Novel functional sequences uncovered through a bovine multi-assembly graph

Danang Crysnanto*, Alexander S. Leonard, Zih-Hua Fang, Hubert Pausch

Animal Genomics, ETH Zürich, Zürich, 8315 Switzerland

Accepted in
Proceedings of the National Academy of Sciences of the United States of America

Abstract

Many genomic analyses start by aligning sequencing reads to a linear reference genome. However, linear reference genomes are imperfect, lacking millions of bases of unknown relevance, and are unable to reflect the genetic diversity of populations. This makes reference-guided methods susceptible to reference-allele bias. To overcome such limitations, we build a pangenome from six reference-quality assemblies from taurine and indicine cattle as well as yak. The pangenome contains an additional 70,329,827 bases compared to the *Bos taurus* reference genome. Our multi-assembly approach reveals 30 and 10.1 million bases private to yak and indicine cattle, respectively, and between 3.3 and 4.4 million bases unique to each taurine assembly. Utilizing liver transcriptomes from 56 cattle, we show that the novel sequences encode transcripts that hitherto remained undetected from the *Bos taurus* reference genome. We uncover novel genes, primarily encoding proteins contributing to immune response and pathogen-mediated immunomodulation, differentially expressed between *Mycobacterium bovis*-infected and non-infected cattle that are also undetectable in the *Bos taurus* reference genome. Using whole-genome sequencing data of cattle from five breeds, we show that reads which were previously misaligned against the bovine reference genome now align accurately to the novel sequences. This enables us to discover 83,250 polymorphic sites that segregate within and between breeds of cattle and capture genetic differentiation across breeds. Our work makes a so far unused source of variation amenable to genetic investigations and provides methods and a framework for establishing and exploiting a more diverse reference genome.

Keywords:Genetic diversity, Genome graphs, Pangenome

Significance

Most sequence variant analyses rely on a linear reference genome that is assumed to lack millions of bases that occur in the genomes of other individuals. To quantify the extent and functional relevance of such missing bases, we integrate six genome assemblies from cattle and related species into a pangenome. This allows us to uncover more than 70 million bases that are not included in the *Bos taurus* reference genome. Through complementary bioinformatics, genomics, and transcriptomics methods we discover novel genes that are differentially expressed and thousands of polymorphic sites that were unused so far. Our work provides a computational framework, broadly applicable to many species, to make a so far neglected source of genomic variation amenable to genetic investigations.

0.1 Introduction

A well-annotated reference genome enables systematic characterization of sequence variation within and between populations, as well as across species. The reference genome of domestic cattle (Bos taurus taurus) was generated from the inbred Hereford cow *L1 Dominette 01449* (Sequencing and Consortium, 2009). Long-read sequencing and sophisticated genome assembly methods have enabled spectacular improvements in the contiguity and quality of the *Bos taurus* reference genome. The contig (contiguous sequence formed by overlapping reads without gaps) N50 size (i.e., 50% of the genome is in contigs of this size or greater) of the bovine reference genome has increased from kilo- to megabases over the past five years (Rosen et al., 2020). Recent method and sequencing technology developments have facilitated the assembly of multiple reference-quality genomes. The application of trio-binning (Koren et al., 2018) resulted in chromosome-scale haplotype-resolved assemblies for three taurine (Hereford, Angus, Highland cattle) and one indicine (Brahman) cattle breeds, as well as for yak (*Bos grunniens*), a closely related species to domestic cattle (Low et al., 2020; Rice et al., 2020).

DNA sequences from taurine and indicine cattle are typically aligned to the Hereford-based reference genome to discover and genotype variable sites. Reference-guided read alignment and variant genotyping has revealed millions of polymorphic variants that segregate within and between taurine and indicine cattle breeds (Kim et al., 2020; Daetwyler et al., 2014; Koufariotis et al., 2018). However, using the linear reference in this alignment approach is susceptible to reference allele bias, particularly for DNA samples that are greatly diverged from the reference (Ballouz et al., 2019; Pritt et al., 2018). Moreover, reference-guided methods are blind to variations in sequences that are not present in the

Table 1: **Details of six bovine genome assemblies**

Assembly (Species)	Sex[1]	Primary data[2]	Assembly type	Assembler	Contig N50(Mb)	Scaffold N50(Mb)	Autosomes lengths (Gb)
Hereford (*Bos taurus taurus)*	F	PacBio (80-fold CLR)	Primary	Falcon	21	108	2.489
Angus (*Bos taurus taurus)*	M	PacBio (136-fold CLR)	Haplotype resolved	TrioCanu	29.4	102.8	2,468
Highland (*Bos taurus taurus)*	F	PacBio (125-fold CLR)	Haplotype resolved	TrioCanu	71.7	86.2	2,483
Original Braunvieh (*Bos taurus taurus)*	F	PacBio (28-fold HiFi)	Primary	Hifiasm	86.0	96.3	2,607
Brahman (*Bos taurus indicus)*	F	PacBio (136-fold CLR)	Haplotype resolved	TrioCanu	23.4	104.5	2,478
Yak (*Bos grunniens*)	F	PacBio (125-fold CLR)	Haplotype resolved	TrioCanu	70.9	94.7	2,478

[1] Female (F) and male (M) assemblies contain either X or Y chromosomal sequences.
[2] Additional data may have been used to polish the assemblies and facilitate scaffolding; CLR: continuous long reads; HiFi: high-fidelity.

reference genome (Wong et al., 2020). Recent estimates suggest that millions of bases are missing in mammalian reference genomes (Sherman et al., 2019; Whitacre et al., 2015), indicating a high potential for bias.

Efforts to mitigate reference allele bias and increase the genetic diversity of reference genomes have led to graph-based references (Garrison et al., 2018; Eggertsson et al., 2017). We have previously shown that a genome graph, which integrates linear reference coordinates and pre-selected variants, improves the mapping of reads and enables unbiased variant genotyping in different breeds of cattle (Crysnanto et al., 2019; Crysnanto and Pausch, 2020). However, previous attempts focused on augmenting the *Bos taurus* reference genome with small variations (<50bp), not the larger class of structural variations. Despite being an important source of genotypic and phenotypic diversity (Song et al., 2020; Kehr et al., 2017), little is known about the prevalence and functional impact of structural variations in the cattle genome. The availability of reference-quality assemblies and long read sequencing data from different breeds of cattle now provides an opportunity to characterize sequence diversity beyond small variations (Hickey et al., 2020; Li et al., 2020). In this paper, we integrate reference-quality assemblies from multiple taurine breeds as well as two close relatives into a multi-assembly graph with minigraph (Li et al., 2020) Table 1. We detect autosomal sequences that are missing in the *Bos taurus* reference genome and investigate their functional significance using transcriptome data. We show that the non-reference sequences contain novel transcripts that are differentially expressed as well as polymorphic sites that segregate within and between breeds of cattle.

0.2 Results

Construction of a bovine multi-assembly graph

We considered the Hereford-based Bos taurus reference genome and five reference-quality assemblies from three breeds of taurine (*Bos taurus taurus*) cattle (Angus, Highland, Original Braunvieh) (Koren et al., 2018; Low et al., 2020; Rice et al., 2020) and their close relatives Brahman (*Bos taurus indicus*) (Low et al., 2020) and yak (*Bos grunniens*) (Rice et al., 2020). All assemblies, except for the Original Braunvieh, were generated prior to this study. The reference-quality assembly for an Original Braunvieh female calf was created with 28-fold PacBio HiFi read coverage (see *SI Appendix*, **??**). The contig and scaffold N50 values of the six assemblies ranged from 21 to 80 Mb and 86.2 to 108 Mb, respectively 1.

The six assemblies were integrated into a multi-assembly graph with minigraph. We only considered autosomal sequences because the haplotype-resolved assemblies represent either paternal or maternal haplotypes, thus lacking either X or Y chromosomal sequences. The Hereford-based linear reference genome (ARS-UCD1.2) formed the backbone of the bovine multi-assembly graph. The graph was then augmented with the five additional assemblies, added in order of increasing Mash-distance from the ARS-UCD1.2 reference (Ondov et al., 2016) Fig. 1. Constructing this multi-assembly graph took 4.1 CPU hours and 58 GB of RAM, taking 36 minutes of wall-clock time when using 10 threads.

Recovery of non-reference sequences from the multi-assembly graph

Our bovine multi-assembly graph represents 2,558,596,439 nucleotides, spread across 182,940 nodes connected by 258,396 edges. On average, a node spans 13,985 nucleotides and is connected by 1.4 edges. Of the edges, 141,086, 113,332, and 3,978 connect two reference nodes, a reference and non-reference node, or two non-reference nodes, respectively.

The vast majority (2,489,385,779 or 97.29%) of nucleotides in the multi-assembly graph originate from the linear reference backbone, covered in 123,483 nodes. These reference nodes span 23,088 bases on average, ranging from 100 to 1,398,882 bases. The incremental integration of the Highland, Angus, Original Braunvieh, Brahman, and yak assemblies added 8,847, 4,613, 3,555, 11,996, and 30,446 non-reference nodes, respectively containing 14,679,286, 5,537,769, 7,013,258, 11,116,220, and 30,864, 127 non-reference bases. The resulting multi-assembly graph contained 59,457 non-reference nodes spanning 69,210,660 bases.

