Animal Genetics

WILEY

RESEARCH ARTICLE OPEN ACCESS

Haplotype-Resolved Genome Assemblies for Norwegian Red Cattle

Thea Johanna Hettasch[1] | Matthew Peter Kent[1] | Arne Bjørke Gjuvsland[2] | Mariann Árnyasi[1] | Dag Inge Våge[1]

[1]Department of Animal and Aquacultural Sciences, Faculty of Biosciences, Norwegian University of Life Sciences (NMBU), Ås, Norway | [2]Geno SA, Hamar, Norway

Correspondence: Dag Inge Våge (daginge.vage@nmbu.no)

Received: 18 December 2025 | **Revised:** 18 December 2025 | **Accepted:** 10 March 2026

Keywords: *Bos taurus* | centromere | long-read sequences | trio-binning

ABSTRACT

Norwegian Red (NR) cattle are the main dairy breed in Norway, bred according to a broad breeding goal including health and fertility since the 1970s. Genomic studies on NR cattle have relied on the public Hereford reference, thus increasing the risk of missing or misrepresenting NR breed-specific variation. Moreover, the Hereford reference is a pseudohaploid assembly, representing homologous chromosomes in a collapsed manner, which results in loss of haplotype-specific alleles and misrepresentation of complex variants. To develop more refined NR-specific resources, we utilised long-read sequencing (PacBio HiFi + ONT) and trio-binning to construct six new haplotype-resolved assemblies representing NR genomes. These six NR2025 assemblies show high completeness (BUSCO: 95.82%–98.11%) and contiguity (N_{50}: 73.8–88.5 Mb) and are accurately phased (hamming error rate: 0.46%–2.52%). Most autosomes have been assembled into the acrocentric centromere, and mapping of bovine satellite sequences reveals distinct organisational patterns of different satellite units across this highly repetitive region. The NR2025 assemblies provide a valuable resource for identification of novel variants and haplotypes in the NR population, which will enable more accurate association studies of genotype-to-phenotype relations and genomic predictions in NR cattle, ultimately enhancing the efficacy of selection for desirable traits.

1 | Introduction

A genome assembly is a digital representation of an organism's genome and is an essential tool for understanding genome structure and identifying genetic variants associated with phenotypes of interest. In domestic cattle, high-quality genome assemblies enable more accurate selection, as they are the basis for genetic variant detection and the construction of SNP arrays used for genomic selection and GWAS. A highly accurate description of the cattle genome improves the identification of genotypes associated with traits such as meat and milk production, sustainability, health, and fertility (Guo et al. 2024; Smith et al. 2023).

Recent advances in long-read sequencing methods, such as the Pacific Biosciences (PacBio) and Oxford Nanopore (ONT) platforms achieving read lengths of several thousand bp with high-accuracy (>99%), enable the assembly of highly continuous genomes (Espinosa et al. 2024; Hu et al. 2021). Together with long-read sequencing techniques, the development of algorithms for genome assembly, phasing, scaffolding, and evaluation has revolutionised the construction of near-complete assemblies for non-model organisms (Li and Durbin 2024; Rhie et al. 2021).

The first reference genome for domestic cattle (*Bos taurus*), assembled from the genome of a Hereford cow, was published in 2009 (Elsik et al. 2009). Since then, updated versions have been released (Rosen et al. 2020), with the latest reference (ARS-UCD2.0, GCA_002263795.4) published in 2023 showing increased completeness and contiguity. The Hereford reference has been widely used for genomic analyses in the NR

population (van Den Berg et al. 2020; Kirsanova et al. 2020; Olsen et al. 2016, 2017), which, despite its utility, raises two concerns. First, there is evidence that the use of a non-breed-specific reference can cause bias towards reference alleles (Chen et al. 2021; Lin et al. 2024; Thorburn et al. 2023), thereby limiting the identification of NR-specific variants. Second, the Hereford reference provides a collapsed representation of the diploid cattle genome, with homologous chromosome copies combined into a single 'mosaic' representation. Haplotype-resolved assemblies provide a more precise description of complex structural variation and rare alleles and enable accurate interpretation of allele-specific expression in downstream analyses (Koren et al. 2018; Zhang et al. 2020; Chaisson et al. 2019). A gapless NR genome assembly was released in 2024 (NRF, GCA_963921495.1) answering the need for a breed-specific reference. Motivated by the additional advantages of having a haplotype-resolved representation of the NR genome, this study aimed to construct assemblies representing genome-wide haplotypes.

Using long-read sequencing data and the *trio-binning* assembly algorithm, which uses parental data to reconstruct both haplotypes of a diploid genome (Koren et al. 2018; Cheng et al. 2021), we have successfully assembled six genome-wide haplotypes for NR cattle. The NR2025 assemblies are highly complete and include an initial representation of the acrocentric cattle centromeres. We believe that the NR2025 collection of assemblies will provide a resource for more accurate and comprehensive characterisation of breed- and haplotype-specific variation in the NR population.

2 | Methods

2.1 | Software Tools

Applied software tools, versions and non-default parameters are provided in Table S1. A more detailed description of commands used to run analyses is provided in Note S1.

2.2 | Trio Selection and Sampling

Three trios (dam, sire, daughter) were selected from the NR population. The selection aimed to minimise genetic relatedness and maximise the genetic diversity within the trios. Blood and semen samples were collected from sires by a veterinarian as part of the routine sampling carried out by the breeding company. Collection of blood samples from dams and daughters was done at the Livestock Production Research Centre (SHF) at NMBU.

2.3 | DNA Sequencing, Read Quality-Control and Filtering

DNA for short-read sequencing was isolated from blood and semen using the Qiagen DNAeasy Kit (Qiagen); quality and quantity were assessed using UV spectroscopy and fluorescence, Nanodrop and Qubit, respectively (ThermoFisher). DNA from the sires (semen) was Illumina sequenced (PE150) by BGI Genomics, while short-read (PE150) sequencing of DNA from dams (blood) was performed by Novogene UK, both using the Illumina NovaSeq 6000 platform. Additional short-read sequencing of DNA (blood) from two sires was performed by Novogene UK to obtain a desired coverage of ~30X.

High molecular weight DNA was isolated from blood collected from daughters using the Nanobind CBB kit from Pacific Bioscience. After assessing integrity with agarose gel electrophoresis (gel image available in Figure S1) and concentration with Nanodrop and Qubit, samples ($n = 3$) were size selected by using the SRE kit from Pacific Bioscience to eliminate the short fragments before library preparation. Size selected DNA samples were sequenced both by a commercial provider (Novogene UK) using the PacBio high-fidelity (HiFi) Revio long-read system and in-house using the Ligation Sequencing Kit V14 (SQK-LSK114) and PromethION flow-cells (R10.4.1) from Oxford Nanopore Technologies (ONT). ONT reads were basecalled routinely by the sequencing facility with *Dorado* (7.3.9) using the High-accuracy (HAC) model. Quality-control (QC) of Illumina reads was performed with *FastQC* (Andrews 2010) and filtering of Illumina and PacBio HiFi reads was done with *fastp* (Chen et al. 2018) and *HiFiAdapterFilt* (Sim et al. 2022), respectively, to remove any contaminants or remnant adapter sequences (Table S2). Any ONT reads below 4000 bp, along with the 10% displaying the lowest average *q*-scores, were removed with *Filtlong* (Wick 2017).

