DISS. ETH NO. XXX

Establishing Bovine Pangenome Graphs

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

Danang Crysnanto

M.Sc., Quantitative Genetics and Genome Analysis
The University of Edinburgh

born on 08.01.1992
citizen of Indonesia

accepted on the recommendation of

Prof. Dr. Hubert Pausch	examiner
Prof. Dr. Bernt Guldbrandtsen	co-examiner
Prof. Dr. David MacHugh	co-examiner

2021

Table of Contents

List of Figures

List of Tables

Summary

The assembly of the draft *Bos taurus* reference genome was a milestone for genetics- and genomics-oriented research in cattle. The reference genome of domestic cattle was built from a single animal from the Hereford breed. However, the linear reference sequence does not represent the genetic diversity of global cattle breeds. The lack of diversity causes problems, particularly when DNA sequences from genetically distant animals are aligned and compared to the reference sequence. This issue is widely known as reference bias. Pangenomes are an intriguing novel reference structure to consider the full-spectrum of genetic diversity within a species. A rich, graph-based pangenome reference can integrate multiple genome assemblies and their sites of variations in a coherent and non-redundant data structure. This thesis investigated for the first time the utility of graph-based references for genomic analysis in a livestock population.

Chapter 2 assessed the feasibility of graph-based genomic analysis in cattle. Specifically, a graph-based sequence variant genotyping approach was implemented using the *Graphtyper* software and compared to two widely-used methods (*SAMtools* and *GATK*) that rely on a strictly linear representation of the reference using whole-genome sequencing data of 49 Original Braunvieh cattle. A comparison between sequence variant and array-derived genotypes indicated that the graph-based approach outperformed both *SAMtools* and *GATK* with regard to genotype concordance, non-reference sensitivity, non-reference discrepancy, and Mendelian consistency of genotypes observed in parent-offspring pairs. These findings demonstrated that graph-based genotyping using *Graphtyper* is accurate, sensitive, and computationally feasible in the cattle genome.

Chapter 3 reports on the construction of breed-specific and multi-breed genome graphs for four European cattle breeds (Original Braunvieh, Brown Swiss, Fleckvieh, and Holstein). The *vg toolkit* was used to augment the linear Hereford-based reference sequence with variants that were prioritized based on allele frequency in different breeds. Based on both real and simulated short-read sequencing data, this study showed that

variant prioritization is crucial to build informative genome graphs. Intriguingly, adding many low frequency and rare variants to the genome graphs compromised mapping accuracy. Moreover, this chapter demonstrated that multi-breed graphs and breed-specific graphs enable almost identical mapping improvements over a linear reference genome. Finally, the first whole-genome graph was constructed for the Brown Swiss cattle breed using 14 million variants. The application of this whole-genome graph facilitated accurate short-read mapping and unbiased sequence variant discovery.

Chapter 4 reports on integrating six reference-quality bovine genome assemblies into a unified multi-assembly graph using the *minigraph* software. The pangenome contains 70 megabases that are not present in the current ARS-UCD1.2 *Bos taurus* reference genome. Using complementary bioinformatics approaches, this chapter provides compelling evidence that these non-reference sequences contain functionally active and biologically-relevant elements. Specifically, the analysis of transcriptome data revealed putatively novel genes, including some that are differentially expressed between individual animals. Moreover, variant discovery in the non-reference sequences revealed thousands of yet undetected polymorphic sites capturing genetic differentiation across cattle breeds. This chapter demonstrated that multi-assembly graphs make so far neglected genetic variations amenable to genetic investigations.

Overall, thesis presents a novel analysis paradigm in livestock genomics by leveraging variation-aware reference structures. The analyses presented in this thesis provide a first step towards the transition from linear to graph-based reference structures in order to mitigate inherent biases of the linear reference genome. Importantly, this thesis establishes a computational framework to integrate multiple genome assemblies and their sites of variations into a more diverse reference structure broadly applicable across species.

Zusammenfassung

Das Assembly der Bos taurus Referenzsequenz war ein Meilenstein für genetische und genomische Forschungsfragen beim Rind. Die Referenzsequenz wurde von einem einzigen Tier der Rasse Hereford erzeugt. Allerdings kann die genetische Diversität der globalen Rinderpopulation nicht in einem einzigen linearen Referenzgenom repräsentiert werden. Das ist besonders dann problematisch, wenn Sequenzen von genetisch weit entfernten Tieren mit dem Referenzgenom verglichen werden. Pangenome sind interessante neuartige Referenzstrukturen, die das gesamte Spektrum der genetischen Diversität einer Spezies abbilden. Solche graph-basierte Referenzstrukturen können mehrere Assemblies sowie deren variable Positionen integrieren. Im Rahmen dieser Dissertation werden erstmals graph-basierte Referenzstrukturen für genetische Analysen in einer Nutztierpopulation verwendet.

Im zweiten Kapitel werden erstmals graph-basierte genomische Analysen beim Rind durchgeführt. Die Genomsequenzen von 49 Original Braunvieh Rindern werden mit einem graph-basierten Ansatz nach polymorphen Positionen durchsucht. Mit der *Graph typer* software werden diese Positionen genotypisiert. Die so erhaltenen Genotypen werden mit Genotypen verglichen, die mit zwei weit verbreiteten Methoden (*SAMtools* und *GATK*) bestimmt wurden, welche strikt auf eine lineare Referenzsequenz angewiesen sind. Im Vergleich mit SNP-Chip basierten Genotypen zeigt sich, dass der graph-basierte Ansatz in *Graphtyper* sowohl *SAMtools* wie auch *GATK* im Hinblick auf die Übereinstimmung, die Sensitivität, die Spezifität und die Genauigkeit der Genotypen überlegen war. Daraus lässt sich schlussfolgern, dass die graph-basierte Genotypisierung von Rindergenomen mit *Graphtyper* genau, sensitiv und rechnerisch machbar ist.

Im dritten Kapitel werden rassespezifische und rassenübergreifende graph-basierte Referenzen für vier Europäische Rinderrassen (Original Braunvieh, Brown Swiss, Fleckvieh und Holstein) aufgestellt und verglichen. Das *vg toolkit* wurde verwendet, um die lineare Referenzsequenz mit Varianten zu erweitern, die hinsichtlich ihrer Allelfrequenz

ausgewählt wurden. Sowohl mit realen wie auch simulierten Sequenzdaten konnte gezeigt werden, dass eine Priorisierung der Varianten für informative graph-basierte Referenzgenome ausschlaggebend ist. So beeinträchtigten viele seltene Varianten den Abgleich der ausgelesenen DNA-Abschnitte mit der Referenz. Zusätzlich zeigt dieses Kapitel, dass rassenübergreifende und rassespezifische Referenzgraphen hinsichtlich des Abgleichs der DNA-Abschnitte eine fast identische Verbesserung gegenüber der linearen Referenzsequenz aufweisen. Schlussendlich konnte der erste genomweite Referenzgraph für die Rasse Brown Swiss mit rund 14 Millionen Sequenzvarianten konstruiert werden. Dieses Kapitel zeigt dass Referenzgraphen das Zuordnen von ausgelesenen DNA-Abschnitten verbessern und somit eine unverzerrte Genotypisierung von Sequenzvarianten ermöglichen.