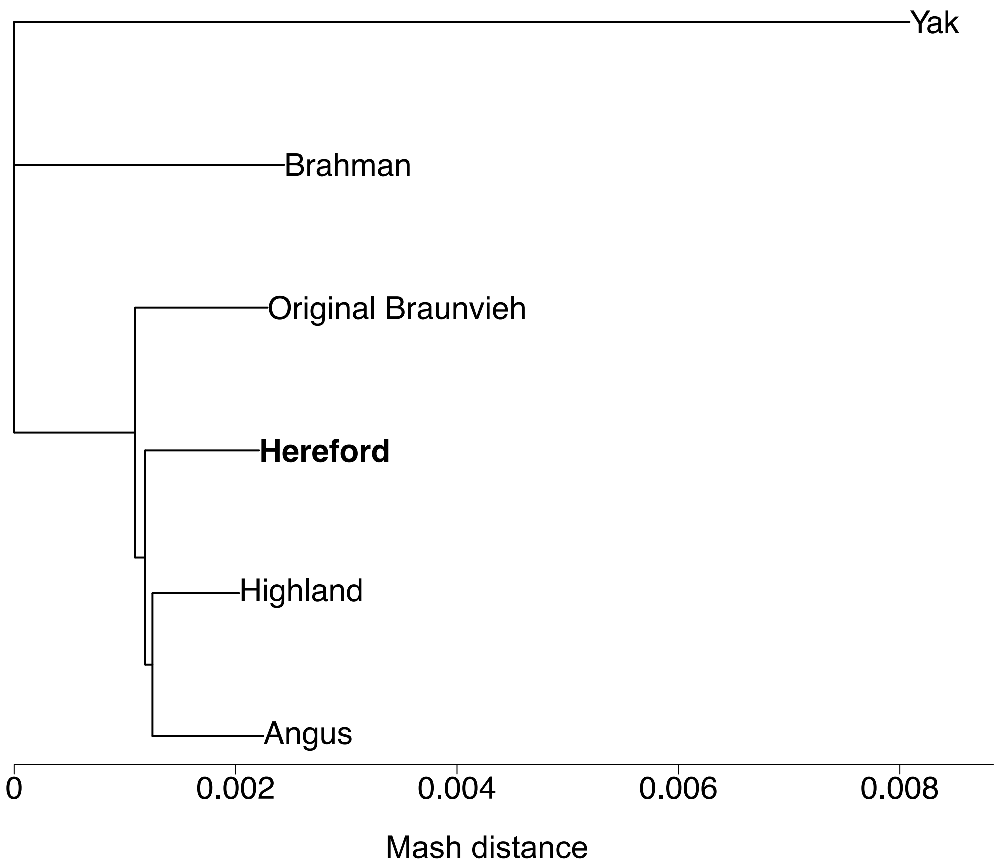

Figure 1: **Phylogenetic distance between six genome assemblies.**
A Mash-based phylogenetic tree derived from six bovine assemblies, including the current Hereford-based *Bos taurus* reference genome (**bold**). The yak assembly was used as the outgroup to root the tree during building.

To determine the support of the non-reference nodes, we aligned individual assemblies back to the multi-assembly graph. Nodes were then labelled according to which assembly path traversed them (see *SI Appendix*, Figs. **??** & **??**). This approach enabled a straightforward confirmation of minigraph's mapping accuracy. Only reference nodes should contain a Hereford label, since this assembly was used as the backbone of the graph. Mapping was highly accurate, as indicated by an F1 score of 99.97%.

The non-reference nodes of the multi-assembly graph had a cumulative length of 43,341,418, 23,644,772, 18,202,102, 14,453,112 and 15,542,368 bases in the yak, Brahman, Original Braunvieh, Angus, and Highland assemblies. Yak and Brahman non-reference nodes were shorter on average compared to the taurine assemblies (*SI Appendix*, Fig. **??**). Most non-reference nodes (41,855 or 70.40%) and non-reference sequences (42.52 Mb, 69.52%) were either private to yak (29,854 nodes, 29.9 Mb), Brahman (7,843 nodes, 8.22

Mb), or shared by both assemblies (4,158 nodes, 3.05 Mb) (S1 Appendix, Fig. S4). The Original Braunvieh, Highland, and Angus assemblies contributed 4.51, 2.78 and 2.39 Mb in 2,016, 1,938 and 1,759 nodes, respectively, that were not detected in any other assembly. The three taurine assemblies shared 668 nodes containing 0.77 Mb not detected in ARS-UCD1.2, yak, or Brahman. There were also 1,318 non-reference nodes with a cumulative length of 4.4 Mb supported by all five additional assemblies.

The core genome of the multi-assembly graph (i.e., nodes shared by all assemblies) is contained in 67,482 nodes with a cumulative length of 2,402,561,410 bases. About 6.10% of the pangenome (115,458 nodes containing 156,035,029 bases) is flexible (i.e., not shared by all assemblies). Of the flexible part, 69,697 nodes containing 97,106,100 bases are shared by at least two assemblies, and 45,761 nodes with 58,928,929 bases are only found in one assembly. The profile of the multi-assembly graph changes markedly when distant assemblies (e.g., Brahman, yak) are added (*SI Appendix*, **??**).

The minigraph approach used to construct the multi-assembly graph does depend on an initial sequence forming a backbone. The choice of backbone consequently impacts the amount of non-reference sequence detected from each additional assembly (see *SI Appendix*, **??**). However, the overall effect on the sequence content of the multi-assembly graph is relatively minor, with 68.72±3.17 Mb of non-reference sequence identified across all possible backbones.

Structural variation discovery from the multi-assembly graph

Using the bubble popping algorithm of gfatools (21), we identified 68,328 structural variations present in the multi-assembly graph. To reveal true alleles within these structural variations, we traversed all possible paths through the bubbles (i.e., alleles) and retained only those that were supported by at least one assembly (*SI Appendix*, Fig. **??**). Most of the structural variations had two alleles (64,224 or 94%). The remaining 4,104 structural variations were multi-allelic, most of which had three alleles (3,324 or 81%). We identified 141,747 alleles at the structural variations, including 73,506 non-reference alleles with a cumulative length of 74,453,929 bases.

We overlapped the breakpoints of the structural variations with the Ensembl annotation (build 101) of ARS-UCD1.2. Almost all structural variations are either intergenic (47,642 or 69.81%) or intronic (20,227 or 29.64%). There were 170 and 202 exons and coding sequences, respectively, of 338 unique genes affected by structural variations. A Panther GO-Slim Biological Process (Mi et al., 2019) analysis indicated that these genes are enriched for genes related to the adaptive immune response (4.35−fold, P=0.04), T-cell mediated immunity (6.37−fold, P=0.04), actin filament depolymerization (8.54−fold, P=6.56e−03), microtubule cytoskeleton organization (10.48−fold, P=1.85e−04), and iron-sulfur cluster assembly (9.96−fold, P=0.02).

The non-reference alleles consisted of 40,369 insertions and 33,137 deletions with an average length of 1,181 and 1,210 bases respectively (*SI Appendix*, Table **??**). The cumulative length (absolute difference between reference and non-reference allele) was longer for insertions (47,691,942 bases) than deletions (40,101,303 bases). This pattern was similar for biallelic variations (35,748 and 28,476 biallelic insertions and deletions, respectively, encompassing 37,388,222 and 28,373,582 bases with an average variant length of 1,045 and 996 bases). The multi-assembly graph contained more complete insertions (20,432; i.e., only non-reference sequences present in the bubbles, thus reference length is 0) than alternate insertions (15,316; i.e., both reference and non-reference sequences present but non-reference allele is longer). The pattern was similar for deletions. The multi-allelic structural variations had 13,299 alleles including 9,282 non-reference alleles with 4,621 insertions and 4,661 deletions, respectively, affecting 11,727,721 and 10,303,720 bases. Bubbles with multi-allelic structural variations contained more mixed mutations (1,941; both deletions and insertions detected within the same bubble) than multiple mutations of the same type (994 and 1,082 for multiple insertions and deletions, respectively).

When compared to the ARS-UCD1.2 backbone, the yak, Brahman, Original Braunvieh, Angus, and Highland assemblies contained respectively 49,836, 22,976, 10,965, 10,735, and 10,560 non-reference alleles (Fig. 2). Most non-reference alleles (36,443, total length: 30 Mb) were private to the yak assembly. We detected 9,267, 2,232, 2,133, and 2,037 non-reference alleles, respectively, containing 10.1, 4.9, 3.8, and 3.3 Mb that were private to the Brahman, Original Braunvieh, Highland, and Angus assembly (Fig. 2, *SI Appendix*, Fig. **??**). We also found 1,749 alleles within the 4.4 Mb of non-reference sequence (2.1 Mb of which is non-repetitive) shared by all assemblies except ARS-UCD1.2.

We mapped PacBio HiFi reads from a Nellore (*Bos taurus indicus*) x Brown Swiss (*Bos taurus taurus*) crossbred bull to the multi-assembly graph to examine support for the non-reference alleles. Nearly one third of the structural variation breakpoints had support from the hybrid cattle, while this rose to approximately three-quarters after excluding nodes with only yak labels. Since neither parental breed is present in the multi-assembly graph, this suggests that the discovered structural variation may be prevalent in different breeds of taurine and indicine cattle.

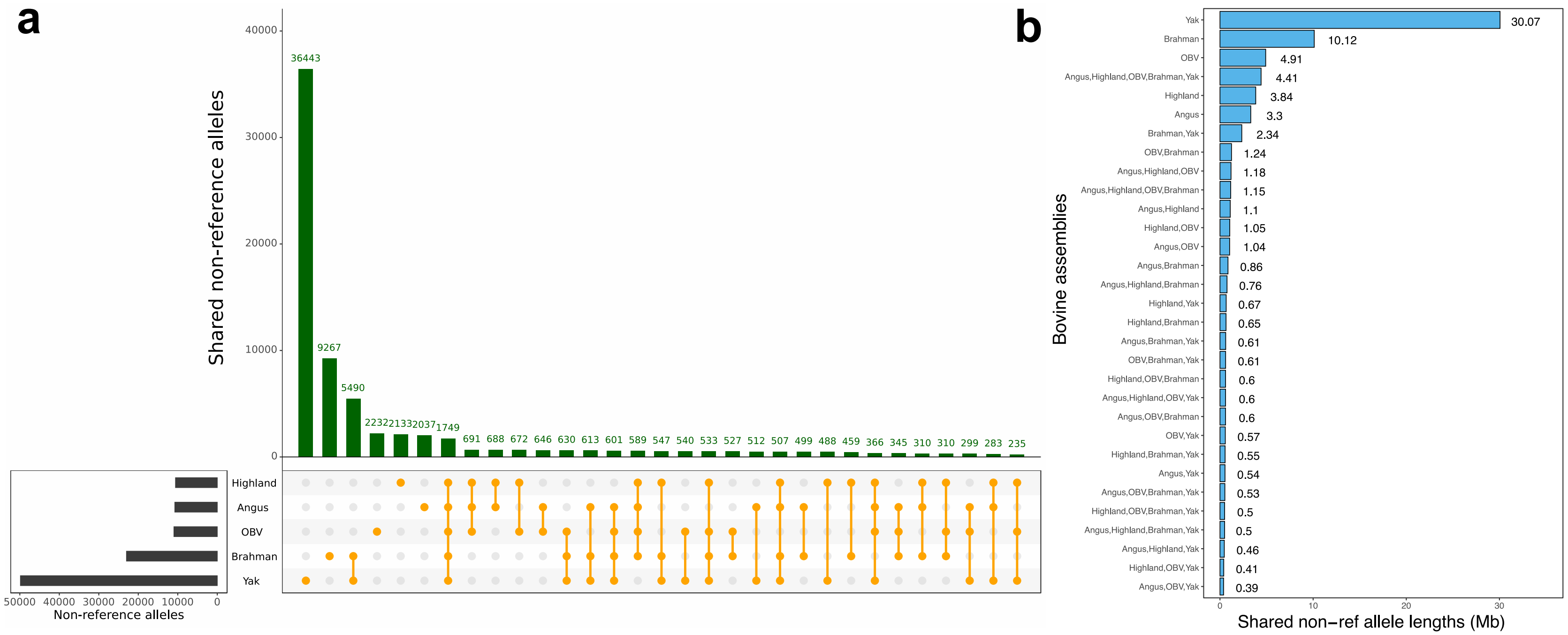

Figure 2: **Non-reference alleles detected across assemblies.**
Intersection of non-reference alleles **(a)** and cumulative length of the alleles **(b)** found in five assemblies when compared to ARS-UCD1.2. OBV = Original Braunvieh.