2.4 | De Novo Assembly With Trio-Binning

HiFi reads from offspring genomes were assembled into haplotypes using the trio-binning feature implemented in *Hifiasm* (Cheng et al. 2021). To increase the continuity of the assemblies, ONT reads were provided as input using the *-ul* option. K-mer dictionaries ($k = 21$) were constructed from parental short-reads with *Yak* (Li 2020) and provided as input for haplotype separation. After assembly, the reference-based scaffolding tool *RagTag* (Alonge et al. 2019) was used to combine contigs into chromosome-level scaffolds, using the gapless NR assembly (NRF, GCA_963921495.1) as a reference.

2.5 | Evaluation of Assembly Quality, Contiguity and Completeness

Assembly statistics were calculated with *gfastats* (Formenti et al. 2022). *Compleasm* (Huang and Li 2023) was used to assess the presence of universal single-copy orthologs (BUSCO genes) against the Mammalia_od10 database (9226 genes).

2.6 | K-Mer-Based Assessment of Haplotype Completeness and Accuracy

The base-level quality and completeness, and the accuracy of haplotype separation was assessed with *Merqury* (Rhie et al. 2020). *K*-mer dictionaries ($k = 21$) were generated from parental short-reads, offspring HiFi reads and the NR2025 assemblies, with *Meryl* (Rhie et al. 2020). Based on the overlap between *k*-mers found in the offspring reads and the assemblies,

consensus quality (QV) and *k*-mer completeness was estimated. Haplotype-specific *k*-mers (hapmers) were identified as the subset of offspring k-mers found uniquely in one of the parental read sets. Based on the presence of hapmers across NR2025 assemblies and assuming that only hapmers from one of the parents should be found in each haplotype assembly, the global hamming error rate was calculated.

$$\text{Hamming error rate} = \frac{\text{erroneous hapmers}}{\text{total number of hapmers} - 1} * 100$$

2.7 | Assembly-To-Assembly Alignment

To identify larger rearrangements and regions of synteny between the NR2025 assemblies, pairwise assembly alignments were performed and visualised. Pairwise whole genome alignments were performed with *minimap2* (Li 2018), thereafter, syntenic regions and structural rearrangements were identified with *syri* (Goel et al. 2019). Finally, *plotsr* (Goel and Schneeberger 2022) was used to visualise assembly alignments and structural annotations.

2.8 | Mapping of Bovine Satellite and Telomere Sequences

Bovine satellite sequences were obtained from NCBI (Note S2) To simplify downstream analyses, complete sequence and partial fragments were manually categorised into 9 main satellite types based on the NCBI sequence description (Table S3). All satellite sequences were mapped to the NR2025 assemblies using *blastn* (Altschul et al. 1990). A simple binomial test was used to identify regions (1 Mb windows) significantly enriched in satellite hits (percent identity > 98) compared to the average number of hits across the genome (Note S3). *Tidk* (Brown et al. 2023) was used to count the number of vertebrate telomeric repeats (TTAGGG) across assemblies.

3 | Results

3.1 | Sequencing Yield

DNA from three NR trios (dam, sire, daughter) was sequenced to give short-read (Illumina PE150) data for parents, and long-read (HiFi and ONT) for offspring. 77–117 Gb (28–42×) of Illumina sequences was generated from parental DNA, together with 108–127 Gb (39–45×) of HiFi, and 147–161 Gb (53–58×) of ONT long-read data for offspring (Table S4). The maximum read length achieved with ONT sequencing was 500–600 kb and 2%–3% of the ONT sequences obtained for each offspring individual were contained in reads over 100 kb (Table S4).

3.2 | Assembly Quality, Contiguity and Completeness

Six haplotypes were assembled with trio-binning from offspring long-reads: NR2025_1P, NR2025_1M, NR2025_2P, NR2025_2M, NR2025_3P and NR2025_3M. On average, per haploid genome, the initial contig-level assemblies included 619–1380 contigs, comprising 3.08–3.19 Gb of sequence, with a contig N50 value of 73.8–88.6 Mb (Table 1). After alignment to the NRF reference, between 82 and 267 contigs per genome were scaffolded into chromosome-level assemblies (29 autosomes + X + Mitochondrion), comprising a total of 2.77 to 2.83 Gb of sequence, with 7–12 gapless chromosomes composed of a single contig (Table 1). The paternal BtaX assemblies are notably smaller than the maternal, suggesting misassembly (Table S5). This issue, which is common in haplotype-resolved assemblies (Carey et al. 2022), is likely due to low-coverage of short-reads from the paternal X chromosome due to its hemizygosity. During hapmer identification, coverage thresholds are applied to exclude low-frequency (potentially erroneous) and high-frequency (non-specific) hapmers (Rhie et al. 2021; Koren et al. 2018). Paternal hapmers from the non-pseudoautosomal region of the X chromosome may fall below the low-coverage threshold and be misclassified as sequencing errors, leading to a misassembled X chromosome in the paternal haplotype.

All NR2025 assemblies have a consensus quality (QV) close to 70, representing a 99.99999% base accuracy based on the consensus between k-mers in the offspring reads and the final assemblies (Table 1). The k-mer completeness is 94.3%–96.1% for individual haplotype assemblies and 99.7% for pairs of haplotypes (Table 1). This suggests that nearly all reliable *k*-mers found in the offspring reads are also found in the haplotype assembly pairs, with some being haplotype-specific.

The NR2025 assemblies achieved BUSCO scores between 95.82% and 98.11%, with a higher number of missing BUSCO genes than the collapsed ARS-UCD2.0 and NRF references (Table 1). However, in pairs of NR2025 haplotypes, no increase in missing BUSCO genes is observed (Table S6). Pairwise alignment between haplotypes reveals larger regions which are present in one haplotype but absent in the other and overlap with missing BUSCO gene positions (Figure S2a). This suggests that the increase in missing BUSCO genes is due to some regions being erroneously collapsed into one haplotype during assembly. Mapping of maternal and paternal hapmers shows that collapsed regions cover fewer hapmers than surrounding regions (Figure S2b). Erroneous collapsing in low heterozygosity regions is a known assembly problem caused by a graph sparsification step implemented in Hifiasm, where read-to-read alignment is used to remove *contained reads* to simplify the read graph (Li and Durbin 2024; Jain 2023; Kamath et al. 2024).

3.3 | Accuracy of Haplotype Separation

To assess the accuracy of haplotype separation, hapmers derived from parental short-reads were mapped to the NR2025 assemblies. Generally, we observe that maternal hapmers map to chromosomes in the maternal haplotype assemblies and paternal hapmers map to chromosomes in the paternal haplotype assemblies, which shows successful separation and assembly of all NR2025 haplotypes (Figure 1).