Im vierten Kapitel werden sechs Rindergenome mit dem Programm *minigraph* zu einen Multi-Referenz-Graphen vereinigt. Dieses Pangenom beinhaltet 70 Megabasen, welche im aktuellen *Bos taurus* Referenzgenom (ARS-UCD1.2) nicht vorhanden sind. Durch die Anwendung von komplementären bioinformatischen Ansätzen liefert dieses Kapitel überzeugende Hinweise, dass diese in der Referenz nicht vorhandenen Sequenzen funktionelle und biologisch-relevante Elemente enthalten. Ausserdem enthalten sie tausende bislang unbekannte Sequenzvarianten, die sich zwischen Rinderrassen unterscheiden. Dieses Kapitel zeigte, dass Multi-Referenzen-Graphen bis anhin nicht berücksichtigte DNA Variation für genetische Untersuchungen zugänglich machen können.

Diese Dissertation präsentiert ein neues Paradigma zur Analyse genomischer Daten mit nicht-linearen Referenzstrukturen. Die verschiedenen Analysen, welche in dieser Arbeit präsentiert werden, sind ein erster Schritt um von linearen zu graphbasierten Referenzgenomen zu wechseln. In dieser Dissertation wurden grundlegende und breit anwendbare Strukturen geschaffen, die es erlauben, mehrere Referenzsequenzen und deren variable Positionen in eine nicht-lineare Datenstruktur zu integrieren.

Thesis Outline

The thesis is structured as follows:

Chapter 1 provides a literature review to introduce the concepts of a reference genome, pangenome, graph-based pangenome, and applications of the pangenome.

Chapter 2 reports on genome-graph based variant discovery and genotyping in a livestock population. This chapter is published in *Genetics Selection Evolution*.

Chapter 3 reports on the construction of the first whole-genome graphs in cattle and their application to read mapping and variant discovery. This chapter is published in *Genome Biology*.

Chapter 4 reports on the construction of a bovine multi-assembly graph from six reference-quality assemblies and its application to investigate sequences not included in the current *Bos taurus* reference genome. This chapter is published in *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*.

Chapter 5 provides a general discussion, and outlook for future research

Chapter 1

General Introduction

1.1 Genomic technologies to assess genetic variations in livestock

Cattle is an important livestock species capable of converting low-quality and human-inedible proteins into high-quality proteins. The global cattle population is highly diverse due to intense selection for specific breeding goals, such as production of milk, beef, or both (dual-purpose), as well as the adaptation to a wide range of environments [1]. Due to selective breeding and improved husbandry conditions, spectacular increases in livestock productivity have been achieved. For example, the average annual milk yield per cow in the United States of America (USA) has increased by more than five-fold from 1,890 kg in 1924 to 9,682 kg in 2011 [2].

Genomic selection had been proposed to further accelerate genetic gain [3]. To this end, the genetic value of an individual is predicted based on genome-wide molecular marker information. Genotyping arrays were developed to assess variation at thousands of polymorphic sites in the cattle genome. The genotype information is then linked to phenotype either to determine markers associated with agriculturally-important traits [4] or to derive the prediction equation for genomic selection [3]. More than 3 million cattle in the USA have already been genotyped [5]. However, variations interrogated by chip-based genotyping are not comprehensive enough to pinpoint causal mutations underlying the traits [6].

This limitation prompted the widespread utilization of whole-genome short-read sequencing. In this approach, the DNA is first fragmented and subsequently read in segments of few hundred bases (Fig. 1.1). Variation discovery typically follows a reference-guided alignment approach. Genotypes are called at positions where the observed nucleotides from the alignments differ from the corresponding reference nucleotides. Sophisticated variant calling algorithms were developed to differentiate between real variants and sequencing errors from noisy short-read data or misalignments [7]. Whole genome sequencing approaches can accurately discover small variants (SNPs and Indels < 50 bp) across the whole genome.

Sequencing costs have dropped substantially over the past decades, faster than Moore's Law (a term in computer hardware that doubling power every two years indicates a well-progressed technology), which has paved the way towards sequencing a gigabase-sized genome for only $100 [8, 9]. The decline in sequencing costs has also enabled the sequencing of individual cattle genomes for agricultural applications. The 1000 Bull Genomes Project was launched to coordinate global sequencing efforts and compile a

reference panel for sequence variant genotype imputation [10]. In their latest (8^{th}) run, the consortium has already catalogued more than 150 million variants from more than 4000 cattle across 200 breeds [11]. This variant database has become a powerful resource to impute sequence variant genotypes into large mapping cohorts, thus accelerating the discovery of causal mutations for complex and monogenic traits and improving the prediction accuracy of genomic selection [10]. Recently, low-pass sequencing (<1x) coupled with genotype imputation were proposed as a cost-effective strategy to enable population-scale whole genome sequencing variant analysis [12].

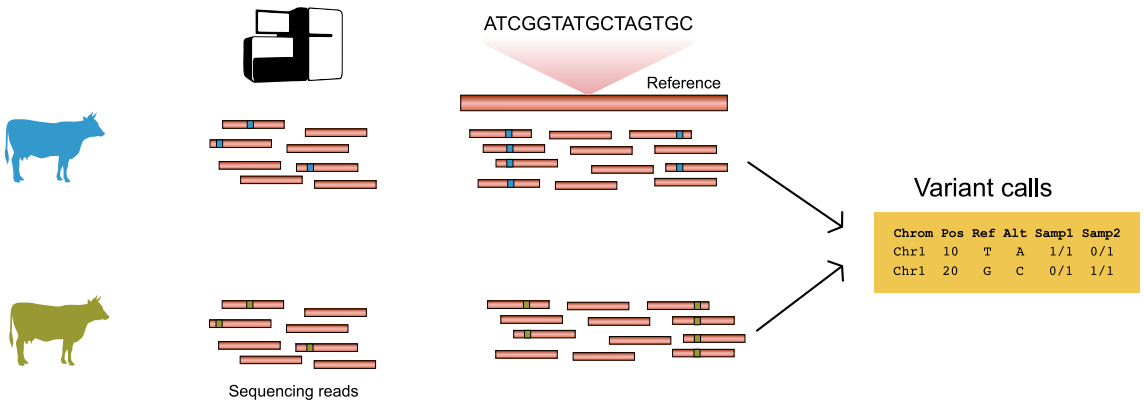


Figure 1.1: **Identification of genetic variants through genome sequencing**
The DNA of individual animals is fragmented into billions of short fragments which are then read by a DNA sequencer in a massively parallel manner. Subsequently, the sequencing reads are compared (aligned) to the reference genome. Genetic variants are identified as nucleotide discordances relative to the reference sequences.