Sequence content of the structural variations

In order to investigate the functional relevance of the non-reference sequences, we extracted 45,357 non-reference alleles from the 70,329,827 non-reference bases in the multi-assembly graph (*SI Appendix*, Fig. **??**). These sequences originate from 38,906 biallelic and 6,451 multiallelic structural variations, respectively, that have a cumulative length of 43,003,591 and 27,326,236 bases. On average, the alleles of multiallelic structural variations were four times longer than that of biallelic bubbles (4,205 versus 1,104 bases).

The non-reference sequences are largely comprised of repetitive elements (53,690,260 bases or 76.34%, *SI Appendix*, Fig. **??**). LINE/L1 and LINE/RTE-BovB account for 28.04 (52.22%) and 6.77 (12.61%) Mb repetitive non-reference bases, respectively. Repetitive sequences (both interspersed and simple repeats) are more evenly distributed across the autosomes than non-repetitive sequences. Both repetitive and non-repetitive non-reference sequences were detected at two regions on bovine chromosomes 18 and 23 that encompass the leukocyte receptor complex and the major histocompatibility complex (*SI Appendix*, Fig. **??**).

We hypothesized that the 16,639,567 non-repetitive non-reference bases contain transcribed sequences. A BLASTX search of these sequences against a protein sequence database of Bos and related species revealed hits for 403 structural variations containing 299,337 non-reference bases. As a complementary approach, we predicted genes from the non-repetitive sequences using the Augustus software tool. The *ab initio* prediction revealed 857 gene models from 768 distinct structural variations that had a minimum coding sequence length of 150 bp, including 374 complete gene models with transcription start site, start codon, exons, stop codon, and transcription termination site (*SI Appendix*, Table **??**). On average, the transcript, coding sequence, and protein length of the complete gene models is respectively 4,742 bp, 794 bp, and 264 aa.

De novo transcript assembly from the non-reference sequences

As the two complementary gene prediction methods indicated that the novel sequences contain transcribed features, we sought experimental evidence. We appended the 70 Mb of repeat masked non-reference sequences contained in 45,357 additional contigs to the ARS-UCD1.2 reference, making an extended reference genome. This renders the non-reference sequences amenable to current methods of linear mapping of transcriptome data. Using HISAT2, we aligned liver transcriptomes from 39 cattle across taurine (Angus, Holstein, Jersey) and indicine (Brahman) breeds to both the linear reference as well as the extended reference. We also aligned transcriptomes from Dominette, the animal sequenced to assemble the *Bos taurus* reference genome. A greater portion of reads mapped to the extended reference compared to the original reference for all examined samples (*SI Appendix*, Fig. **??**). Across the 40 samples, the overall mapping rate increased by 0.037%, which corresponds to approximately 18K reads for a paired-end RNA-seq dataset of 25 million reads. The mapping improvements were larger for samples with great genetic distance from the reference genome. Brahman had the largest improvement (0.060%), followed by the taurine breeds: Angus (0.032%), Holstein (0.026%), and Jersey (0.030%). As expected, Dominette benefitted the least (0.010%), but still demonstrated an improvement over using the original reference.

Next, we used StringTie2 (Kovaka et al., 2019), guided with gene models predicted by Augustus (see above), to assemble reads which aligned to non-reference sequences into 1,431 putatively novel genes. Of these, 885 were expressed at TPM $\geq$ 1 in at least one breed, including 405 that were originally predicted by Augustus. We selected these 405 putatively novel genes, supported by both *ab initio* prediction and de novo transcript assembly for further analyses.

Only 263 of the 405 putatively novel genes were expressed at TPM $\geq$ 1 in Dominette, with BLASTP queries indicating they may be divergent copies of ribosomal proteins or olfactory receptors. The remaining 142 putatively novel genes were expressed at TPM $\geq$ 1 in Angus, Holstein, Jersey or Brahman cattle. Most were expressed in Brahman cattle (Fig. 3a), including 20 putatively novel genes specific to this indicine breed. Among the taurine breeds, Angus contributed more putatively novel genes than either Holstein or Jersey cattle. Putatively novel genes common to all four non-reference breeds accounted for nearly half, 68 of the 142, identified in any non-reference breed (Fig. 3b). The average expression was significantly higher (P=0.004, one-tailed t-test) for genes that were expressed in at least two breeds (N=106, TPM=13.48) than genes expressed in only one breed (N=36, TPM=1.64). BLASTP queries provided additional support for 57 out of 142 putatively novel genes (*SI Appendix*, Fig. **??**). The top hits suggest that the putatively novel genes encode proteins related to: immune response (antigen-presenting glycoprotein, immunoglobulin, BOLA, killer-T-cell, interferon, Ig-like lectin, CMRF35, MHC, cytokine), signalling (G-protein signalling protein, tyrosine-phosphatase), cytoskeleton regulations (myosin, actin, twinfilin, KANTB1), lipid metabolism (apolipoprotein, lipid-binding protein), and protein modifications (heat-shock chaperone, ubiquitin conjugating enzyme, rhoA ubiquitin).

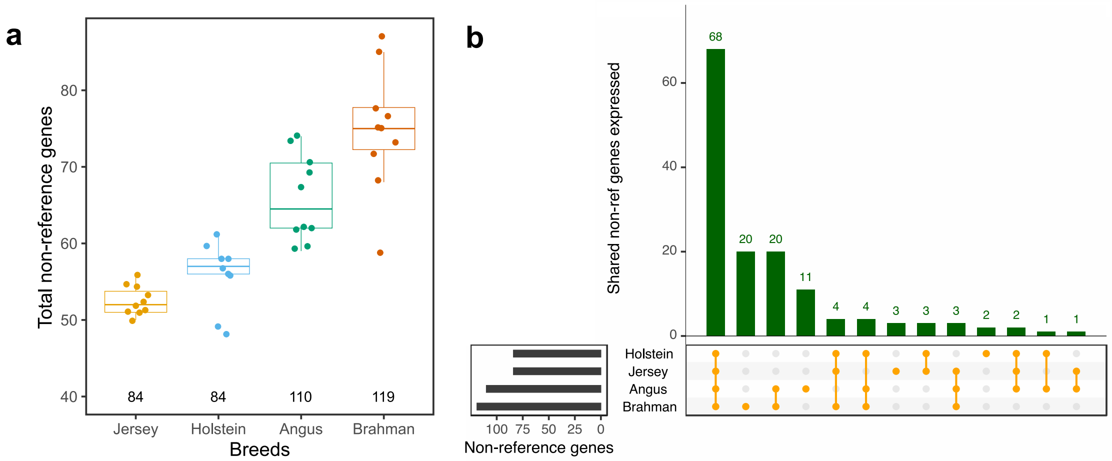

Figure 3: **Novel transcribed genes detected from non-reference sequences.** **(a)** Number of non-reference genes expressed $\geq$1 TPM in liver tissue from taurine (Jersey, Holstein, Angus) and indicine (Brahman) cattle breeds. Each point represents the number of novel genes detected per animal. The number of distinct novel genes detected for each breed is indicated below the boxplots. **(b)** Expression of 142 putatively novel non-reference genes in the four cattle breeds.

Non-reference sequences contain differentially expressed genes

To investigate if the non-repetitive sequences also encode transcripts that are differentially expressed between individual Bos taurus cattle, we obtained publicly available peripheral blood leukocyte transcriptome data for eight Mycobacterium bovis-infected and eight non-infected Holstein cattle (McLoughlin et al., 2014). Following the transcriptome analysis introduced earlier, the RNA sequencing reads were aligned to both the standard and extended ARS-UCD1.2 reference genome sequence. Between 8,616,414 and 23,940,699 RNA sequencing reads aligned to the standard and between 8,631,277 and 23,977,859 RNA sequencing reads aligned to the extended reference genome. The subsequent de novo transcript assembly from the non-reference sequences produced 949 transcripts, encoded by 661 putatively novel genes. We appended them to the Ensembl ARS-UCD1.2 annotation, yielding a total of 28,268 genes. Considering only unique alignments, we detected expression levels $\geq$ 1 CPM in at least eight samples for 13,085 genes, including 272 novel genes. We subsequently tested these genes for differential expression, finding 3,646 genes, including 36 putatively novel genes, which were differentially expressed (FDR $\leq$ 0.05) between Mycobacterium bovis-infected and non-infected cattle (Fig. 4a). The top differentially expressed genes from our extended Ensembl ARS-UCD1.2 annotation, as well as their transcript abundances in cases and controls, agreed well with the original findings from McLoughlin *et al.* that were based on the previous UMD3.1 annotation (Pearson R log2FC: 0.99) as well as with those from the standard ARS-UCD1.2 reference genome annotation (Pearson R log2FC: 0.99, *SI Appendix*, **??**).

Within the 36 putatively novel differentially expressed genes, 28 and 8 are respectively up- and downregulated in peripheral blood leukocytes of *Mycobacterium bovis*-infected cattle, with an average 2-fold change compared to non-infected controls (*SI Appendix*, Fig. **??**). Multidimensional scaling representations of transcript abundance estimates of the 36 differentially expressed genes separated *Mycobacterium bovis*-infected from non-infected cattle (Fig. 4b). BLASTX queries against a protein reference database provided additional support for 13 out of 36 differentially expressed genes (*SI Appendix*, Table **??**). The top upregulated non-reference gene supported by the BLASTX query (4.04-fold increase, P=1.98e-05) encodes the Workshop Cluster (WC) 1.1-like protein, i.e., a receptor expressed on gamma delta T cells that modulates the immune response to *Mycobacterium bovis* infections (McGill et al., 2014; Damani-Yokota et al., 2018; Kennedy et al., 2002).