The global Hamming error rate—defined as the proportion of hapmers incorrectly assigned to the opposite haplotype—ranges

TABLE 1 | Assembly statistics for NR2025, NRF and ARS-UCD2.0 assemblies.

Breed		NR		NR		NR		NR	Hereford
Assembly		NR2025_1P Paternal	NR2025_1M Maternal	NR2025_2P Paternal	NR2025_2M Maternal	NR2025_3P Paternal	NR2025_3M Maternal	NRF	ARS-UCD2.0
ENA Accession		GCA_977927265.1	GCA_977927225.1	GCA_977970235.1	GCA_977927245.1	GCA_977927255.1	GCA_977927285.1	GCA_963921495.1	GCA_002263795.4
Assembly size (contigs)		3.08 Gb	3.11 Gb	3.11 Gb	3.11 Gb	3.14 Gb	3.17 Gb	3.19 Gb	NA
Assembly size (chromosomes)		2.80 Gb	2.83 Gb	2.77 Gb	2.81 Gb	2.77 Gb	2.81 Gb	2.78 Gb	2.77 Gb
# Contigs		846	619	1380	720	1356	810	712	2343
# Contigs placed in chromosomes		266	82	239	124	214	118	NA	NA
Contig N_{50}		81.1 Mb	79.8 Mb	88.6 Mb	73.8 Mb	84.6 Mb	76.9 Mb	93.3 Mb	26.4 Mb
Chromosomes included		29 + X	29 + X + Mito	29 + X	29 + X + Mito	29 + X	29 + X+ Mito	29 + X + Y + Mito	29 + X + Y + Mito
Gapless chromosomes		2, 9, 11, 12, 13, 22, 24, 27	2, 5, 12, 16, 18, 20, 22, 26, 27, Mito	2, 5, 7, 8, 10, 12, 13, 16, 19, 22, 26, 29	7, 8, 10, 13, 22, 23, 24, Mito	2, 5, 7, 8, 10, 13, 16, 22, 23, 25, 26	7, 8, 13, 18, 19, 20, 22, 24, Mito	ALL	NA
K-mer-based QV		69.5	68.1	68.2	68.2	69.4	68.1		
		68.8		68.2		68.7		NA	NA
K-mer completeness		94.7	95.8	94.3	95.0	95.3	96.1		
		99.7		99.7		99.7		NA	NA
Hamming error rate		0.46%	0.51%	1.03%	2.52%	1.05%	2.19%	NA	NA
BUSCO[a]	S	96.95%, 8945	97.27%, 8974	97.63%, 9007	95.82%, 8840	98.11%, 9052	97.70%, 9014	98.63%, 9100	98.53%, 9090
	D	1.12%, 103	1.08%, 100	1.18%, 109	1.25%, 115	1.08%, 100	1.16%, 107	1.14%, 105	1.22%, 113
	F	0.25%, 23	0.22%, 20	0.23%, 21	0.20%, 18	0.21%, 19	0.24%, 22	0.18%, 17	0.21%, 19
	I	0.00%, 0	0.00%, 0	0.00%, 0	0.00%, 0	0.00%, 0	0.00%, 0	0.00%, 0	0.00%, 0
	M	1.68%, 155	1.43%, 132	0.96%, 89	2.74%, 253	0.60%, 55	0.90%, 83	0.04%, 4	0.04%, 4

[a]BUSCO scores: Single-copy complete genes (S). Duplicated complete genes (D). Fragmented genes, subclass 1 – only proportion of gene aligns (F). Fragmented genes, subclass 2—gene aligns to two positions (I). Missing genes (M).

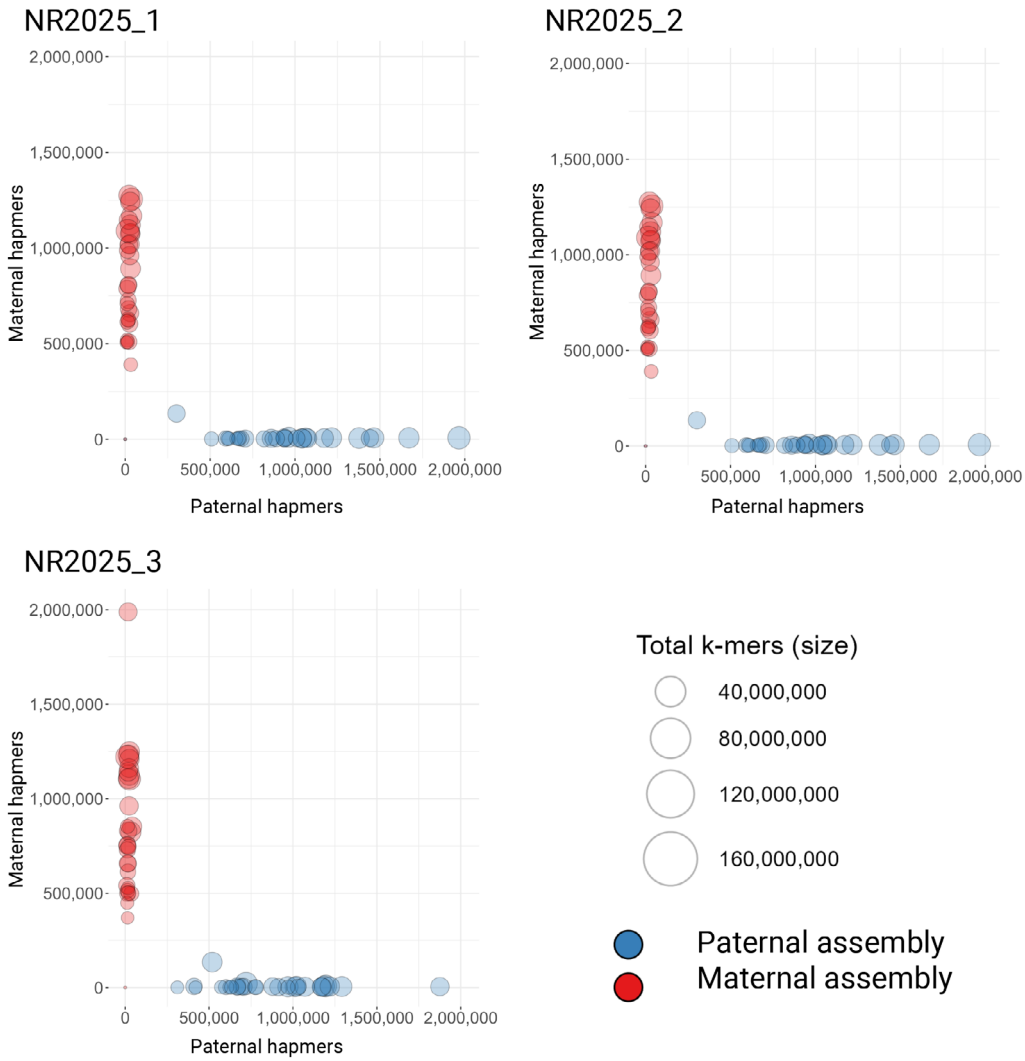

FIGURE 1 | Hapmer blob plots describing accuracy of haplotype separation. Chromosomes from the maternal (red) and paternal (blue) haplotype assemblies are positioned based on number of maternal (*y*-axis) and paternal (*x*-axis) hapmers mapped to the assemblies, while strong alignment with either axis indicates more accurate haplotype separation.

from 0.46% to 2.52% across assemblies (Table S7). The haplotype assembly pair NR2025_1P/NR2025_1M shows a lower error rate than the NR2025_2P/NR2025_2M and NR2025_3P/NR2025_3M pairs. This difference is likely due to differences in read lengths of HiFi reads. HiFi reads generated from the genome of offspring 1 have a read length N50 value ~1500–2000 bp longer than offspring 2 and 3, which improves the local haplotype separation within the read graph during assembly (Cheng et al. 2021).