1.2 Improvements in the cattle reference genome

A well-annotated reference genome is the starting point for many genomic analyses. It serves as a reference point for read alignment, variant calling, gene annotation, and functional analysis. Gene loci are defined at specific genomic coordinates, and alleles are referred to as alternative or reference nucleotides. The ability to compare billions of sequencing reads from hundreds to thousands of individuals to reference sequences has quickly become the gold standard, identifying variants underpinning inherited diseases or other relevant traits, thus accelerating genetic progress [13].

The first cattle reference genome (Btau 3.1 and Btau 4.0) was assembled in 2009 from bacterial artificial chromosome (BAC) and whole-genome shotgun (WGS) sequenc-

ing [14]. The contig and scaffold N50 (i.e., 50% of the genome is in fragments of this size or greater) for this assembly were 48.7 kb and 1.9 Mb respectively. This assembly was further refined in 2014 to close gaps and correct structural errors (UMD_3.1.1) using additional sequencing data and improved assembly approaches [15]. The most recent cattle reference genome (ARS-UCD 1.2) was assembled using single-molecule real-time (SMRT) long-read sequencing data and scaffolded with optical mapping data. The quality of the resulting assembly improved considerably over UMD3.1 with contig and scaffold N50 values of 25.89 Mb and 103 Mb, respectively [16]. Advances in assembly techniques (e.g. trio binning) and the development of highly accurate long-read sequencing technology facilitate constructing assemblies of high continuity, correctness and completeness [13]. The recently generated bovine assemblies exceed in quality the current *Bos taurus* reference genome with contig N50 larger than 70 Mb and could resolve complex genomic regions, e.g. major histocompatibility regions [17]. Trio binning takes advantage of the high heterozygosity observed in hybrids to separate long reads according to parental origins. The assembly is subsequently performed separately from the partitioned reads resulting in two haplotype-resolved assemblies. This approach was first applied to a cross between *Bos taurus* x *Bos indicus* cattle (Angus x Brahman) [18] and now has been applied to a broad range of cattle breeds, including undomesticated and/or cattle relatives (yak, gaur, bison) [19]. Recently, the Bovine Pangenome Consortium [20] was initiated to coordinate genome assembly efforts and characterize the complete diversity from hundreds of global cattle breeds, including their wild-relatives and under-represented breeds.

1.3 One reference genome is not enough

A single linear genome cannot fully represent species diversity

Despite recent spectacular quality improvements, linear reference genomes are unable to represent the full genomic diversity within a species. A linear reference genome typically represents a mosaic haplotype assembled from either one or a few individuals. For example, the current cattle reference genome (ARS-UCD1.2) was assembled from a DNA sample from a single highly-inbred animal from the Hereford breed named Dominette, which was initially selected to simplify the assembly process [16]. Reference assemblies from other livestock species were generated using a similar approach, e.g., an animal from the Duroc breed was used for the Sscrofa11.1 pig reference [21], an animal from San Clemente breed was used for the domestic goat reference [22], and an animal from Boxer breed was used for the CanFam 3.1 dog reference [23]. While the selec-

tion of reference animals seems to be a trivial process, the resulting reference sequences do not necessarily reflect the most common alleles that segregate in the population or from samples with breed-defining phenotypes [24]. Reference-guided variant discovery might reflect some properties of the reference animal rather than the population; e.g. variant calling will discover more variants when the reference contains rare alleles. Low et al. [25] found a striking difference in the number of detected polymorphic sites when calling Angus variants from an Angus reference than from a Brahman reference. Additionally, the reference genome might carry the lower frequency variants or variants private to the reference animals. Shukla et al. [26] and Ballouz et al. [24] estimated that 2 million bases in the human reference genome are minor alleles.

Insufficient representation of genetic diversity by linear genomes cause reference bias

Because alignment algorithms compare the reads towards the reference and try to minimize differences, the reference-guided variant discovery is biased towards the reference bases. In other words, it is easier to align DNA fragments without differences to the reference bases than DNA fragments that contain non-reference bases. Comparison of the sequencing reads with variants, even if they are the true representation of that species, will be penalized, resulting in sub-optimal alignments, misalignments, or unmapped reads (Fig. 1.2) [27]. Together, this limitation is referred to as **soft reference bias**, which hampers genomic analyses that depend on the ratio between reference and alternate alleles such as heterozygous variant calling [28], allelic-specific expression [29], or analysis in the highly polymorphic regions [30]. Wu et al. [31] observed that reference bias caused a lower estimate of divergence among *Bos* species due to mapping of cattle-relatives data to the *Bos taurus* reference genome, which tends to overlook the diverged regions.

Another limitation is referred to as **hard reference bias**, whereby a single reference is a poor representation of large structural variations that diverged between individuals in the population (Fig. 1.2) [32]. Reads originating from these highly diverged segments will remain unmapped and all subsequent genomic analyses will be blind to variations in these "missing" regions. In cattle, the comparison between two taurine assemblies revealed 10.9 Mb of Angus-specific sequences that were not present in the Hereford-based reference assembly [25]. This number increases to 21.8 Mb when the Angus assembly is compared to an indicine cattle genome. Reference genomes lacking millions of bases have been observed in many species. Ameur et al. [33] and Audano et al. [34] estimated that each human genome carries about 10 Mb non-reference bases. Long read data analysis across global ancestries discovered 8.5 Mb insertions observed in the majority of the

human population [34]. Remarkably, an analysis of the unmapped reads of the African pangenome revealed 300 Mb non-reference insertions, suggesting that the existing human reference genome might lack diversity spanning 10% of the genome [35].

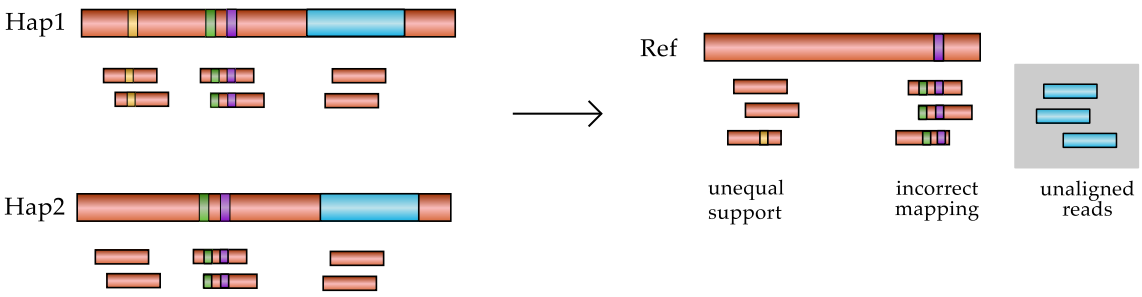


Figure 1.2: **Illustration of the reference allele bias.**
The origin of short sequencing reads of the sample (hap1 and hap2) are determined by alignments to the reference nucleotides. Thus, the comparison will always be biased towards nucleotides in the reference. Alignment of reads with alleles differing from reference nucleotides might receive lower support than alleles matching to the reference nucleotides (yellow stripe), results in incorrect alignments with multiple variations (green and purple stripes), or remains unmapped if the regions are not present in the reference (e.g. large insertion, light blue box). Orange background denotes reference sequences.