The top downregulated non-reference gene supported by the BLASTX query encodes a protein with high similarity (79.80%) to leukocyte immunoglobulin-like receptor A5 (LILRA5). LILRA5 triggers the strength of the innate immune response to *Mycobacterium* infections (Bah et al., 2018) and might serve as a target for pathogen-mediated immunomodulation. Many genes of the leukocyte receptor complex are missing in the assembled chromosomes of the ARS-UCD1.2 reference (Bakshy et al., 2021); instead, *LILRA5* (LOC100139766) is annotated on a 236 kb long unplaced scaffold (NW_020190675). A non-reference gene encoding a protein similar to *LILRA5* is located within a 20.4 kb insertion of the multi-assembly graph at 62,471,732 bp on chromosome 18. Both taurine (Original Braunvieh) and indicine (Brahman) assemblies support this insertion. The putatively novel gene encoding *LILRA5* is expressed at 9.59±2.54 and 23.10±8.30 CPM, respectively, in *Mycobacterium bovis*-infected and non-infected cattle, corresponding to a 2.19−fold decrease (P=1e-04) in infected cattle (*SI Appendix*, Table **??**).

Variant discovery from the non-reference sequences

Next, we mapped short sequencing reads, with an average of 19-fold sequencing coverage, from 45 cattle representing five taurine breeds against ARS-UCD1.2 and the extended ARS-UCD1.2 reference genome. An average number of 34,342 reads per sample mapped perfectly within 50 bp of the breakpoints of the newly added contigs indicating that the addition of 100 bp flanking sequence was sufficient to facilitate accurate alignments. Across 45 samples, the average mapping rate increased by 0.0176% over ARS-UCD1.2, corresponding to approximately 100K sequencing reads for a DNA sample sequenced at 30-fold coverage. The mapping rate increased more noticeably for Brown Swiss (0.024%) and Original Braunvieh (0.021%) than Holstein (0.015%) and Simmental (0.016%) cattle (*SI Appendix*, Fig. **??**). Similarly, to the transcriptome mapping, sequence reads from

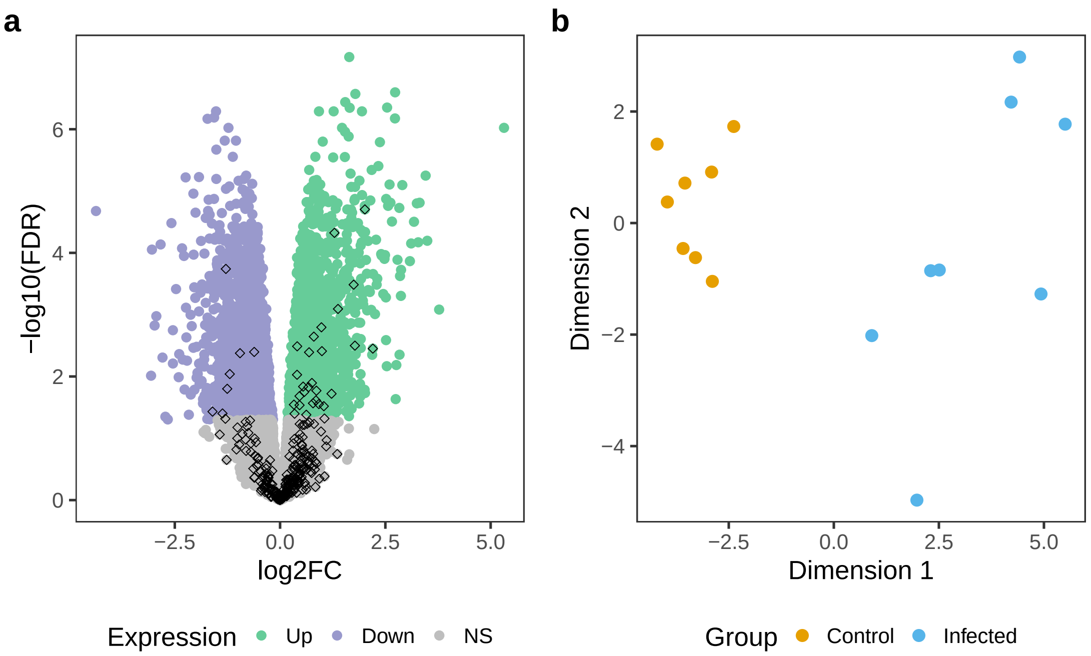

Figure 4: **Differentially expressed non-reference genes.**
(a) Volcano plot representing results from the differential expression analysis. Green and purple color indicates genes that are up- and downregulated (FDR ≤ 0.05), respectively, in peripheral blood leukocytes of *Mycobacterium bovis*-infected cattle. Diamond shapes indicate the 272 putatively novel genes found in non-reference sequences. **(b)** Multidimensional scaling plot of 36 differentially expressed non-reference genes in *Mycobacterium bovis*-infected (blue) and non-infected (orange) Holstein cattle.

Dominette benefitted the least from the extended reference genome (0.006%). However, the increase in mapping rate was greater (0.013%) for other Hereford cattle. For all breeds, the extended reference genome also enabled more perfect alignments (alignments without difference from the reference), less partially mapped (i.e., clipped) reads, and less reads with supplementary alignments. However, the proportion of reads with unique alignment was lower for the extended than standard reference genome (*SI Appendix*, Table **??**).

We next investigated the alignments against the 2,115,702 non-repetitive non-reference bases detected in all assemblies except ARS-UCD1.2. Among these, 919,761 bases were covered by confident alignments (≥10-fold) from Dominette. This suggests that, although absent from the autosomal assembly, these sequences do occur in the animal used to construct the reference. However, 1,195,941 bp were not covered with reads from Dominette, but instead from Brown Swiss, Holstein, Original Braunvieh or Simmental samples. Strikingly, reads from non-Dominette Hereford samples covered 745,392 of the 1,195,941 bases. This directly implies that Dominette has individual-specific deletions, which are either rare or absent in other Hereford cattle.

Mapping against the extended reference resulted in many reads changing alignment location to the non-reference additions. Most (85.55%) of the reads mapping at non-reference sequences already mapped to the original ARS-UCD1.2 reference genome, although 5% of these mapped to unplaced contigs, while 14.45% were previously unmapped. These mappings displayed an increase in the average mapping quality (22 to 44), alignment score (110 to 142), and alignment identity (0.975 to 0.995). The proportion of clipped reads decreased from 39% to 4%. The subset of these reads which were previously unmapped showed even greater improvements (*SI Appendix*, Fig. **??**).

Using reads with mapping quality greater than 10 for reference-guided sequence variant genotyping yielded 83,250 filtered variants (73,709 SNPs, 9,541 Indels) in non-reference sequences that were identified by both *SAMtools* and *GATK*. These variants formed 80,995 biallelic and 2,255 multi-allelic sites, with a Ti:Tv ratio of 1.91, averaging 1.18 variants per kb. 3890 small variations (Ti:Tv ratio: 1.79) were detected within 50 bp of the breakpoints of the newly added contigs. On average each Brown Swiss, Original Braunvieh, Holstein, Simmental, and Hereford animal respectively had 31,028, 29,685, 29,851, 30,309, and 15,845 variant sites in non-reference bases (Fig. 5a). A DNA sample from Dominette had considerably fewer polymorphic sites at non-reference bases, only 7,531. Most variants (32.67%) had alternate allele frequency less than 0.1, and 193 were fixed for the alternate allele (*SI Appendix*, Fig. **??**). The top principal components from a genomic relationship matrix that was built from the 83,250 non-reference variants separated the animals by breeds (Fig. 5b,c). Functional annotation based on the gene models predicted from Augustus indicated that most non-reference variants were either intergenic (83%) or intronic (7.5%). 1138 variants (Ti:Tv ratio: 1.83) were in putative coding sequences, of which 54 were classified as "HIGH IMPACT" variants (*SI Appendix*, Table **??**).

0.3 Discussion

We utilize a bovine multi-assembly graph to uncover sequences that are not included in the Bos taurus reference genome. Novel contigs can also be assembled from unmapped reads, but placing them onto reference coordinates is difficult (Sherman et al., 2019; Golicz et al., 2016). Our approach provides physical coordinates for the novel sequences because the breakpoints anchor them onto the reference genome. Despite including the genetically distant yak, constructing the multi-assembly graph using minigraph (Li et al., 2020) was computationally efficient and scalable. Our multi-assembly graph utilizes a well-annotated backbone assembly to identify non-reference sequences from other assemblies. We show that the choice of the backbone as well as its genetic distance to all other assemblies influences the amount of non-reference bases uncovered through the multi-assembly graph.

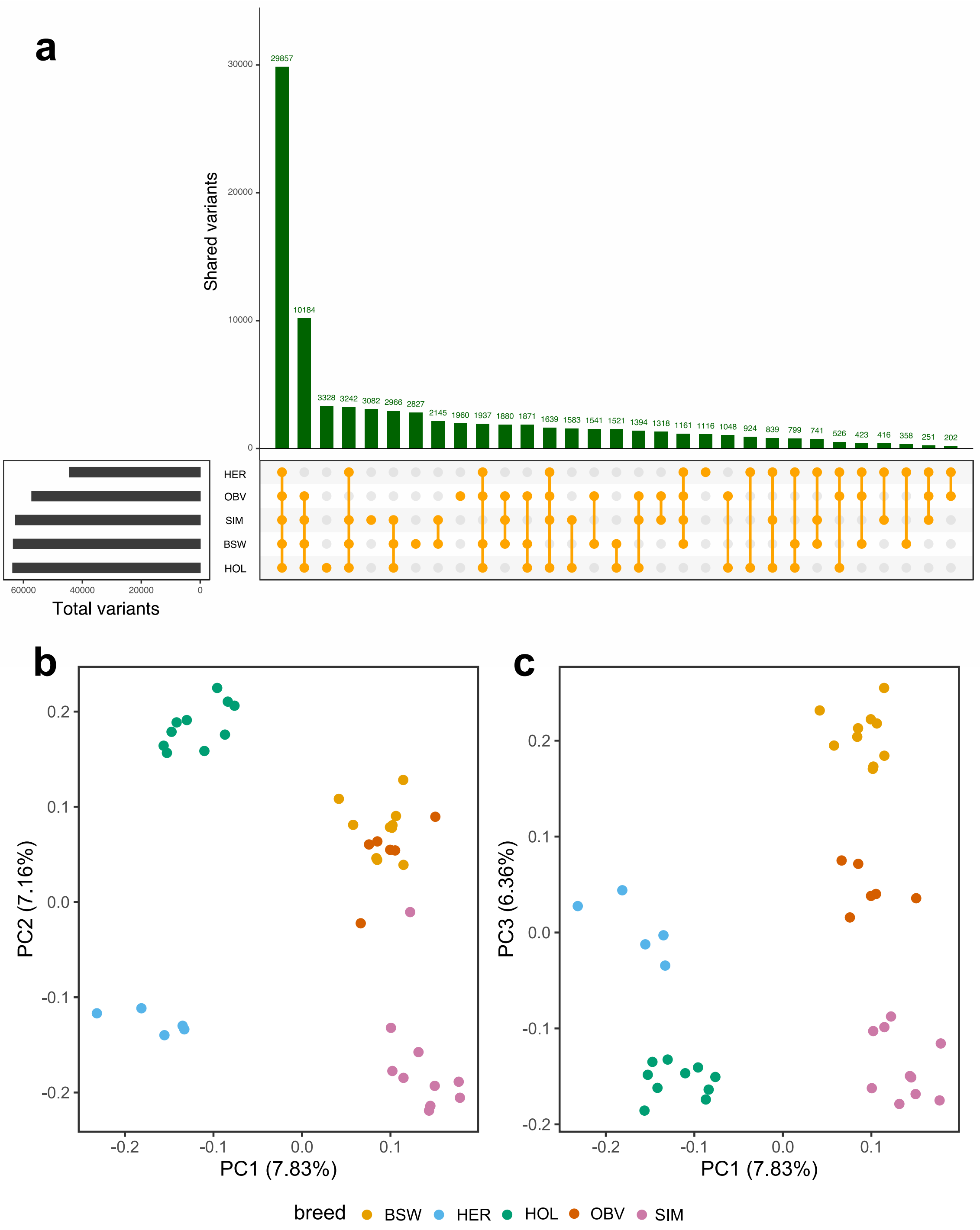

Figure 5: **Polymorphic sites detected from non-reference sequences in five breeds.**