All paternal X chromosomes show a higher hamming error rate compared to remaining chromosomes (Table S7), in line with the previously described mis-assembly of the paternal X chromosome due to an imbalance in hapmers identified from paternal short-reads.

3.4 | Characterisation of Centromere and Telomere Regions

Assembly-to-assembly alignments were performed to identify regions of diverging sequence and structural rearrangements between the NR2025 assemblies. Across each haplotype pair, ~2.6 Gb of sequence was found to be syntenic (Table S8). Most non-syntenic regions, including larger structural rearrangements and unaligned sequence, are found at the beginning of the autosomal chromosomes (Figure 2a).

The complex architecture at the start of chromosomes is consistent with their acrocentric nature with terminal centromeres consisting of highly variable and repetitive sequence (Escudeiro, Adega, et al. 2019; Frohlich et al. 2017). This also implies that the NR2025 assemblies successfully span into the centromeric regions, which are notoriously difficult to resolve due to their repetitive architecture, and thus represent a significant advancement towards complete telomere-to-telomere (T2T) representation of the NR genome.

To further characterise the composition of the centromeres, bovine satellite sequences were mapped to the NR2025 assemblies. Across the assemblies, 22–27 of the 29 autosomes show enrichment of satellites within the first 25 Mb coinciding with the highly divergent regions observed in the alignment plot (Figure 2 and Table S9).

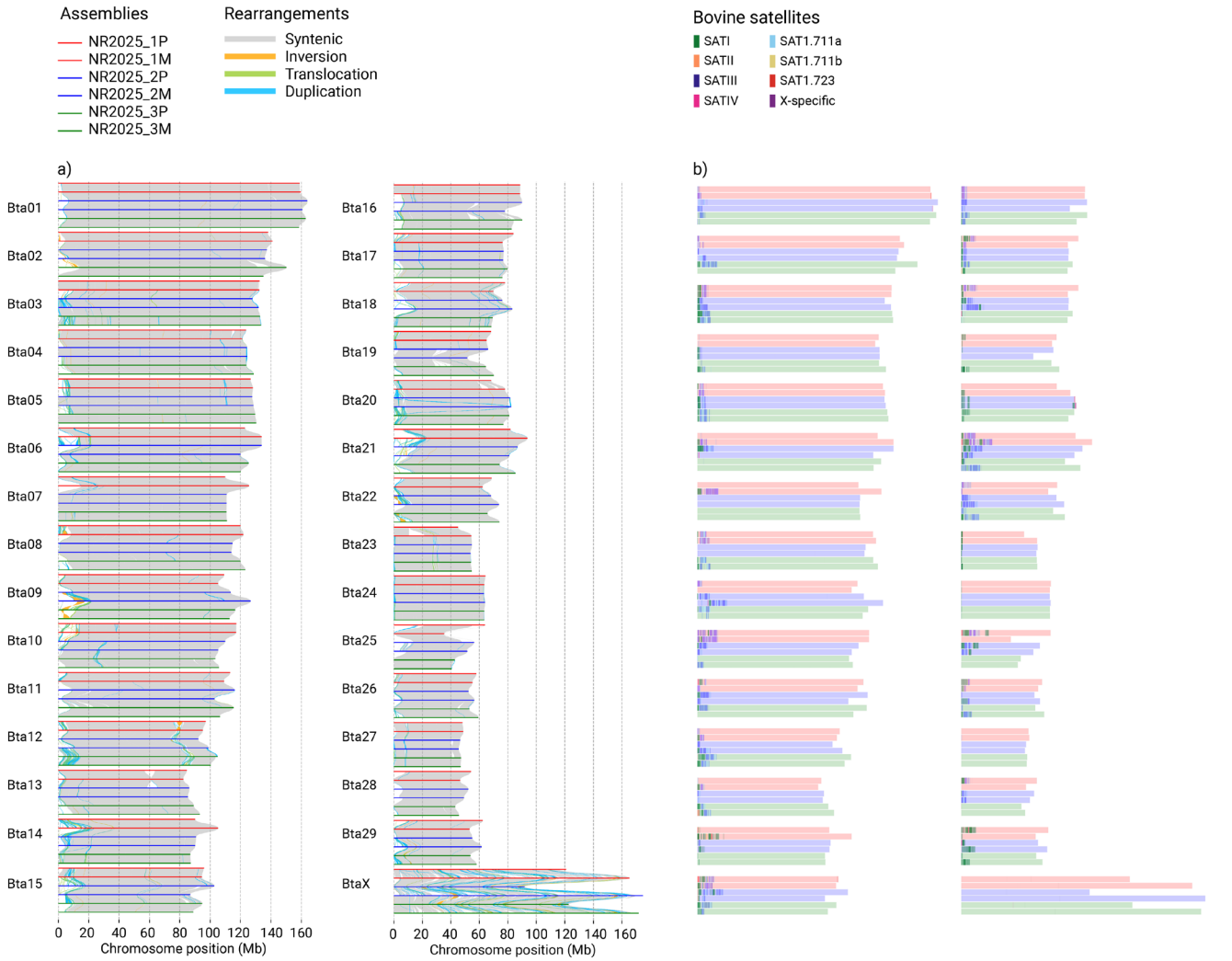

FIGURE 2 | (a) Assembly-to-assembly alignment between NR2025 assemblies including annotation of syntenic regions and larger rearrangements. (b) Mapping of bovine satellites (SATI, SATII, SATIII, SATIV, SAT1.711a, SAT1.711b, SAT1.723, X-specific) to individual NR2025 assemblies.

Satellite types are not randomly dispersed across the centromeres but tend to cluster into higher order repeat (HOR) structures composed of one or two tandemly repeated units (Table S10). Although the length of these HORs varies between chromosomes and assemblies, they appear to follow a somewhat consistent organisational pattern along the centromeres. Across the assemblies, 130 of the 174 autosomes can be assigned to ten centromeric categories based on the general order of satellites along the centromere (Figure 3 and Table S10). The remaining 44 autosomes either show no clear enrichment of satellites (23 cases) or exhibit unique organisational patterns (21 cases), which may represent additional centromeric categories or possible mis-assemblies (Table S10).