The problem of reference bias is pronounced in a species with high genetic diversity

The effect of reference bias will be more pronounced in a highly diverged species like in cattle. The genetic architecture of the bovine genomes has been shaped by various processes related to domestication, introgression, local adaptation, and human-mediated selection [1], resulting in the creation of more than 600 subpopulations (known as breeds) adapted for a variety of environmental conditions and selected for various breeding goals. Moreover, genetic diversity in cattle is higher than in human population [36]. The bovine species formed the bovine tribe which subdivided into three sub-tribes diverged about 10-15 million years ago: the *Pseudorygina*, *Bubalina* (buffalo), and *Bovina* (genus *Bison* and *Bos*). Specifically, the subtribe bovina is comprised of three subtribes split about 3-5 million years ago: (i) yak, bison; (ii) gaur, gayal, and banteng; and (iii) taurine and indicine cattle [37]. Generally, taurine breeds (*Bos taurus taurus*) are intensively selected for production traits (milk and beef) and have higher fertility than indicine breeds. Indicine breeds (*Bos taurus indicus*) generally have lower production traits and fertility, but still possess desirable traits related to heat tolerance, parasite and disease resistance [25]. However, these characteristics are not strict as there are numerous local cattle breeds optimized for specialized breeding goals [38, 39]. Series of introgressions and hybridizations created specialized breeds with mosaic genomes, such as Brahman, composed of 10 % taurine and 90% indicine origin [40]. African cattle are generally

admixed between *Bos taurus* x *Bos indicus*, where the introgressed regions are selected for African pastoralism [41]. On average, each individual cattle carries more than 5 million variants that differ from the *Bos taurus* reference, which is higher than reported for human genome [10, 42]. The number of variants is higher in more diverged, indicine [40] or under-studied African cattle [41, 43]. Additionally, this amount likely underestimates the actual genetic diversity as it does not consider structural variations, which are poorly characterized with short-read sequencing technology [44, 45].

1.4 Strategies to mitigate reference bias

Modification of the existing linear reference genome

Some strategies have been proposed to mitigate the reference bias. The most straightforward solution is to create a so-called consensus reference genome, whereby each minor allele in the reference sequence is replaced by the most frequent allele in the population. Since the transformed reference is still in the linear space, the downstream genetic analyses can still use the tools currently developed for linear genomes. However, a coordinate lift-over is needed when indels are included in the substitutions. Ballouz et al. [24] built a consensus human reference by replacing 2 million minor alleles with the corresponding major allele, that reduced mapping error by a factor of three and improved the quantification of transcripts [46]. Chen et al. [47] extended this idea into a so called reference flow approach, which re-aligns sub-optimally mapped reads into a set of genomes from multiple population, that could reduce strongly heterozygous sites by 22%. Another strategy is to continuously expands reference sequences with alternative contigs representing alleles at polymorphic regions that are impossible to assemble with a single haplotype. For instance, there were 13 updates that add 109 Mb total length to the current human reference sequences. However, this strategy is not sustainable with more diversity included. Additionally, the lack of tools that can properly handle these additional overlapping contigs will likely not be able to mitigate the reference bias and will suffer from mapping ambiguity [48].

Creation of population-specific genome assemblies

The reduced cost of long-read sequencing and improved assembly techniques make it easier to generate high-quality, near error-free, and near-complete genome assemblies [49, 50]. Thus, more studies have now shifted from species-level references to population-specific reference genomes, effectively creating personalized genomes. Large genomic initiatives such as the Vertebrate Genome Project (VGP, https://vertebratege

nomesproject.org/) [51], Darwin Tree of Life (https://www.darwintreeoflife.org/), or Earth Bio-genome Project [52] contribute to the explosion of the number of genome assemblies across the tree of life deposited in the public domain. The first phase of VGP generated 268 vertebrate genomes using long-read data, that were further scaffolded with optical mapping to produce chromosome-scale assemblies, fulfilling their strict high-quality criteria [51]. On the other hand, some genomic initiatives focus on deeply characterizing the diversity of a single species, such as the Human Pangenome Reference Consortium (HPRC) that plans to generate 350 human assemblies representing global ancestries (see https://humanpangenome.org/) [53]. A similar internationally coordinated effort was recently initiated for cattle with the Bovine Pangenome Consortium [20], which aims at generating reference-quality assemblies across global cattle breeds. There are already dozens of genomes from livestock species available in public repositories. As of April 2021, there are chromosome-level assemblies of 22 cattle (*Bos*) and its relatives (gaur, gayal, yak, bison), 19 pigs (*Sus*), 7 sheeps (*Ovis*), 4 goats (*Capra*), 9 dogs (*Canis*), with many more continuing to be added.

1.5 Transition from genomics to pangenomics

Definition of the pangenome

A pangenome refers to a structure used to integrate multiple genomes, reflecting the complete species diversity rather than collapsing all variations into a single haplotype, (see recent reviews [48, 54, 55]). The term pan-genome (pan – whole, Greek) was first introduced by Tettelin et al. [56] to describe the complete gene repertoire across *Streptococcus agalactiae* strains where 20% of the genes are variable across isolates. This concept was quickly adopted for many species across the tree of life, including pig [57, 58], goat [59], and human [35, 60]. There has been rapid growth in the number of pangenome publications across years [54], with close to 8000 studies indexed by PubMed, although most currently focus on bacterial pangenomes.

Categorization of the pangenome

The content of a pangenome can be divided into the core and flexible genome (also known as dispensable or accessory genome, Fig. 1.3a). The core genome contains sequences common to all individuals that are responsible for maintaining essential functions (e.g., DNA replication, cellular homeostasis and cellular processes). This part of the genome is under purifying selection, thus having less diversity. The dispensable genome contains segments that vary across individuals. They are under less evolutionary con-

straint, which allows for contributions to numerous adaptive phenotypes, mainly disease, biotic, and abiotic resistance, survival, immunity, defence response, adaptation to new environments, communications, and signalling [61]. Thus, dispensable genomes are of particular interest to study adaptive traits that might underpin genetic differentiation and give populations their distinguishing characteristics. In mammals, the pangenome is largely dominated by the core component (e.g. 96.67% of genes in the human) [60]. A recent report in the Mediterranean mussel *Mytilus galloprovincialis*, with high-stress tolerance and lineage-specific duplications, indicates that up to 25% of the total genome is variable [62]. Pangenomes have been extensively characterized in plants, for instance in rice [63], tomato [64], and wheat [65]. Plant pangenomes seem to have a larger proportion of accessory genomes (>20%), particularly in polypoids, outcrossing species, or in species with history of whole-genome duplications [66]. Higher ratio of flexible to core genome is typically found in species with higher adaptability [67].

