059-020-02134-9

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., 2020. Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. GigaScience 9, giaa029. https://doi.org/10.1093/gigascience/giaa029

Rosen, B.D., Bickhart, D.M., Schnabel, R.D., Koren, S., Elsik, C.G., Tseng, E., Rowan, T.N., Low, W.Y., Zimin, A., Couldrey, C., Hall, R., Li, W., Rhie, A., Ghurye, J., McKay, S.D., Thibaud-Nissen, F., Hoffman, J., Murdoch, B.M., Snelling, W.M., McDaneld, T.G., Hammond, J.A., Schwartz, J.C., Nandolo, W., Hagen, D.E., Dreischer, C., Schultheiss, S.J., Schroeder, S.G., Phillippy, A.M., Cole, J.B., Van Tassell, C.P., Liu, G., Smith, T.P.L., Medrano, J.F., 2020. De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience 9. https://doi.org/10.1093/gigascience/giaa021

Rubin, C.-J., Megens, H.-J., Barrio, A.M., Maqbool, K., Sayyab, S., Schwochow, D., Wang, C., Carlborg, Ö., Jern, P., Jørgensen, C.B., Archibald, A.L., Fredholm, M., Groenen, M.A.M., Andersson, L., 2012. Strong signatures of selection in the domestic pig genome. PNAS 109, 19529–19536. https://doi.org/10.1073/pnas.1217149109

Sahlin, K., 2022. Strobealign: flexible seed size enables ultra-fast and accurate read alignment. Genome Biology 23, 260. https://doi.org/10.1186/s13059-022-02831-7

Sirén, J., Monlong, J., Chang, X., Novak, A.M., Eizenga, J.M., Markello, C., Sibbesen, J.A., Hickey, G., Chang, P.-C., Carroll, A., Gupta, N., Gabriel, S., Blackwell, T.W., Ratan,

A., Taylor, K.D., Rich, S.S., Rotter, J.I., Haussler, D., Garrison, E., Paten, B., 2021. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science 374, abg8871. https://doi.org/10.1126/science.abg8871

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, 2023. The Bovine Pangenome Consortium: democratizing production and accessibility of genome assemblies for global cattle breeds and other bovine species. Genome Biology 24, 139. https://doi.org/10.1186/s13059-023-02975-0

Talenti, A., Powell, J., Wragg, D., Chepkwony, M., Fisch, A., Ferreira, B.R., Mercadante, M.E.Z., Santos, I.M., Ezeasor, C.K., Obishakin, E.T., Muhanguzi, D., Amanyire, W., Silwamba, I., Muma, J.B., Mainda, G., Kelly, R.F., Toye, P., Connelley, T., Prendergast, J., 2022. Optical mapping compendium of structural variants across global cattle breeds. Sci Data 9, 618. https://doi.org/10.1038/s41597-022-01684-w

Trigo, B.B., Utsunomiya, A.T.H., Fortunato, A.A.A.D., Milanesi, M., Torrecilha, R.B.P., Lamb, H., Nguyen, L., Ross, E.M., Hayes, B., Padula, R.C.M., Sussai, T.S., Zavarez, L.B., Cipriano, R.S., Caminhas, M.M.T., Lopes, F.L., Pelle, C., Leeb, T., Bannasch, D., Bickhart, D., Smith, T.P.L., Sonstegard, T.S., Garcia, J.F., Utsunomiya, Y.T., 2021. Variants at the ASIP locus contribute to coat color darkening in Nellore cattle. Genetics Selection Evolution 53, 40. https://doi.org/10.1186/s12711-021-00633-2

Venhoranta, H., Pausch, H., Wysocki, M., Szczerbal, I., Hänninen, R., Taponen, J., Uimari, P., Flisikowski, K., Lohi, H., Fries, R., Switonski, M., Andersson, M., 2013. Ectopic KIT copy number variation underlies impaired migration of primordial germ cells associated with gonadal hypoplasia in cattle (Bos taurus). PLoS ONE 8, e75659. https://doi.org/10.1371/journal.pone.0075659

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., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Whitacre, L., 2014. Structural variation at the KIT locus is responsible for the piebald phenotype in Hereford and Simmental cattle (Thesis). University of Missouri--Columbia. https://doi.org/10.32469/10355/44434

Wick, R.R., Schultz, M.B., Zobel, J., Holt, K.E., 2015. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics 31, 3350–3352. https://doi.org/10.1093/bioinformatics/btv383

Zinovieva, N.A., Dotsev, A.V., Sermyagin, A.A., Deniskova, T.E., Abdelmanova, A.S., Kharzinova, V.R., Sölkner, J., Reyer, H., Wimmers, K., Brem, G., 2020. Selection signatures in two oldest Russian native cattle breeds revealed using high-density single nucleotide polymorphism analysis. PLOS ONE 15, e0242200. https://doi.org/10.1371/journal.pone.0242200