From the main centromeric categories, the predominant physical order of satellites can be inferred. SATI generally occurs upstream of SATIII + SAT1.711a, while SATIV is observed at the distal end of the centromere, upstream of SATIII + SAT1.711a (Figure 3). Notably, this observed order aligns with previously reported physical mapping of bovine SATI, SATIII, and SATIV satellites on domestic cattle chromosomes (Escudeiro, Ferreira, et al. 2019). The X-specific satellite unit is detected exclusively within BtaX (Figure 3 and Table S10), confirming the assembly of the metacentric centromere in all NR2025 X chromosomes.

In a complete T2T assembly, we expect a region of tandemly repeated telomere sequence at both chromosome ends. Across the 180 NR2025 haploid chromosomes, only Bta15 (NR2025_1P) and BtaX (NR2025_3M) show evidence of being assembled from T2T (Figure 4). Most autosomes (15–22 out of 29) include a region (0.5–10 kb) of tandemly repeated telomeric sequence at the non-centromere end (Figure 4 and Table S11). The absence of telomeres at the centromere end in remaining autosomes indicates incomplete assembly of the acrocentric short arms, requiring additional efforts to achieve complete T2T representation.

4 | Discussion

Genomic studies on Norwegian Red (NR) cattle have largely been based on the public cattle reference—a pseudohaploid representation of the genome of a Hereford cow. To provide a more accurate representation of the NR genome, we present a collection of six NR breed-specific and haplotype-resolved

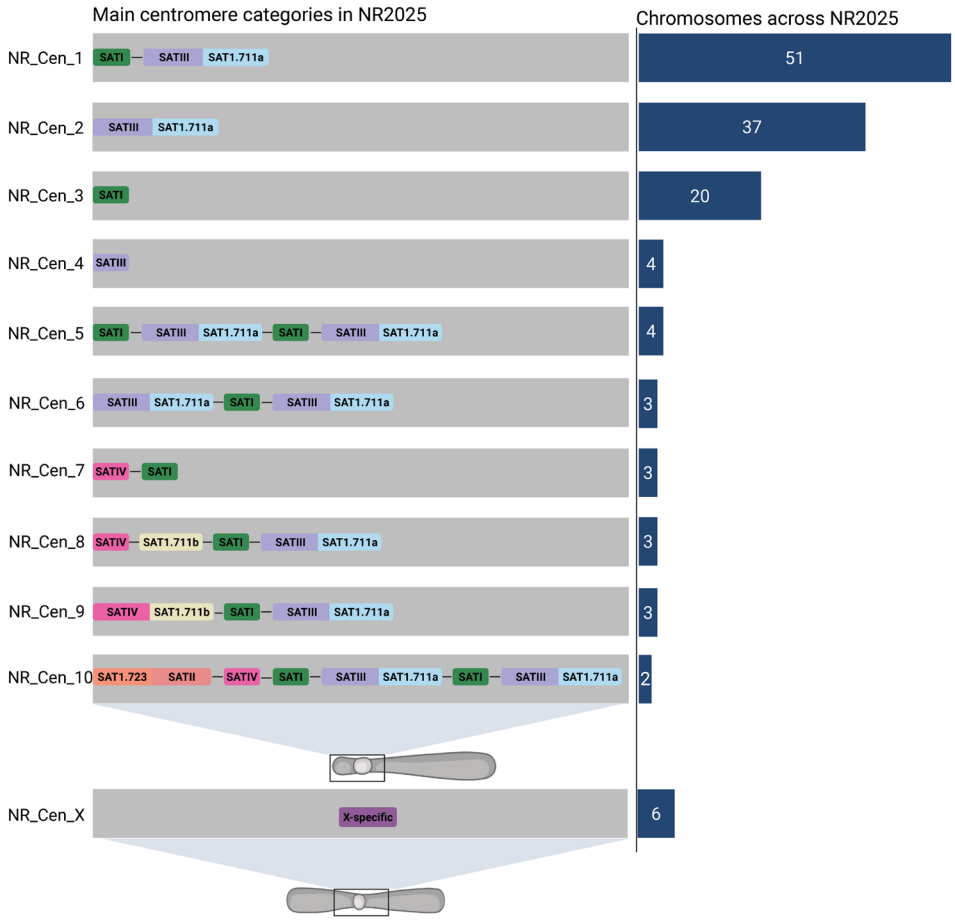

FIGURE 3 | Main categories of centromere organisation observed across the acrocentric autosomes and metacentric X chromosomes in NR2025. Only organisational patterns observed in more than one chromosome are included.

assemblies (NR2025). *De novo* assembly of long-read data (HiFi + ONT) with the trio-binning method achieves highly contiguous and complete assemblies representing genome-wide NR haplotypes.

NR2025 is the first set of assemblies representing the diploid NR genome in an uncollapsed manner. The haplotype resolution is necessary to capture the true spectrum of structural variation between a set of genomes (Chaisson et al. 2019; Huddleston et al. 2017; Ebert et al. 2021) and can enable more precise identification of causal variants in NR cattle, as shown in other cattle breeds and subspecies (Low et al. 2020; Leonard et al. 2024, 2022). Previous studies have used data from an F1 hybrid of two different subspecies or breeds to construct haplotype-resolved assemblies, due to a higher level of within-genome heterozygosity (Koren et al. 2018; Low et al. 2020; Rice et al. 2020). However, our results show that the haplotype-specific information contained in short-reads from the NR parents is generally enough to successfully assemble both haplotypes of the NR offspring. A few exceptions are observed in regions with low inter-haplotype heterozygosity, which are erroneously collapsed into one haplotype, and the paternal assembly of the X chromosome.

The NR2025 assemblies feature an improved overview of the highly repetitive centromeric regions, which have been absent or only partially assembled in previous cattle assemblies (Kalbfleisch et al. 2024). The acrocentric centromere is represented in the assembly of most autosomes, offering the current best overview of the autosomal centromere structure in NR cattle. Furthermore, mapping of bovine satellite sequences reveals that the centromeres are composed of distinct clusters of satellite units forming higher order repeat (HOR) structures. These HORs are further arranged along the centromere, following specific organisational patterns. Although centromeres are usually "gene-free" regions (Talbert and Henikoff 2020), they play an essential role in the segregation of chromosomes during mitosis and meiosis, and proper function is critical for genome stability, fertility, and development (McKinley and Cheeseman 2016). In NR cattle, specifically, there have been incidents of fusions between the acrocentric Bta01 and Bta29 chr, called Robertsonian translocations. In 1966, rob(1;29) was discovered in an NR sire, causing low fertility in daughters (Gustavsson 1979). Consequently, all AI bulls have been tested for the translocation, eliminating it from the NR population (Refsdal et al. 2014). Having a detailed characterisation of the composition and organisation of bovine centromeres will enhance our understanding of such chromosomal abnormalities. Thus, the NR2025 assemblies represent a valuable starting point to further explore the effect of centromere structure on bovine fertility and health traits.