It is important to differentiate between closed or open pangenomes. In a closed type pangenome, the sequencing of sufficient samples will capture the whole pangenome, and thus the size of the complete pangenome can be computationally predicted. On the other hand, sequencing more individuals will recover more pangenome content in an open pangenome. Thus, the size of pangenome keeps increasing as more samples are added [61]. Many plant and animal pangenomes are a closed type in terms of the number of genes but an open type in terms of total sequence content [60, 61], which also suggests that the non-coding segments primarily drive the sequence variability across individuals. Bacterial pangenomes are generally open type due to the prevalence of horizontal gene transfer [68]. Sampling bias of underrepresented diversity (such as genetically related samples) could lead to the falsely concluding that the pangenome is complete [67]. With additional, sufficiently diverged samples, the pangenome would continue to grow. Thus, the sampling strategy in a pangenome study should maximize diversity to fully retrieve the complete pangenome.

Approaches to build a pangenome

There are two commonly used approaches to build a pangenome (Fig. 1.3bc): “assemble-then-map" and “map-then-assemble" (also known as map-to-pan) [61]. In the “assemble-then-map"-strategy, each genome is assembled and annotated independently, which is then followed by pairwise alignment of all assembled genomes to determine shared and non-shared segments [59, 60, 69]. This assembly-based strategy is supposed to recover the full-length non-reference sequences and resolve repetitive and complex structural variants. However, this approach depends on the assembly contiguity and completeness. Assembly and annotation errors make the comparison difficult and may lead to erro-

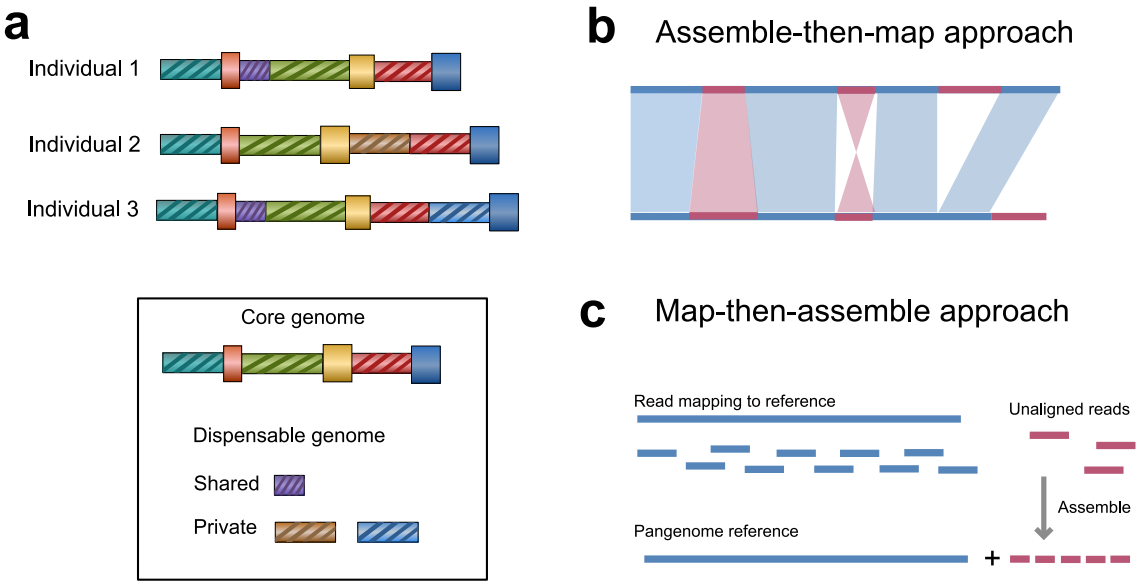


Figure 1.3: **The concept of a pangenome.**
(a) A pangenome is a collection of individual genomes, which is further divided into core (shared by all individuals) and flexible parts (the presence varies across individuals). Different strategies to build the pangenome: **(b)** Assemble-then-map: Genomes from multiple individuals are assembled, which are then compared to the reference assembly; **(c)** Map-then-assemble: sequencing reads from multiple individuals are aligned to the reference. Unmapped sequences are subsequently assembled and added as additional contigs to the reference sequences. Figures are adapted from [48] and [54].

neous identification of structural variations. Additionally, high-quality genome assemblies are still too expensive to be created at the population-scale, limiting analysis only to a subset of individuals. To take advantages the massive amount of short-read sequencing data, the majority of recent pangenome studies utilize the “map-then-assemble"-approach [35, 70, 71]. Sequencing reads from each sample are independently mapped to the reference genome. The unmapped (or poorly mapped) reads are subsequently assembled to obtain non-reference contigs. However, due to the nature of short-read-based assembly, most of the resulting contigs are fragmented, making it difficult to locate the breakpoints (origin) in the reference genome [35].

1.6 Graph-based pangenomics

Graphs as rich reference structures that integrate genetic diversity

The pangenome approaches that are based on either unmapped reads or an assembly comparison, as discussed above, rely on collections of linear genomes and do not attempt to provide coherent and comprehensive representation of genomic variation

across individuals. Considering the prevalence of genetic variations across individuals in the population and abundance of genomic resources, the linear representation is clearly an oversimplification. Emerging pangenome methods are developed to build richer variation-aware reference structures that unify the complete genetic diversity of a species in a non-redundant way. These efforts led to a new genomic discipline known as *Computational Pangenomics*, see review [72, 73, 74].

Graph-based models (also known as genome graphs or sequence graphs) are data structures to unify a collection of related sequences in a compact way (Fig. 1.4). In a sequence graph, nodes are commonly labelled with sequences and directed edges connect nodes with continuous sequences. Regions without differences are collapsed into a single node allowing compression of redundant sequences. Regions where the samples differ from each other form bubbles, where alternate paths represent different alleles [75]. Traversing (walk through) the graph recovers the initial input sequences as well as all possible recombinations.

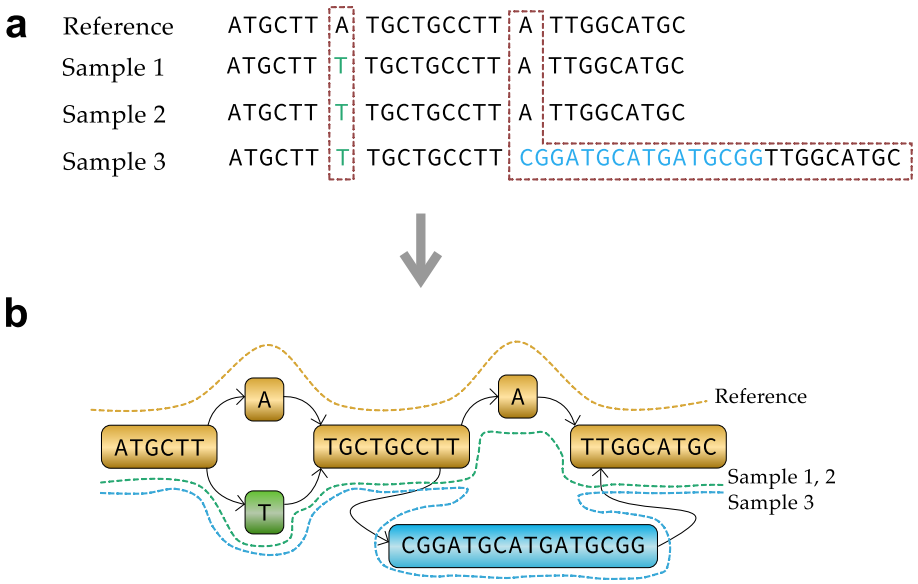


Figure 1.4: **Graph-based pangenome approach.**
(a) Most pangenomes follow the classical pangenome approach, where multiple linear genomes are compared without compressing redundant information and might lack orthologous relationships. **(b)** A graph-based pangenome approach unifies multiple genomes into a compact and rich reference representation. Nodes contain DNA sequences and nodes with continuous sequences are connected with directed edges. Redundant information across genomes is compacted by collapsing invariant regions into a single node. Alternative nodes in the bubbles (green and blue nodes) are alleles in the population. Thus, graphs allow sequence comparison to occur in the context of variations. Walks through the graph might retrieve the original sequences from which the graph was built (dashed lines).