(a) Sharing of 83,250 variants across five taurine cattle breeds (BSW: Brown Swiss, HER: Hereford, HOL: Holstein, OBV: Original Braunvieh, SIM: Simmental). **(b, c)** The top three principal components (PC) of a genomic relationship matrix constructed from non-reference sequence variants separate the animals by breeds.

Sophisticated algorithms facilitate the reference-free alignment of thousands of assemblies (Armstrong et al., 2020). To determine the origin of the non-reference sequences, we developed an approach to assign labels to all nodes in the multi-assembly graph. Our evaluation showed that this strategy is highly accurate.

By systematically characterizing structural variations in multiple assemblies from domestic cattle and their close relatives, we detect 45,357 autosomal segments with a cumulative length of 70,329,827 bases that are novel when compared to the *Bos taurus* reference genome. To obtain continuous non-reference sequences spanning multiple non-reference nodes, we recovered the non-reference alleles from structural variations. The number of bases detected in our study that are not in the *Bos taurus* reference genome is comparable to values reported for pigs (72.5 Mb (Tian et al., 2019) and goats (38.3 Mb (Li et al., 2019), based on multi-assembly graphs constructed from 11 and 8 animals representing different breeds respectively. In our study, many non-reference sequences originate from yak. Hybridizing between yak and cattle is widely practiced, and results in fertile female descendants. However, multiple generations of backcrossing are required for males to resume fertility (Qi et al., 2010). A pangenome constructed from domestic cattle and their extant relatives as recently proposed by the Bovine Pangenome Consortium (Smith et al., 2020) will reveal variants that were lost during domestication and the separation of cattle into specialized breeds (Khan et al., 2020). For instance, some of the 8 million non-reference bases specific to Brahman might contribute to the adaptation of indicine cattle to harsh environments. Individual taurine assemblies also contain between 14 and 18 million bases that are missing in the Hereford-based reference assembly, many of which are shared between individuals. This value is somewhat higher than the 5-10 million non-reference bases detected per human genome (Ameur et al., 2018; Audano et al., 2019; Duan et al., 2019), possibly because cattle breeds diverged more strongly than human populations due to intense artificial selection. Each of the three taurine assemblies contains approximately 3 million autosomal non-reference bases that were not detected in any other assembly. There were also 4.4 million non-reference bases, of which 2.1 million were non-repetitive, that were present in all assemblies except the reference. This includes 1.2 million bases that are either specifically deleted in the Hereford breed or the animal used to build the reference, inadvertently propagating reference-bias.

A reference graph may integrate linear reference coordinates, non-reference sequences, and shorter variants (Hickey et al., 2020). However, as many genome analysis tools still rely on a linear coordinate system, we append the novel non-reference sequences linearly to the ARS-UCD1.2 reference genome. Adding 100 bp flanking sequence on either side of the breakpoints facilitated accurate alignment of sequencing reads at the boundaries of the novel contigs. A graph-based approach might enable the mapping of sequencing reads spanning breakpoints (Hickey et al., 2020). We considered only variations larger than 100 bp because integrating smaller variations increases the complexity of the resulting reference with limited benefit for downstream analyses (Li et al., 2020). We show that our extended ARS-UCD1.2 reference genome leads to improved DNA and RNA sequence read mapping in indicine and taurine cattle, even for breeds that did not contribute to the multi-assembly graph. However, excessively adding novel sequences to the reference genome carries the risk of increasing the number of ambiguous alignments.

The non-reference sequences comprise more repetitive elements than the overall ARS-UCD1.2 reference genome (76% versus 48%), but less than non-reference insertions detected from human pangenomes (88%) (Sherman et al., 2019; Ameur et al., 2018). Many non-reference sequences with repetitive elements were observed at immune gene complex loci, corroborating that these regions are highly repetitive (Schwartz et al., 2017). The immune gene complex loci also contain many non-repetitive non-reference sequences suggesting great allelic diversity which may cause assembly problems (Bakshy et al., 2021), thus resulting in gaps and missing sequences in the primary ARS-UCD1.2 assembly.

We show that the 16.6 million non-repetitive non-reference bases encompass transcribed features. An *ab initio* approach predicted 857 gene models from these sequences. The *de novo* assembly of RNA sequencing read alignments from liver samples provided additional support for more than 400 of these gene models. As these analyses were only conducted on liver transcriptomes, it is highly likely that the non-reference sequences contain additional coding sequences that are transcribed in different tissues. The discovery of distinct putatively novel genes in an independent RNA sequencing dataset from peripheral blood leukocytes of Holstein cattle supports this hypothesis. Some of the putatively novel genes, including genes encoding olfactory receptors, were also present in the animal used to build the reference genome. Olfactory receptors have been observed to undergo frequent duplication and rapid evolution in mammalian genomes (Li et al., 2017; Hughes et al., 2018). Segments encompassing duplicated genes may either be collapsed in primary assemblies or result in unplaced contigs that represent variants of the sequence in the assembled chromosomes (Vollger et al., 2019; Kelley and Salzberg, 2010), hence the presence of paralogous copies among non-reference genes is expected. In order to obtain a confident set of non-reference genes, we retained only genes that were not expressed in Dominette. Many of the proteins encoded by these non-reference genes are predicted to play roles in the immune response. Pangenome analyses in species other than cattle have also revealed non-reference genes with immune-related functions (Li et al., 2017; Gordon et al., 2017; Golicz et al., 2020). Our findings show that more novel transcripts can be assembled in breeds that contribute to the multi-assembly graph (Brahman, Angus) than those not included (Holstein, Jersey), suggesting that individual assemblies contain breed-specific, functionally relevant bases. We detect the largest number of non-reference genes using RNA samples from Brahman, suggesting that breeds with great genetic distance from the reference benefit the most from a more diverse reference genome. Importantly, some putatively novel genes are differentially expressed between Mycobacterium bovis-infected and non-infected cattle, including genes that encode proteins that either contribute to the immune response against Mycobacterium infections or may serve as targets for immunomodulation by the pathogen. These differentially expressed genes remained undetected when the transcriptomes were aligned against the standard linear reference genome (McLoughlin et al., 2014). Thus, our multi-assembly graph uncovers functionally active and biologically relevant genomic features that are missing in the Bos taurus reference genome.

Our extended reference genome also leads to substantial improvements over ARS-UCD1.2 in reference-guided alignment and variant discovery. First, the sequence read mapping rate increases for samples from all breeds investigated. Using the extended reference genome would enable mapping approximately 100K previously unmapped reads for samples sequenced at 30-fold coverage. Second, the mapping quality increases for reads that were previously aligned to other positions in ARS-UCD1.2, suggesting that the novel sequences resolve misalignments. These findings agree well with results from species other than cattle, including goats, pigs, and humans (Tian et al., 2019; Li et al., 2019; Audano et al., 2019). In addition, we show that the novel sequences contain polymorphic sites that remained hitherto undetected; we discover 83,250 variants that segregate within and between breeds of cattle. A cluster analysis based on these variants separated individuals by breed, suggesting that variable non-reference bases might be associated with breed-specific traits. This hypothesis is further supported by the "HIGH IMPACT" classification of 54 variants affecting non-reference bases. Considering that the Ti/Tv ratio of the variants in putatively novel coding sequences was only 1.83, they need to be scrutinized for false positives (DePristo et al., 2011). In any case, our multi-assembly graph makes a previously neglected source of inherited variation amenable to genetic investigations.

The size of the bovine multi-assembly graph will grow as additional reference-quality assemblies from the Bovinae subfamily become available. Assemblies which are more distant will contribute correspondingly to the overall pangenome growth, increasing the flexible part of graph, and reducing the size of the core genome (*SI Appendix*, **??**). In its current implementation, our multi-assembly graph only contains insertions and deletions, as other types of structural variations (e.g., translocations, inversions) that distort the collinearity of the assembly graph cannot be integrated accurately with minigraph. We provide a versatile workflow that facilitates constructing and characterizing multi-assembly graphs for a flexible number of assemblies (`https://github.com/AnimalGenomicsETH/bovine-graphs`, *SI Appendix*, **??**). Our workflow provides tools to determine the origin of non-reference bases, derive structural variations from multi-assembly graphs, predict putatively novel genes and append the novel sequences linearly to a reference genome. We anticipate that the latter will become obsolete as soon as accurate and fast base-level alignment and split-read graph mapping enables the full-suite of genome analyses from a reference graph (Sirén et al., 2020).

0.4 Methods

Construction of the multi-assembly graph

We used minigraph (Li et al., 2020) (version 0.12-r389) with option *-xggs* to integrate six reference-quality genome assemblies into a multi-assembly graph. The current bovine reference genome (Bos taurus taurus, ARS−UCD1.2, GCF_002263795.1) and four assemblies that were generated previously are accessible at NCBI: Angus (*Bos taurus taurus*, UOA_Angus_1, GCA_003369685.2)(Low et al., 2020), Brahman (*Bos taurus indicus*, UOA_Brahman_1, GCF_003369695.1) (Low et al., 2020), Highland *(Bos taurus taurus*, ARS_UNL_Btau-highland_paternal_1.0_alt, GCA_009493655.1) (Rice et al., 2020), yak (*Bos grunniens*,
ARS_UNL_BGru_maternal_1.0_p, GCA_009493645.1) (Rice et al., 2020). Additionally, we constructed an assembly from a female Original Braunvieh calf (*Bos taurus taurus*) using PacBio high-fidelity (HiFi) reads (*SI Appendix*, **??**). The sampling of blood from the Original Braunvieh animal and its parents was approved by the veterinary office of the Canton of Zurich (animal experimentation permit ZH 200/19).