The NR2025 collection of haplotype-resolved assemblies provides a unique basis for an NR *pangenome*. While a single

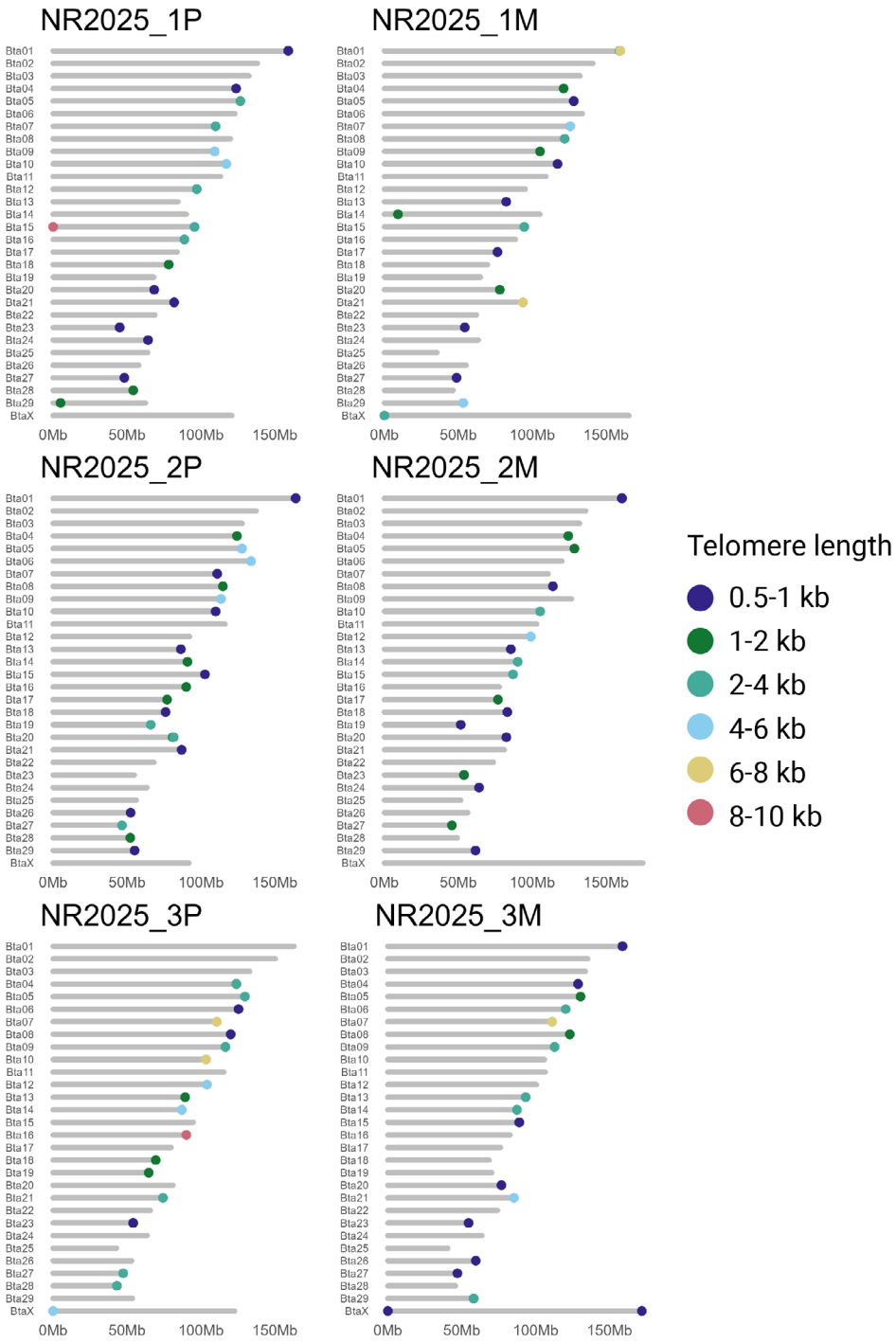

FIGURE 4 | Regions of tandemly repeated telomeric sequences (TTAGGG/CCCTAA). Only regions with a telomere repeat density (#telomere repeats × 6/length) above 50% are indicated.

linear reference is limited in the detection of large complex variants and variants significantly different from the reference sequence, a pangenome is able to represent the full range of genetic variation covered by the assemblies included (Hickey et al. 2024; Ebler et al. 2022; Gong et al. 2023). Its graph-based structure preserves the genome-wide haplotype information contained in the assemblies, enabling more accurate genotyping through haplotype-aware sequence alignment (Ebler et al. 2022; Chandra et al. 2024). The NR2025 collection of haplotype-resolved assemblies can be combined into a common pangenome and further expanded with additional haplotypes to capture the full spectrum of genetic variation in NR cattle. The utility of using a pangenome to uncover genotype–phenotype relationships has already been demonstrated across cattle breeds (Leonard et al. 2024; Sorin et al. 2025). Hence, we believe that an NR pangenome will provide an important resource to identify and link novel structural variants to phenotypes of interest in the NR population.

Author Contributions

Thea Johanna Hettasch: investigation, writing – original draft, methodology, writing – review and editing, formal analysis, data curation, conceptualization, visualization, validation, software. **Matthew Peter Kent:** writing – review and editing, supervision, conceptualization, investigation. **Arne Bjørke Gjuvsland:** conceptualization, writing – review and editing, investigation, supervision, resources. **Mariann Árnyasi:** methodology, writing – review and editing, investigation, resources. **Dag Inge Våge:** conceptualization, investigation, funding acquisition, writing – review and editing, project administration, supervision.

Acknowledgments

Computational analyses were performed on the Orion HPC provided by the Norwegian University of Life Sciences (NMBU) (https://orion.nmbu.no). We also want to thank Thu-Hien To from Elixir Norway (https://elixir.no/) for her help with submitting assemblies and read files to ENA.

Funding

The study was funded by the European Union (EU) as a part of the RUMIGEN project (grant no. 7551000128). TJH was funded by a PhD grant from the Norwegian University of Life Sciences (NMBU).

Ethics Statement

Collection of blood samples from trio dams and daughters was approved by the Norwegian Food Safety Authorities (Mattilsynet, FOTS-ID: 30618). Collection of blood and semen samples from sires was approved and provided by the NR breeding company Geno as part of their routine sampling at the AI bull farm.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Assemblies and read files are available on the European Nucleotide Archive (ENA) under project accession number PRJEB105120 (https://www.ebi.ac.uk/ena/browser/view/PRJEB105120).

References

Alonge, M., S. Soyk, S. Ramakrishnan, et al. 2019. "RaGOO: Fast and Accurate Reference-Guided Scaffolding of Draft Genomes." *Genome Biology* 20: 1–17.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215, no. 3: 403–410.

Andrews, S. 2010. "FastQC: A Quality Control Tool for High Throughput Sequence Data." Cambridge, United Kingdom.

Brown, M., P. González De la Rosa, and B. Mark. 2023. "A Telomere Identification Toolkit." *Zenodo*.