Implementations of graph genomes

The first pangenome graph implementation was based on the DBG (*De Bruijn Graphs*). Sequencing reads from all samples were fragmented into *k*-mers (where $k <$ read length). The graph was constructed by inducing the first and second node where $k-1$ bp end of first node that overlap with the $k-1$ bp start of the second node. Nodes are "coloured" where each colour represents samples. Iqbal et al. [76] developed *Cortex*, a coloured DBG-based pangenome tool. They used it to construct a population graph from 164 human samples and identified 3.2 Mb novel sequences that are absent in the human reference genome. Because the genomic coordinates are discarded by fragmenting the reads, DBG-based approaches are not suitable for genetic variant discovery that relies on reference coordinates, although a recent study attempts to embed long-range path information into the graph [77].

Current well-established graph genome implementations establish a variation graph as an extension of the linear reference genome [28, 78, 79, 80, 81]. This implementation utilizes the existing linear reference genome as a backbone, which is then augmented with known variants. To build the graph, reference sequences are split at variable sites, and variants are added as alternative nodes of the reference bases in the graph (Fig. 1.5). The linear reference coordinates are embedded in the graphs as a path, and the nodes are referred to relative to this reference path. Thus, the reference path provides a stable coordinate system that can be used as a basis for alignment and annotation [28]. However, large nodes containing sequences absent from the reference genome cannot be represented with the linear genome coordinate, prompting the development of coordinate systems that directly encode the topology of the graph [72, 74].

Graphtyper is the first open-source variation graph-based software designed for genotyping from a local (region-specific) graph [78, 82]. It uses a variant file (VCF) to add variant sites and a reference assembly as backbone of the graph. Because of the limited variations included in a VCF file, the output graph is directed and acylic containing insertions and deletions but lacks complex variations (Fig. 1.6b). *Graphtyper* applies a two-step genotyping processes. The "discovery step" is similar to linear reference-guided variant analysis. Sequencing reads are mapped to the linear genome and variants are identified from these alignments. This step is then followed by read realignment towards local graphs. To this end, *Graphtyper* first constructs small regional graphs within 10 kb windows that are subsequently augmented with variants discovered during the first step. Then, *Graphtyper* extracts reads that were initially mapped by the linear mapper, realigns them onto the local graph and performs the variant genotyping from the refined alignments. This approach, however, does not fully eliminate reference bias be-

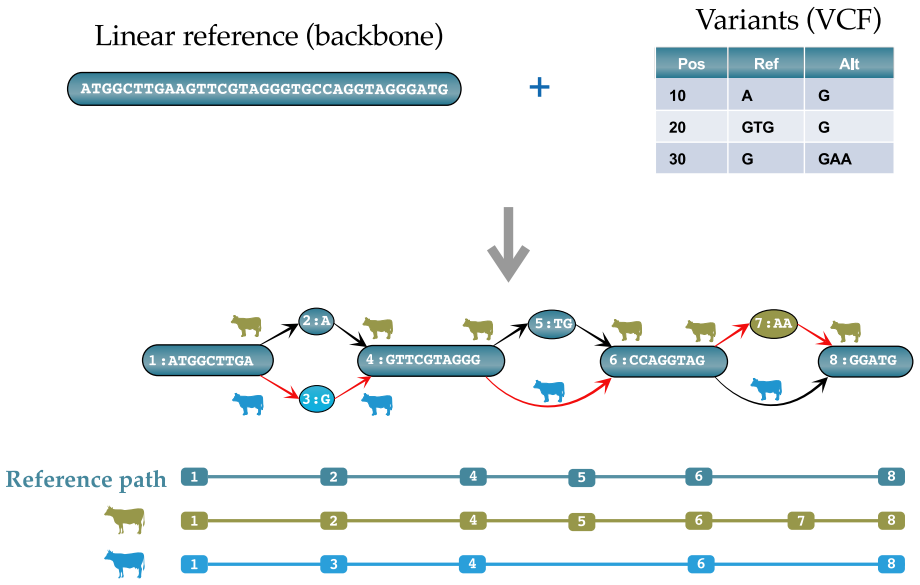


Figure 1.5: **Construction of variation graphs**
A variation graph augments a reference sequence backbone with previously identified genetic variants as alternative nodes (green and light blue nodes). Colored lines represent the path of the reference sequences and haplotypes of the animals included in the graph.

cause it relies on the global read placement by a linear mapper. However, this design makes the graph-based sequence variant analysis become highly efficient as evidenced with scalable joint genotyping of close to 50,000 human genomes [82]. Additionally, *Graphtyper* outperformed current state-of-the-art linear genome-based tools (e.g., *SAMtools* and *GATK*) with regard to genotyping accuracy, particularly from more refined variants surrounding Indels with considerably reduced Mendelian errors observed in parent-offsprint pairs [78].

Construction of the whole-genome variation graphs with the *vg toolkit*

The variation-graph toolkit (*vg*) is the first open-source toolkit designed to perform the full suite of genome analyses from genome graphs in species with gigabase-sized genomes [28]. The basic structure of *vg* is a bidirected sequence graph that can express the stranded-ness of the input sequences (Fig. 1.6c). Each edge endpoint has an independent orientation to indicate whether the forward or reverse sequences are spelled out when visiting the node [72]. Therefore, *vg* can also represent variations with complex topology e.g., inversions or translocations. An auxiliary index is used to store the phasing information so that analysis from the graph can consider haplotypes of the samples [83]. Graph mapping in *vg* is optimized for short sequencing reads that follows the seed-and-extend paradigm. It relies on a *GCSA2* graph index (a generalization of linear genome-based *BWT* index to graphs) for a fast seed query [84]. The index construction is the computationally most demanding step because all *k*-bp paths in the graphs need

to be enumerated, which are intractable in complex regions with high variant density. In practice, *vg* can handle complex regions by indexing them on a simplified graph e.g. retaining only biologically plausible paths informed by the haplotype index [84]. Graph mapping is computationally more expensive than linear mapping because multiple alternative paths need to be explored. To make graph-based mapping competitive to linear mapping, *vg mapper* is currently being improved to utilize the minimizer-based mapping paradigm and by restricting the mapping that conforms the haplotype paths. It can achieve the same mapping speed as the *BWA* (linear mapper) with more accurate alignment performance [85].