The genetic distance among the six assemblies was estimated using Mash (version 2.2) (Ondov et al., 2016). We performed genomic sketching separately for each assembly with *mash sketch* using a sketch and k-mer size of s=1000 and k=21, respectively. Sketches were combined using *mash paste*, and *mash dist* was used to estimate the distances between the assemblies. A phylogenetic tree was built from the estimated pairwise distances using the neighbor-joining method (Saitou and Nei, 1987) as implemented in the R package ape (version 5.4) (Paradis and Schliep, 2019). The tree was visualized with the *phylo.plot* function, using the yak assembly as the outgroup to root the tree.

Identification of non-reference segments from the multi-assembly graph

We refer to nodes that are not in the Hereford-based reference genome (ARS-UCD1.2) as non-reference nodes. We separately aligned (with minigraph parameters "–cov -x asm") each of the six assemblies back to the multi-assembly graph to determine the support for non-reference nodes. For each alignment, all nodes with non-zero coverage, i.e., nodes traversed by this specific assembly, were labelled. After iterating through all the alignments, each node then contained labels for every assembly which passed through it. As such, each node necessarily had at least one label, while a node traversed by all six assemblies would have six labels (*SI Appendix*, Fig. **??**).

It was possible to assess minigraph's alignment accuracy for the path of the Hereford-based reference genome (ARS-UCD1.2), because all reference nodes in the multi-assembly graph were from this assembly. Nodes were considered true positive (TP) and true negative (TN) when reference and non-reference nodes were correctly assigned Hereford labels, respectively. Reference nodes aligned as non-reference nodes were assigned false negative (FN) and non-reference nodes aligned as reference nodes were assigned false positive (FP). We characterized alignment recall (TP / (TP+FN)), precision (TP / (TP+FP)), and overall F1 score (2 * (precision * recall) / (precision + recall)).

Identification of structural variations from the multi-assembly graph

We used the bubble popping algorithm of gfatools (version 0.4) (Li et al., 2020) to derive the structural variations from the multi-assembly graph. In the reference graph model of minigraph, a bubble is a branching region in the graph for which the start and end node are reference sequences. A path traversing the start and end nodes represents an allele of a structural variant.

The version of gfatools considered in our study reports the shortest and longest path for each bubble. To detect and classify all paths within a bubble, we applied the following stepwise procedure (*SI Appendix*, Fig. **??**):

- Determine the start and stop node for each bubble using the bubble popping algorithm of gfatools.
- Traverse all possible paths in the bubble using a recursive depth-first search.
- Retain only paths with color-consistent labels (see above).

- Classify a path as a reference path when all nodes and edges are part of the Hereford-based reference assembly, and as non-reference otherwise.
- Compare reference and non-reference paths to classify the type of the structural variations.

Structural variations were classified as biallelic if two paths were observed in a bubble and multi-allelic if a bubble contained more than two paths. The structural variations were further classified into:

- Alternate deletion, when the non-reference path was shorter than the reference path (but the reference path has nonzero length).
- Complete deletion, when the non-reference path has a length of zero.
- Alternate insertion, when the non-reference path was longer than the reference path.
- Complete insertion, when the reference path has a length of zero.

Breakpoints of structural variations were determined according to ARS-UCD1.2 reference coordinates. We overlapped the breakpoints with annotations from Ensembl (build 101) to identify structural variations in coding sequences. Affected genes were subjected to a gene set enrichment analysis using PANTHER (http://pantherdb.org/) (Mi et al., 2019) for which the Bos taurus reference gene list was supplied as a baseline.

To validate the structural variations, we mapped 6,803,270 (46-fold coverage) PacBio HiFi reads to the multi-assembly graph using GraphAligner (version 1.0.12) (Rautiainen and Marschall, 2020) with preset -x vg (variation graph mapping). The HiFi reads were generated from a Nellore x Brown Swiss crossbred bull (SAMEA7765441), representing taurine and indicine breeds that were not used to build the multi-assembly graph. The veterinary office of the Canton of Zurich approved the sampling of blood from the crossbred animal and its parents (animal experimentation permit ZH 200/19). The mean read length was 20,612 bases with an average accuracy of 99.76%. We calculated coverage (number of reads aligned) at each node and edge in the graph based on the GAF (Graphical Alignment Format) output from GraphAligner.

We combined all non-reference alleles (excluding complete deletions, paths without non-reference bases, and paths with length less than 100 bp) to obtain a comprehensive set of non-reference bases from the multi-assembly graph. To facilitate the mapping of short reads to the segment edges, we added 100 bp of flanking sequences (derived from sequences at the source and sink nodes) on either side of the structural variations. The flanking sequences were not considered for length calculations or gene predictions (see below).

To investigate the repeat content of the non-reference sequences, we used the RMBlastn search engine (version 2.10.0) to run RepeatMasker version 4.1.1 (option -species cow) (Smit et al., 2015) using the database of repetitive DNA elements from Repbase (release 20181026) (Bao et al., 2015).

Bioinformatic characterization of non-reference sequences

In order to reveal functionally active non-reference sequences, we performed two complementary analyses:

First, we compared the repeat masked non-reference sequences against a local protein database using DIAMOND BLASTX (version 0.9.30) (Buchfink et al., 2015). Using DIAMOND makedb, the local protein database was built from the RefSeq protein sequences of

- Taurine cattle (*Bos taurus taurus*, GCF_002263795.1_ARS-UCD1.2_protein.faa)
- Indicine cattle (*Bos taurus indicus*, GCF_003369695.1_UOA_Brahman_1_protein.faa)
- Yak (*Bos mutus*, GCF_000298355.1_BosGru_v2.0_protein.faa)
- Human (*Homo sapiens*, GCF_000001405.39_GRCh38.p13_protein.faa)
- Mouse (*Mus musculus*, GCF_000001635.26_GRCm38.p6_protein.faa)
- Bison (*Bison bison,* GCF_000754665.1_Bison_UMD1.0_protein.faa)
- Water buffalo (*Bubalus bubalis*, GCF_003121395.1_ASM312139v1_protein.faa)
- Goat (*Capra hircus*, GCF_001704415.1_ARS1_protein.faa)
- Sheep (*Ovis aries*, GCF_002742125.1_Oar_rambouillet_v1.0_protein.faa)
- the curated protein databases of SwissProt and PDB (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/)

To query the non-reference sequences against the local protein database we ran BLASTX with the parameters "–more-sensitive –e-value 1e-10 –outfmt 6". We considered only the top hit for each queried sequence with minimum coverage and identity of 80%.

Second, we performed an ab initio gene structure prediction from the repeat masked non-reference sequences using a local instance of Augustus (version 3.3.3) (Stanke and Waack, 2003) using default parameters trained on the human genome. From the Augustus GTF output file, we extracted the number of gene models, the number of gene models with transcription start and termination site, transcript length, exon count, and length per gene, coding sequence count and length per gene, and protein length of the putatively novel protein-coding sequences. To classify the domain and family of the putatively novel proteins, we converted the Augustus GTF output to the fasta format and performed a query against the local protein database (as above) using DIAMOND BLASTP with the same parameters and thresholds as the BLASTX query.

De novo transcript assembly from non-reference sequences

We downloaded between 12,361,440 and 34,421,106 paired-end RNA-sequencing reads from liver tissue from 10 Angus (Xiang et al., 2018), 10 Brahman (Nguyen et al., 2016), 9 Holstein and 10 Jersey (Salleh et al., 2018) cattle, as well as from Dominette - the animal used to construct the ARS-UCD1.2 reference genome (Rosen et al., 2020). Adapter sequences and low-quality bases were removed from the raw RNA sequencing data using default parameters of fastp (version 0.19.4) (60). The filtered reads were then aligned using HISAT2 (version 2.1.0) (Kim et al., 2019), with option "–dta" to facilitate the downstream transcriptome assembly, to the original ARS-UCD1.2 reference as well as the extended version of the ARS-UCD1.2 reference. The extended reference was constructed by appending repeat masked non-reference sequences as unplaced contigs.

Putative novel transcripts were assembled de novo using StringTie2 (version 2.1.1) (Kovaka et al., 2019) from RNA-seq reads that aligned to the non-reference sequences. To facilitate transcript assembly, we supplied the ARS-UCD1.2 Ensembl annotation (build 101) and the novel gene models predicted by Augustus (see above). Transcripts were assembled *de novo* separately for all RNA sequencing samples. Subsequently, we used StringTie2 *merge* to create a unique set of transcripts across all samples and facilitate the assembly of full-length transcripts from partially assembled transcripts. We quantified gene expression for each sample with StringTie2 using a fixed (merged) GTF file that was generated previously (without predicting novel transcripts, option -e). Gene abundance was quantified in transcript per million (TPM).

Differential gene expression analysis

We utilized publicly available peripheral blood leukocyte transcriptomes of eight *Mycobacterium bovis*-infected and eight age-matched healthy Holstein cattle (McLoughlin et al., 2014) to detect differentially expressed genes from non-reference sequences. The RNA-sequencing data contain between 9,272,629 and 25,358,979 single-end reads of length 78 bp. We performed quality control on the raw sequencing reads using fastp (version 0.19.4) (Chen et al., 2018) with default parameters. The filtered reads were then mapped to the extended ARS-UCD1.2 reference genome that contained the non-reference sequences using HISAT2 (Kim et al., 2019). Potential non-reference transcripts were assembled de novo with StringTie2 (see above). Gene-level read counts were estimated based on a custom annotation file that contained the Ensembl (build 101) ARS-UCD1.2 genome annotation and the non-reference annotation as generated by StringTie2 using the *featurecounts* function of the Rsubread package (option countMultiMappingReads =FALSE to exclude multi-mapping reads). The read count matrix was used as input for EdgeR version 3.24.3 (Robinson et al., 2010). We normalized transcript abundance by sequencing depth using the trimmed-mean of M-values (TMM) approach. Genes that were expressed at $\geq$1 count per million (CPM) in at least eight samples were tested for differential expression in peripheral blood leukocytes between *Mycobacterium bovis*-infected and control animals using a generalized linear model (GLMQfit) with dispersion parameter estimated using the Cox-Reid method. Genes were considered to be differentially expressed at a Benjamini-Hochberg-corrected FDR$\leq$0.05. Multidimensional scaling of the normalized read count matrix of the differentially expressed genes was performed using the *cmdscale* function in R.