Carey, S. B., J. T. Lovell, J. Jenkins, et al. 2022. "Representing Sex Chromosomes in Genome Assemblies." *Cell Genomics* 2, no. 5: 100132.

Chaisson, M. J., A. D. Sanders, X. Zhao, et al. 2019. "Multi-Platform Discovery of Haplotype-Resolved Structural Variation in Human Genomes." *Nature Communications* 10, no. 1: 1784.

Chandra, G., D. Gibney, and C. Jain. 2024. "Haplotype-Aware Sequence Alignment to Pangenome Graphs." *Genome Research* 34, no. 9: 1265–1275.

Chen, N.-C., B. Solomon, T. Mun, S. Iyer, and B. Langmead. 2021. "Reference Flow: Reducing Reference Bias Using Multiple Population Genomes." *Genome Biology* 22: 1–17.

Chen, S., Y. Zhou, Y. Chen, and J. Gu. 2018. "Fastp: An Ultra-Fast All-In-One FASTQ Preprocessor." *Bioinformatics* 34, no. 17: i884–i890.

Cheng, H., G. T. Concepcion, X. Feng, H. Zhang, and H. Li. 2021. "Haplotype-Resolved de Novo Assembly Using Phased Assembly Graphs With Hifiasm." *Nature Methods* 18, no. 2: 170–175.

Ebert, P., P. A. Audano, Q. Zhu, et al. 2021. "Haplotype-Resolved Diverse Human Genomes and Integrated Analysis of Structural Variation." *Science* 372, no. 6537: eabf7117.

Ebler, J., P. Ebert, W. E. Clarke, et al. 2022. "Pangenome-Based Genome Inference Allows Efficient and Accurate Genotyping Across a Wide Spectrum of Variant Classes." *Nature Genetics* 54, no. 4: 518–525.

Elsik, C. G., R. L. Tellam, K. C. Worley, et al. 2009. "The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution." *Science* 324, no. 5926: 522–528.

Escudeiro, A., F. Adega, T. J. Robinson, J. S. Heslop-Harrison, and R. Chaves. 2019. "Conservation, Divergence, and Functions of Centromeric Satellite DNA Families in the Bovidae." *Genome Biology and Evolution* 11, no. 4: 1152–1165.

Escudeiro, A., D. Ferreira, A. Mendes-da-Silva, J. Heslop-Harrison, F. Adega, and R. Chaves. 2019. "Bovine Satellite DNAs–a History of the Evolution of Complexity and Its Impact in the Bovidae Family." *European Zoological Journal* 86, no. 1: 20–37.

Espinosa, E., R. Bautista, R. Larrosa, and O. Plata. 2024. "Advancements in Long-Read Genome Sequencing Technologies and Algorithms." *Genomics* 20: 110842.

Formenti, G., L. Abueg, A. Brajuka, et al. 2022. "Gfastats: Conversion, Evaluation and Manipulation of Genome Sequences Using Assembly Graphs." *Bioinformatics* 38, no. 17: 4214–4216.

Frohlich, J., S. Kubickova, P. Musilova, et al. 2017. "Karyotype Relationships Among Selected Deer Species and Cattle Revealed by Bovine FISH Probes." *PLoS One* 12, no. 11: e0187559.

Goel, M., and K. Schneeberger. 2022. "Plotsr: Visualizing Structural Similarities and Rearrangements Between Multiple Genomes." *Bioinformatics* 38, no. 10: 2922–2926.

Goel, M., H. Sun, W.-B. Jiao, and K. Schneeberger. 2019. "SyRI: Finding Genomic Rearrangements and Local Sequence Differences From Whole-Genome Assemblies." *Genome Biology* 20: 1–13.

Gong, Y., Y. Li, X. Liu, Y. Ma, and L. Jiang. 2023. "A Review of the Pangenome: How It Affects Our Understanding of Genomic Variation, Selection and Breeding in Domestic Animals?" *Journal of Animal Science and Biotechnology* 14, no. 1: 73.

Guo, M. H., L. C. Francioli, S. L. Stenton, et al. 2024. "Inferring Compound Heterozygosity From Large-Scale Exome Sequencing Data." *Nature Genetics* 56, no. 1: 152–161.

Gustavsson, I. 1979. "Distribution and Effects of the 1/29 Robertsonian Translocation in Cattle." *Journal of Dairy Science* 62, no. 5: 825–835.

Hickey, G., J. Monlong, J. Ebler, et al. 2024. "Pangenome Graph Construction From Genome Alignments With Minigraph-Cactus." *Nature Biotechnology* 42, no. 4: 663–673.

Hu, T., N. Chitnis, D. Monos, and A. Dinh. 2021. "Next-Generation Sequencing Technologies: An Overview." *Human Immunology* 82, no. 11: 801–811.

Huang, N., and H. Li. 2023. "Compleasm: A Faster and More Accurate Reimplementation of BUSCO." *Bioinformatics* 39, no. 10: btad595.

Huddleston, J., M. J. Chaisson, K. M. Steinberg, et al. 2017. "Discovery and Genotyping of Structural Variation From Long-Read Haploid Genome Sequence Data." *Genome Research* 27, no. 5: 677–685.

Jain, C. 2023. "Coverage-Preserving Sparsification of Overlap Graphs for Long-Read Assembly." *Bioinformatics* 39, no. 3: btad124.

Kalbfleisch, T. S., S. D. McKay, B. M. Murdoch, et al. 2024. "The Ruminant Telomere-To-Telomere (RT2T) Consortium." *Nature Genetics* 56, no. 8: 1566–1573.

Kamath, S. S., M. Bindra, D. Pal, and C. Jain. 2024. "Telomere-To-Telomere Assembly by Preserving Contained Reads." *Genome Research* 34, no. 11: 1908–1918.

Kirsanova, E., B. Heringstad, A. Lewandowska-Sabat, and I. Olsaker. 2020. "Identification of Candidate Genes Affecting Chronic Subclinical Mastitis in Norwegian Red Cattle: Combining Genome-Wide Association Study, Topologically Associated Domains and Pathway Enrichment Analysis." *Animal Genetics* 51, no. 1: 22–31.

Koren, S., A. Rhie, B. P. Walenz, et al. 2018. "De Novo Assembly of Haplotype-Resolved Genomes With Trio Binning." *Nature Biotechnology* 36, no. 12: 1174–1182.

Leonard, A. S., D. Crysnanto, Z.-H. Fang, et al. 2022. "Structural Variant-Based Pangenome Construction Has Low Sensitivity to Variability of Haplotype-Resolved Bovine Assemblies." *Nature Communications* 13, no. 1: 3012.

Leonard, A. S., X. M. Mapel, and H. Pausch. 2024. "Pangenome-Genotyped Structural Variation Improves Molecular Phenotype Mapping in Cattle." *Genome Research* 34, no. 2: 300–309.

Li, H. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." *Bioinformatics* 34, no. 18: 3094–3100.

Li, H. 2020. "yak: Yet Another k-Mer Analyzer." GitHub. https://github.com/lh3/yak.