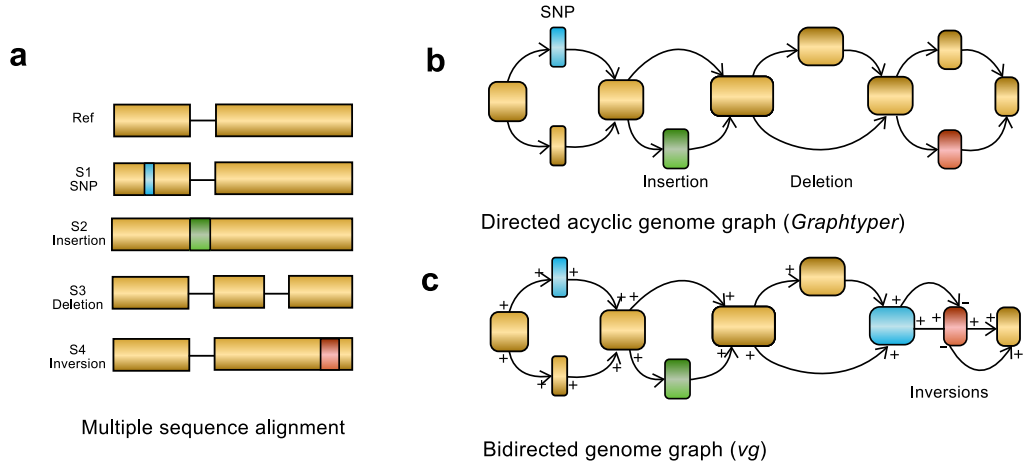


Figure 1.6: **Different genome graph implementations and representations of variations in the graphs**
(a) multiple sequence alignments capturing sequence relationships. **(b)** directed genome graphs underlying the data structure of *Graphtyper*; a similar to multiple sequence alignments but with compressing redundant information. **(c)** general bidirected sequence graph as implemented in *vg* that each edge endpoint has independent orientation. Note forward (+) and reverse strand (-) to indicate inversions (orange). Figures are adapted from Eizenga et al. [74].

1.7 Genome graph construction from a collection of reference-quality assemblies

Multi-assembly graphs as a framework to integrate multiple genome assemblies

The construction of graphs by augmenting a reference genome with a predefined set of variants is still somewhat biased to the reference allele, because the variations are discovered with respect to the reference genome. Additionally, variant identification based on the read alignment is limited by the read length, and thus cannot reliably identify large structural changes between individual genomes [86]. Moreover, the input variant

file format (VCF) can only model simple variations and is not suitable for representing complex structural variations (e.g. SNPs nested inside long insertions) [87]. Building a graph directly from a collection of genome assemblies is a better approach to capture genetic variations. Such a graph will encompass more types of genetic variations, including large structural changes that differ between assemblies (so-called non-reference sequences) that are currently not accessible from linear genomes. This effort will be highly relevant to exploit an ever-increasing number of reference-quality genome assemblies that are being produced at unprecedented rate in order to perform integrative and comprehensive comparative genomics from these resources.

In the multi-assembly graph approach, a graph is constructed from multiple whole-genome alignments (Fig. 1.7). Segments which are present in multiple assemblies (without sufficient variation) are collapsed into a common node, representing conserved regions or core genomes shared in all input samples. The variable regions form bubbles containing multiple paths of the segments that differ (at poorly or non-aligned sequences) between assemblies. Thus, bubbles in the graph represent structural variations across assemblies, with different paths being different alleles.

Strategies to build multi-assembly graphs

Accurate multi-genome alignment is the key for the multi-assembly-based graph approach. However, the alignment of multiple gigabase-sized genomes is computationally demanding and scales poorly with the number of genomes. Recently an efficient multiple-genome alignment approach has been implemented in the *Cactus Progressive* software [88] that scales to hundreds or even thousands of genomes while maintaining high alignment accuracy. The key to its computational efficiency and accuracy is dividing a large whole-genome alignment problem into smaller sub-alignment problems using a guide tree. Whole-genome alignment of more than 600 mammals and birds using *Cactus* enabled a thorough comparative analysis of vertebrate phylogeny [89]. Hickey et al. [90] applied the *vg toolkit* to induce graphs from the *Cactus* alignment of 12 yeast strains. Using this approach, they could map more reads with higher mapping quality, mostly due to mapping improvement in the regions harbouring complex structural variations missed from read alignment-based method.

The approximate mapping between assemblies is another approach to construct multi-assembly graphs. *Minigraph* [91] has recently been developed as a multi-genome graph constructor that extends the minimizer-mapping capability of *minimap2* into a graph [92]. It can establish a pangenome graph from 20 human assemblies in under 3 hours with less than 100 GB of memory. The tool applies an incremental graph gener-

ation. It uses a selected genome as a backbone of the graph which is then iteratively augmented with unaligned or poorly mapped segments from all other assemblies. *Minigraph* implements a hierarchical coordinate system is still retained when more assemblies are added into the graph. Additionally, *minigraph* simplifies the general bidirected sequence graph data model resulting in faster and more straightforward graph analysis. For example, it enforces linearity of the input genomes that produces a graph containing insertions and deletions between genomes but ignoring events that breaks the linearity, such as translocations. Constraining the alignment to an anchor genome also ensures that the graphs lacks of highly tangled parts which are difficult to interpet [93]. Comparative analysis of a pangenome graph built with *minigraph* containing genomes of human and closely related ape species revealed insights into the evolution of repeat-rich regions in primates [91], which was inaccessible from a linear genome. An unpublished graph pipeline (Pangenome Graph Builder, https://github.com/pangenome/pggb) aims at building a comprehensive reference-free graph containing all classes of genetic variations with paths that can reconstruct the entire input sequences. However, this method is still in its infancy and requires further testing.