Mapping and variant calling from whole-genome short read data

We considered the original ARS-UCD1.2 reference genome and an extended version of the reference that additionally contained 70,329,827 non-reference bases detected from five assemblies. We used paired-end short read sequencing data from 45 samples representing five breeds: Original Braunvieh, Brown Swiss, Holstein, Simmental (Häfliger et al., 2020), and Hereford (including Dominette, the animal used to construct the ARS-UCD1.2 reference genome) (Rosen et al., 2020; Young et al., 2020) that had average sequencing coverage of 18.94-fold. Quality control of the short-read sequencing reads was performed using fastp (version 0.19.4) (Chen et al., 2018) with default parameter settings. The filtered reads were subsequently mapped to the original ARS-UCD1.2 reference and the extended ARS-UCD1.2 reference that also contained non-reference sequences using the mem-algorithm of BWA (version 0.7.17) (Li, 2013) with default parameters. Duplicate reads were marked with Samblaster (version 0.1.24) (Faust and Hall, 2014).

We performed multi-sample variant calling (SNP and Indels) on the non-reference sequences using SAMtools (version 1.10) (Li et al., 2009) and GATK (version v4.1.9.0) (**?**) as detailed in Crysnanto *et al.* (Crysnanto et al., 2019). Base quality scores were recalibrated using known variants from the 1000 bull genomes project database (`http://www.1000bullgenomes.com/doco/ARS1.2PlusY_BQSR_v3.vcf.gz`). We applied the GATK modules *HaplotypeCaller*, *GenomicsDBImport* and *GenotypeGVCFs* to discover and genotype polymorphic sites. The variants were subsequently hard-filtered using recommended parameters (SNP filters: $QD < 2||QUAL < 30||FS > 60||MQ < 40||MQRankSum < -12.5||ReadPosRankSum < -8||AN < 10$, Indel filters: $QD < 2||QUAL < 30||FS > 200||ReadPosRankSum < -20.0||AN < 10$) (Crysnanto et al., 2019). A second independent variant discovery and genotyping approach was performed using SAMtools mpileup and bcftools call (Li et al., 2009). The resulting genotypes were subsequently hard-filtered according to parameters recommend by the 1000 bulls genomes project ($QUAL < 20||MQ < 30||DP < 10||AN < 10$) (Daetwyler et al., 2014). To create a consistent variant representation across both datasets, variants were normalized using vt version 0.5 (Tan et al., 2015). We retained only filtered variants, which were identified by both SAMtools and GATK. Functional consequences of variants affecting non-reference bases were predicted based on the GTF-file from Augustus (see above) using Ensembl's Variant Effect Predictor (McLaren et al., 2016).

Data availability

Short sequencing reads are available at the European Nucleotide Archive (ENA) (`http://www.ebi.ac.uk/ena`) with study accession PRJNA436715 (Transcriptome - Brahman), PRJNA392196 (Transcriptome - Angus), PRJNA357463 (Transcriptome – Holstein, Jersey), PRJNA294306 (Transcriptome - Dominette), PRJNA257841 (Differential expression analysis – Holstein), PRJEB18113 (WGS – BSW, OBV, HOL, SIM), PRJNA494431 (WGS - Hereford), PRJNA391427 (WGS - Dominette). Accession numbers for all samples are provided in Dataset S1. PacBio HiFi reads for an Original Braunvieh animal used to construct a de novo assembly are available at study accession PRJEB42335 under sample accession SAMEA7759028. PacBio HiFi reads for a Nelore x Brown Swiss bull are available at study accession PRJEB42335 under sample accession SAMEA7765441. Data supporting this study, including the multi-assembly graph, non-reference sequences, putatively novel genes, transcript abundances and sequence variants detected from non-reference sequences are available via Zenodo (`https://doi.org/10.5281/zenodo.4385983`) (Crysnanto et al., 2021).

Code availability

Workflows to construct multi-assembly graphs and custom scripts to characterize non-reference sequences are available via Github
(https://github.com/AnimalGenomicsETH/bovine-graphs).
All workflows were built using Snakemake (version 5.30.1)(72) and custom scripts were written in R (version 3.5.1) (R Core Team, 2017) and Python (version 3.7.1).

Acknowledgements

We are thankful for the excellent technical support provided by the ETH Zürich functional genomics platform FGCZ (https://fgcz.ch/). Computing was done at the Leonhard High Performance Compute cluster at ETH Zürich. This study was supported by grants from the Swiss National Science Foundation (310030_185229) and the Swiss Federal Office for Agriculture (FOAG), Bern.

References

A. Ameur, H. Che, M. Martin, I. Bunikis, J. Dahlberg, I. Höijer, S. Häggqvist, F. Vezzi, J. Nordlund, P. Olason, et al. De novo assembly of two swedish genomes reveals missing segments from the human grch38 reference and improves variant calling of population-scale sequencing data. *Genes*, 9(10):486, 2018.

J. Armstrong, G. Hickey, M. Diekhans, I. T. Fiddes, A. M. Novak, A. Deran, Q. Fang, D. Xie, S. Feng, J. Stiller, et al. Progressive cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, 587(7833):246–251, 2020.

P. A. Audano, A. Sulovari, T. A. Graves-Lindsay, S. Cantsilieris, M. Sorensen, A. E. Welch, M. L. Dougherty, B. J. Nelson, A. Shah, S. K. Dutcher, et al. Characterizing the major structural variant alleles of the human genome. *Cell*, 176(3):663–675, 2019.

S. Y. Bah, T. Forster, P. Dickinson, B. Kampmann, and P. Ghazal. Meta-analysis identification of highly robust and differential immune-metabolic signatures of systemic host response to acute and latent tuberculosis in children and adults. *Frontiers in genetics*, 9:457, 2018.

K. Bakshy, D. Heimeier, J. Schwartz, E. Glass, S. Wilkinson, R. A. Skuce, A. Allen, J. Young, J. McClure, J. Cole, et al. Development of polymorphic markers in the immune gene complex loci of cattle. *Journal of Dairy Science*, 2021.

S. Ballouz, A. Dobin, and J. A. Gillis. Is it time to change the reference genome? *Genome biology*, 20(1): 1–9, 2019.

W. Bao, K. K. Kojima, and O. Kohany. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile Dna*, 6(1):1–6, 2015.

B. Buchfink, C. Xie, and D. H. Huson. Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59–60, 2015.

S. Chen, Y. Zhou, Y. Chen, and J. Gu. fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics*, 34(17):i884–i890, 2018.

D. Crysnanto and H. Pausch. Bovine breed-specific augmented reference graphs facilitate accurate sequence read mapping and unbiased variant discovery. *Genome biology*, 21(1):1–27, 2020.

D. Crysnanto, C. Wurmser, and H. Pausch. Accurate sequence variant genotyping in cattle using variation-aware genome graphs. *Genetics Selection Evolution*, 51(1):1–15, 2019.

D. Crysnanto, A. S. Leonard, Z. H. Fang, and H. Pausch. Supporting data for Novel functional sequences uncovered through a bovine multi-assembly graph (version 1.0) [Dataset], 2021. URL https://doi.org/10.5281/zenodo.4385983.

H. D. Daetwyler, A. Capitan, H. Pausch, P. Stothard, R. Van Binsbergen, R. F. Brøndum, X. Liao, A. Djari, S. C. Rodriguez, C. Grohs, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature genetics*, 46(8):858–865, 2014.

P. Damani-Yokota, J. C. Telfer, and C. L. Baldwin. Variegated transcription of the wc1 hybrid prr/co-receptor genes by individual $\gamma\delta$ t cells and correlation with pathogen responsiveness. *Frontiers in immunology*, 9:717, 2018.

M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. Del Angel, M. A. Rivas, M. Hanna, et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43(5):491, 2011.

Z. Duan, Y. Qiao, J. Lu, H. Lu, W. Zhang, F. Yan, C. Sun, Z. Hu, Z. Zhang, G. Li, et al. Hupan: a pan-genome analysis pipeline for human genomes. *Genome biology*, 20(1):1–11, 2019.

H. P. Eggertsson, H. Jonsson, S. Kristmundsdottir, E. Hjartarson, B. Kehr, G. Masson, F. Zink, K. E. Hjorleifsson, A. Jonasdottir, A. Jonasdottir, et al. Graphtyper enables population-scale genotyping using pangenome graphs. *Nature genetics*, 49(11):1654, 2017.

G. G. Faust and I. M. Hall. Samblaster: fast duplicate marking and structural variant read extraction. *Bioinformatics*, 30(17):2503–2505, 2014.

E. Garrison, J. Sirén, A. M. Novak, G. Hickey, J. M. Eizenga, E. T. Dawson, W. Jones, S. Garg, C. Markello, M. F. Lin, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature biotechnology*, 36(9):875–879, 2018.

A. A. Golicz, P. E. Bayer, G. C. Barker, P. P. Edger, H. Kim, P. A. Martinez, C. K. K. Chan, A. Severn-Ellis, W. R. McCombie, I. A. Parkin, et al. The pangenome of an agronomically important crop plant brassica oleracea. *Nature communications*, 7(1):1–8, 2016.

A. A. Golicz, P. E. Bayer, P. L. Bhalla, J. Batley, and D. Edwards. Pangenomics comes of age: from bacteria to plant and animal applications. *Trends in Genetics*, 36(2):132–145, 2020.

S. P. Gordon, B. Contreras-Moreira, D. P. Woods, D. L. Des Marais, D. Burgess, S. Shu, C. Stritt, A. C. Roulin, W. Schackwitz, L. Tyler, et al. Extensive gene content variation in the brachypodium distachyon pan-genome correlates with population structure. *Nature communications*, 8(1):1–13, 2017.

I. M. Häfliger, M. Sickinger, M. Holsteg, L. M. Raeder, M. Henrich, S. Marquardt, C. Drögemüller, and G. Lühken. An il17ra frameshift variant in a holstein cattle family with psoriasis-like skin alterations and immunodeficiency. *BMC genetics*, 21:1–10, 2020.

G. Hickey, D. Heller, J. Monlong, J. A. Sibbesen, J. Sirén, J. Eizenga, E. T. Dawson, E. Garrison, A. M. Novak, and B. Paten. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome biology*, 21(1):1–17, 2020.

G. M. Hughes, E. S. Boston, J. A. Finarelli, W. J. Murphy, D. G. Higgins, and E. C. Teeling. The birth and death of olfactory receptor gene families in mammalian niche adaptation. *Molecular biology and evolution*, 35(6):1390–1406, 2018.