Li, H., and R. Durbin. 2024. "Genome Assembly in the Telomere-To-Telomere Era." *Nature Reviews Genetics* 25, no. 9: 658–670.

Lin, M.-J., S. Iyer, N.-C. Chen, and B. Langmead. 2024. "Measuring, Visualizing, and Diagnosing Reference Bias With Biastools." *Genome Biology* 25, no. 1: 101.

Low, W. Y., R. Tearle, R. Liu, et al. 2020. "Haplotype-Resolved Genomes Provide Insights Into Structural Variation and Gene Content in Angus and Brahman Cattle." *Nature Communications* 11, no. 1: 2071.

McKinley, K. L., and I. M. Cheeseman. 2016. "The Molecular Basis for Centromere Identity and Function." *Nature Reviews. Molecular Cell Biology* 17, no. 1: 16–29.

Olsen, H. G., T. M. Knutsen, A. Kohler, et al. 2017. "Genome-Wide Association Mapping for Milk Fat Composition and Fine Mapping of a QTL for de Novo Synthesis of Milk Fatty Acids on Bovine Chromosome 13." *Genetics Selection Evolution* 49, no. 1: 20.

Olsen, H. G., T. M. Knutsen, A. M. Lewandowska-Sabat, et al. 2016. "Fine Mapping of a QTL on Bovine Chromosome 6 Using Imputed Full Sequence Data Suggests a Key Role for the Group-Specific Component (GC) Gene in Clinical Mastitis and Milk Production." *Genetics Selection Evolution* 48, no. 1: 79.

Refsdal, A. O., P. Gillund, and K. Karlberg. 2014. *Fruktbarhet i fjøset*. Fagbokforl.

Rhie, A., S. A. McCarthy, O. Fedrigo, et al. 2021. "Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species." *Nature* 592, no. 7856: 737–746.

Rhie, A., B. P. Walenz, S. Koren, and A. M. Phillippy. 2020. "Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies." *Genome Biology* 21: 1–27.

Rice, E. S., S. Koren, A. Rhie, et al. 2020. "Continuous Chromosome-Scale Haplotypes Assembled From a Single Interspecies F1 Hybrid of Yak and Cattle." *GigaScience* 9, no. 4: giaa029.

Rosen, B. D., D. M. Bickhart, R. D. Schnabel, et al. 2020. "De Novo Assembly of the Cattle Reference Genome With Single-Molecule Sequencing." *GigaScience* 9, no. 3: 128.

Sim, S. B., R. L. Corpuz, T. J. Simmonds, and S. M. Geib. 2022. "HiFiAdapterFilt, a Memory Efficient Read Processing Pipeline, Prevents Occurrence of Adapter Sequence in PacBio HiFi Reads and Their Negative Impacts on Genome Assembly." *BMC Genomics* 23, no. 1: 157.

Smith, T. P., D. M. Bickhart, D. Boichard, et al. 2023. "The Bovine Pangenome Consortium: Democratizing Production and Accessibility of Genome Assemblies for Global Cattle Breeds and Other Bovine Species." *Genome Biology* 24, no. 1: 139.

Sorin, V., M. M. Naji, C. Birbes, et al. 2025. "Application of a French Cattle Pangenome, From Structural Variant Discovery to Association Studies on Key Phenotypes." *Genetics, Selection, Evolution* 57, no. 1: 61.

Talbert, P. B., and S. Henikoff. 2020. "What Makes a Centromere?" *Experimental Cell Research* 389, no. 2: 111895.

Thorburn, D. M. J., K. Sagonas, M. Binzer-Panchal, et al. 2023. "Origin Matters: Using a Local Reference Genome Improves Measures in Population Genomics." *Molecular Ecology Resources* 23, no. 7: 1706–1723.

van Den Berg, I., R. Xiang, J. Jenko, et al. 2020. "Meta-Analysis for Milk Fat and Protein Percentage Using Imputed Sequence Variant Genotypes in 94,321 Cattle From Eight Cattle Breeds." *Genetics Selection Evolution* 52, no. 1: 37.

Wick, R. 2017. "Filtlong."

Zhang, X., R. Wu, Y. Wang, J. Yu, and H. Tang. 2020. "Unzipping Haplotypes in Diploid and Polyploid Genomes." *Computational and Structural Biotechnology Journal* 18: 66–72.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Figure S1:** Image of agarose gel electrophoresis done on high molecular weight DNA prior to ONT sequencing. 400 ng DNA was loaded from each sample and run for 2 h at 50 V on 0.5% agarose gel. Left hand side ladder is a GeneRuler 1 kb Plus DNALadder and right-hand side ladder is a GeneRuler HighRange DNALadder. **Figure S2:** (a) Assembly-to-assembly alignment between NR2025 assemblies. Examples of larger regions missing from one haplotype assembly and present in the other is observed in the assembly of Bta13 in NR2025_1M and in the assembly of Bta23 in NR2025_1P. (b) Pairwise assembly alignment of region missing in one haplotype. The position of BUSCO genes reported as missing from one haplotype assembly and present in the other are represented by grey lines in the plot. Unique maternal (red) and paternal (blue) hapmers are mapped to the assemblies and counted across 100 kb windows. Regions which are presumably collapsed erroneously into one haplotype cover most missing BUSCO genes and seem to coincide with less heterozygous regions with few parental hapmers. **Note S1** Description of commands used to perform analyses related to QC and filtering of reads, de novo assembly with trio-binning and evaluation of assembly quality and structure. **Note S2** Search setting for identification of bovine satellite sequences from NCBI. **Table S1:** Software tools applied in study. **Table S2:** Contaminants reported in Illumina short-reads by FastQC. **Table S3:** Bovine satellites identified from NCBI. All satellites, including fragments and partial sequences are categorised into common satellite types based on the NCBI description. Shared identity of satellite sequences with different names reported in literature is referenced. **Table S4:** Statistics of raw sequencing data. **Table S5:** Length of chromosomes (bp) included in NR2025 assemblies. **Table S6:** BUSCO genes (from OrthoDB v10.1) reported as missing from assemblies by Compleasm. **Table S7:** Maternal and paternal hapmers found in NR2025 assemblies using the k-mer based tool Merqury. The

hamming error rate across the whole assembly (global) and individual chromosomes is calculated based on the number of correct and erroneous parental hapmers found in the assemblies. **Table S8:** Syntenic regions and rearrangements reported by SyRI based on pairwise assembly-to-assembly alignments performed by Minimap2 between NR2025 haplotype pairs. **Table S9:** Regions of satellite enrichment across NR2025 assemblies in 1 Mb windows. Enriched regions have a significantly higher number of satellite sequence mappings than the average number of mappings across the genome. The enrichment analysis is based on a simple binomial with an adjusted *P*-value threshold of 0.05. **Table S10:** Satellites mapped to centromere region of acrocentric autosomes and metacentric X chromosome. The organisation of satellite units into HORs and HORs into centromere categories is provided. **Table S11:** Regions with telomere repeat density > 0.5. The repeat density is calculated as: (#repeats * 6)/length.