B. Kehr, A. Helgadottir, P. Melsted, H. Jonsson, H. Helgason, A. Jonasdottir, A. Jonasdottir, A. Sigurdsson, A. Gylfason, G. H. Halldorsson, et al. Diversity in non-repetitive human sequences not found in the reference genome. *Nature Genetics*, 49(4):588–593, 2017.

D. R. Kelley and S. L. Salzberg. Detection and correction of false segmental duplications caused by genome mis-assembly. *Genome biology*, 11(3):1–11, 2010.

H. E. Kennedy, M. D. Welsh, D. G. Bryson, J. P. Cassidy, F. I. Forster, C. J. Howard, R. A. Collins, and J. M. Pollock. Modulation of immune responses to mycobacterium bovis in cattle depleted of wc1+ $\gamma\delta$ t cells. *Infection and immunity*, 70(3):1488–1500, 2002.

A. W. Khan, V. Garg, M. Roorkiwal, A. A. Golicz, D. Edwards, and R. K. Varshney. Super-pangenome by integrating the wild side of a species for accelerated crop improvement. *Trends in plant science*, 25 (2):148–158, 2020.

D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nature biotechnology*, 37(8):907–915, 2019.

K. Kim, T. Kwon, T. Dessie, D. Yoo, O. A. Mwai, J. Jang, S. Sung, S. Lee, B. Salim, J. Jung, et al. The mosaic genome of indigenous african cattle as a unique genetic resource for african pastoralism. *Nature Genetics*, 52(10):1099–1110, 2020.

S. Koren, A. Rhie, B. P. Walenz, A. T. Dilthey, D. M. Bickhart, S. B. Kingan, S. Hiendleder, J. L. Williams, T. P. Smith, and A. M. Phillippy. De novo assembly of haplotype-resolved genomes with trio binning. *Nature biotechnology*, 36(12):1174–1182, 2018.

L. Koufariotis, B. Hayes, M. Kelly, B. Burns, R. Lyons, P. Stothard, A. Chamberlain, and S. Moore. Sequencing the mosaic genome of brahman cattle identifies historic and recent introgression including polled. *Scientific reports*, 8(1):1–12, 2018.

S. Kovaka, A. V. Zimin, G. M. Pertea, R. Razaghi, S. L. Salzberg, and M. Pertea. Transcriptome assembly from long-read rna-seq alignments with stringtie2. *Genome biology*, 20(1):1–13, 2019.

H. Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*, 2013.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

H. Li, X. Feng, and C. Chu. The design and construction of reference pangenome graphs with minigraph. *Genome biology*, 21(1):1–19, 2020.

M. Li, L. Chen, S. Tian, Y. Lin, Q. Tang, X. Zhou, D. Li, C. K. Yeung, T. Che, L. Jin, et al. Comprehensive variation discovery and recovery of missing sequence in the pig genome using multiple de novo assemblies. *Genome research*, 27(5):865–874, 2017.

R. Li, W. Fu, R. Su, X. Tian, D. Du, Y. Zhao, Z. Zheng, Q. Chen, S. Gao, Y. Cai, et al. Towards the complete goat pan-genome by recovering missing genomic segments from the reference genome. *Frontiers in genetics*, 10:1169, 2019.

W. Y. Low, R. Tearle, R. Liu, S. Koren, A. Rhie, D. M. Bickhart, B. D. Rosen, Z. N. Kronenberg, S. B. Kingan, E. Tseng, et al. Haplotype-resolved genomes provide insights into structural variation and gene content in angus and brahman cattle. *Nature communications*, 11(1):1–14, 2020.

J. L. McGill, R. E. Sacco, C. L. Baldwin, J. C. Telfer, M. V. Palmer, and W. R. Waters. Specific recognition of mycobacterial protein and peptide antigens by $\gamma\delta$ t cell subsets following infection with virulent mycobacterium bovis. *The Journal of Immunology*, 192(6):2756–2769, 2014.

W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. The ensembl variant effect predictor. *Genome biology*, 17(1):1–14, 2016.

K. E. McLoughlin, N. C. Nalpas, K. Rue-Albrecht, J. A. Browne, D. A. Magee, K. E. Killick, S. D. Park, K. Hokamp, K. G. Meade, C. O'Farrelly, et al. Rna-seq transcriptional profiling of peripheral blood leukocytes from cattle infected with mycobacterium bovis. *Frontiers in immunology*, 5:396, 2014.

H. Mi, A. Muruganujan, D. Ebert, X. Huang, and P. D. Thomas. Panther version 14: more genomes, a new panther go-slim and improvements in enrichment analysis tools. *Nucleic acids research*, 47(D1): D419–D426, 2019.

L. Nguyen, A. Reverter-Gomez, A. Canovas, B. Venus, A. Islas-Trejo, S. Lehnert, J. Medrano, S. Moore, and M. Fortes. P1012 liver transcriptome from pre versus post-pubertal brahman heifers. *Journal of Animal Science*, 94(suppl_4):20–21, 2016.

B. D. Ondov, T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren, and A. M. Phillippy. Mash: fast genome and metagenome distance estimation using minhash. *Genome biology*, 17(1):1–14, 2016.

E. Paradis and K. Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in r. *Bioinformatics*, 35(3):526–528, 2019.

J. Pritt, N.-C. Chen, and B. Langmead. Forge: prioritizing variants for graph genomes. *Genome biology*, 19(1):1–16, 2018.

X. Qi, H. Jianlin, G. Wang, J. Rege, and O. Hanotte. Assessment of cattle genetic introgression into domestic yak populations using mitochondrial and microsatellite dna markers. *Animal genetics*, 41(3): 242–252, 2010.

R Core Team. R: A Language and Environment for Statistical Computing. 2017. URL https://www.r-project.org.

M. Rautiainen and T. Marschall. Graphaligner: rapid and versatile sequence-to-graph alignment. *Genome biology*, 21(1):1–28, 2020.

E. S. Rice, S. Koren, A. Rhie, M. P. Heaton, T. S. Kalbfleisch, T. Hardy, P. H. Hackett, D. M. Bickhart, B. D. Rosen, B. V. Ley, et al. Continuous chromosome-scale haplotypes assembled from a single interspecies f1 hybrid of yak and cattle. *Gigascience*, 9(4):giaa029, 2020.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

B. D. Rosen, D. M. Bickhart, R. D. Schnabel, S. Koren, C. G. Elsik, E. Tseng, T. N. Rowan, W. Y. Low, A. Zimin, C. Couldrey, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience*, 9(3):giaa021, 2020.

N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.

S. Salleh, G. Mazzoni, P. Løvendahl, and H. N. Kadarmideen. Gene co-expression networks from rna sequencing of dairy cattle identifies genes and pathways affecting feed efficiency. *BMC bioinformatics*, 19(1):1–15, 2018.

J. C. Schwartz, M. S. Gibson, D. Heimeier, S. Koren, A. M. Phillippy, D. M. Bickhart, T. P. Smith, J. F. Medrano, and J. A. Hammond. The evolution of the natural killer complex; a comparison between mammals using new high-quality genome assemblies and targeted annotation. *Immunogenetics*, 69(4): 255–269, 2017.

B. G. Sequencing and A. Consortium. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science (New York, NY)*, 324(5926):522, 2009.

R. M. Sherman, J. Forman, V. Antonescu, D. Puiu, M. Daya, N. Rafaels, M. P. Boorgula, S. Chavan, C. Vergara, V. E. Ortega, et al. Assembly of a pan-genome from deep sequencing of 910 humans of african descent. *Nature genetics*, 51(1):30–35, 2019.

J. Sirén, J. Monlong, X. Chang, A. M. Novak, J. M. Eizenga, C. Markello, J. Sibbesen, G. Hickey, P.-C. Chang, A. Carroll, et al. Genotyping common, large structural variations in 5,202 genomes using pangenomes, the giraffe mapper, and the vg toolkit. *Biorxiv*, 2020.

A. Smit, R. Hubley, and P. Green. RepeatMasker Open-4.0, 2015. URL http://www.repeatmasker.org.

T. Smith, D. Bickhart, and B. Rosen. Genome assemblies of global cattle breeds to create a cattle pangenome. In *Plant and Animal Genome XXVIII Conference (January 11-15, 2020)*. PAG, 2020.

J.-M. Song, Z. Guan, J. Hu, C. Guo, Z. Yang, S. Wang, D. Liu, B. Wang, S. Lu, R. Zhou, et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of brassica napus. *Nature Plants*, 6(1):34–45, 2020.

M. Stanke and S. Waack. Gene prediction with a hidden markov model and a new intron submodel. *Bioinformatics*, 19(suppl_2):ii215–ii225, 2003.

A. Tan, G. R. Abecasis, and H. M. Kang. Unified representation of genetic variants. *Bioinformatics*, 31 (13):2202–2204, 2015.

X. Tian, R. Li, W. Fu, Y. Li, X. Wang, M. Li, D. Du, Q. Tang, Y. Cai, Y. Long, et al. Building a sequence map of the pig pan-genome from multiple de novo assemblies and hi-c data. *Science China Life Sciences*, pages 1–14, 2019.

M. R. Vollger, P. C. Dishuck, M. Sorensen, A. E. Welch, V. Dang, M. L. Dougherty, T. A. Graves-Lindsay, R. K. Wilson, M. J. Chaisson, and E. E. Eichler. Long-read sequence and assembly of segmental duplications. *Nature methods*, 16(1):88–94, 2019.

L. K. Whitacre, P. C. Tizioto, J. Kim, T. S. Sonstegard, S. G. Schroeder, L. J. Alexander, J. F. Medrano, R. D. Schnabel, J. F. Taylor, and J. E. Decker. What's in your next-generation sequence data? an exploration of unmapped dna and rna sequence reads from the bovine reference individual. *BMC genomics*, 16(1):1–7, 2015.

K. H. Wong, W. Ma, C.-Y. Wei, E.-C. Yeh, W.-J. Lin, E. H. Wang, J.-P. Su, F.-J. Hsieh, H.-J. Kao, H.-H. Chen, et al. Towards a reference genome that captures global genetic diversity. *Nature communications*, 11(1):1–11, 2020.

R. Xiang, B. J. Hayes, C. J. Vander Jagt, I. M. MacLeod, M. Khansefid, P. J. Bowman, Z. Yuan, C. P. Prowse-Wilkins, C. M. Reich, B. A. Mason, et al. Genome variants associated with rna splicing variations in bovine are extensively shared between tissues. *BMC genomics*, 19(1):1–18, 2018.

A. E. Young, T. A. Mansour, B. R. McNabb, J. R. Owen, J. F. Trott, C. T. Brown, and A. L. Van Eenennaam. Genomic and phenotypic analyses of six offspring of a genome-edited hornless bull. *Nature biotechnology*, 38(2):225–232, 2020.