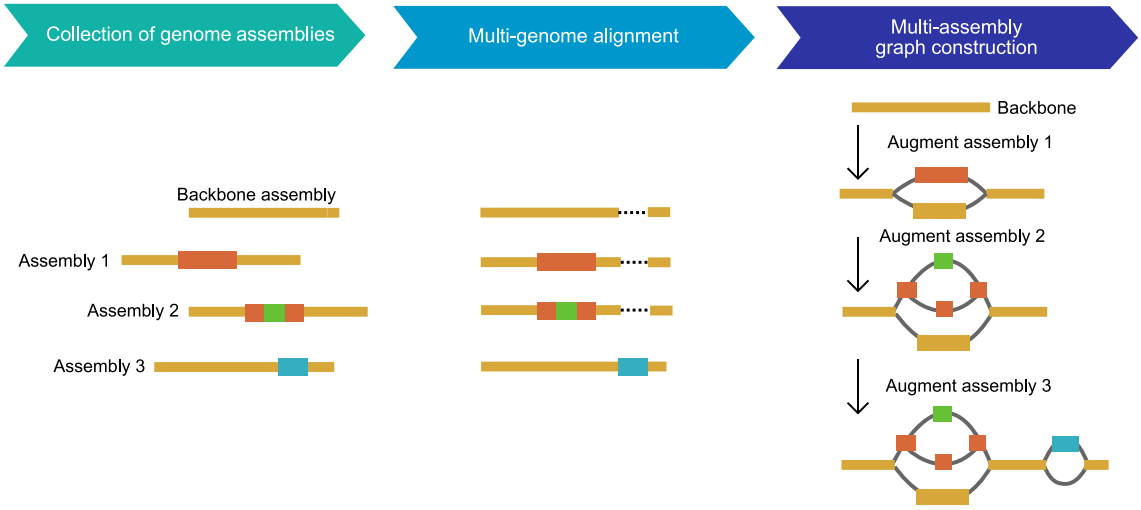


Figure 1.7: **Construction of multi-assembly graphs**
A multi-assembly graph is built based on multi-genome alignment from collection of genome assemblies (*left*, *middle*). In the *minigraph* approach (*right*), the graph is built iteratively from alignments of the genome to the backbone or to the existing graph, which is then augmented with diverged sequences from the alignment.

1.8 Utilization of graph genomes for genomic analyses

Graph genome approaches have been applied to a wide range of genomic analyses, mostly focusing on human or plant genomes. These analyses were initially restricted to challenging regions such as the highly polymorphic Human Leukocyte Antigen (HLA) region [30, 94], where graph-based methods outperform gold-standard linear genome-based genotyping. Several studies [28, 78, 79, 80] then assessed the performance of graph-based methods on whole-genome variant discovery and genotyping. For instance, Garrison et al. [28] constructed a global human graph that contained 80 million variants catalogued by the 1000 Human Genome Project. They showed that genome graphs enable a considerable improvement in read mapping over linear genomes, particularly for reads that differ from the reference, resulting in substantial reduction in the bias of calling genotypes at large indels.

Pritt et al. [27] estimated that with carefully selected variants, genome graphs could rescue 1.2 million reads from 30-fold coverage of human whole-genome sequencing data that are incorrectly mapped to the linear reference. Martiniano et al. [95] applied the *vg* graph framework to an ancient DNA sample to mitigate reference bias due to short and degraded DNA fragments. The benefit of graph-based mapping translates to substantial improvements in calling indels with sufficient accuracy for population genomic inference. Grytten et al. [96] extended the *vg* graph capability to analyse ChIP-Seq data. Using a pangenome of *A. thaliana*, they discovered transcription factor binding sites that are absent in the linear genome. Studying transcription-factor binding motifs from *vg* graphs, Tognon et al. [97] identified variations in regulatory regions affecting gene expression that remained undetected from linear genomes.

Graph genomes were also rigorously exploited to investigate large (structural) variations (SV). SV genotyping mainly relies on the indirect inference of abnormal read alignment profiles (such as depth or split mapping) because the alleles are not present in the reference assembly [44]. Known structural variants can be reliably genotyped once included in the graph, even with short-read data, because the sequencing reads can be directly aligned to the corresponding variants. Sirén et al. [83] constructed a *vg* graph from 167 thousand structural variations detected from long-read data across diverse human ancestry. Reanalysis of 5202 short-read sequencing data using this graph considerably improves the SV genotyping. Subsequent analyses led to the identification of thousands of expression quantitative trait loci (eQTLs) driven by these large variations, largely undetectable from the linear reference genome. Liu et al. [98] applied a similar strategy in the recent soybean pangenome. Re-genotyping of 2898 sequenced samples

from diverse accessions using a pangenome graph integrated from 26 assemblies enabled the identification of a hitherto unknown 10 kb insertion that is associated with a seed phenotype.

1.9 Applications of the pangenomes

Pangenome analyses in plant genomes

Pangenome studies in plants successfully identified a large number of genes not included in the reference and highlight a substantial contribution of large variations to genetic and phenotypic diversity. For example, a pangenome analysis involving 3000 rice accession identified more than 10,000 genes not included in the reference [99]. A considerable number of non-reference insertions associated with agronomic traits, including seed weight and flowering time, were found in a *Brassica* pangenome constructed from eight long-read-based assemblies [100]. Interestingly, GWAS signals from these insertions are significantly stronger than the standard SNPs-based association. Using pangenome constructed from 725 tomato accessions, Gao et al. [64] revealed 4873 genes absent from the reference genome and discovered a 2 kb promoter insertion regulating fruit flavour that was lost during domestication. An increasing number of studies shift towards the graph-based approach, which is pioneered by construction of a graph-based soybean pangenome [98].

Pangenome analyses in human genomes

In humans, pangenome analyses following large scale re-sequencing initiatives revealed several important insights. The 1000 Genomes Project revealed that each genome carries more regions affected by structural variations (8.9 Mb) than small variations (3.6 Mb) [101]. Importantly, they discovered 240 non-reference genes related to immunoglobulin and glycoprotein with homozygous (knock out) deletions in multiple populations, suggesting their dispensable role [42]. Pangenome analyses focusing on the Icelandic population [102] found a common 766 bp insertion (allele frequency of 0.65) associated with decreased risk of myocardial infarctions. A follow-up study based on 3622 samples sequenced using Nanopore (the largest long-read-based pangenome study to date) found that each Icelandic genome carries on average large insertions covering 10.02 Mb genomic regions and identified a tandem repeat motif strongly associated with height [103]. Application of a customized pangenome of the Chinese Han population detected 29.5 Mb non-reference sequences, including 185 genes missing from the reference genome [60]. Sherman et al. [35] reported markedly larger non-reference se-

quences from an African pangenome suggesting the substantial underrepresentation of the African genetic diversity in the current reference genome.

Pangenome analyses in animal genomes

Pangenome approaches have also been applied to livestock and domestic animals, although at a lower extent than in plants or humans. The most notable analysis involves the 44 genomes spanning all extant ruminant families which revealed insights into evolutionary processes [104]. Holden et al. [70] identified 4.6 Mb novel insertions in contigs assembled from non-aligned reads from three dog breeds, that include novel insertions at six known disease-associated loci. Analysis of unmapped reads from the reference individual in song bird (*Parus major*), Laine et al. [71] uncovered 1822 genes missing in the reference annotation, including *TRY1*, which is highly expressed in the reference bird. A similar effort to characterize unmapped reads in the cattle reference animal discovered a number of parasite genomes which are likely to be associated with the reference animal as a host [105].

With a rapid influx of high-quality assemblies, pangenome analyses in animals now transition to the comparison between assemblies. Comparison between Angus (*Bos taurus*) and Brahman (*Bos indicus)* haplotypes-resolved assemblies [25] uncovered an extra copy of *FADS2P1* gene in *Bos indicus*, which is proposed to confer heat resistance. Analysis of unaligned sequences between 10 goat assemblies [59] recovered 38.3 Mb non-reference insertions and identified a large mis-assembled regions in the ARS-1 goat reference genome that includes the prolactin gene region. Analysis of 12 Eurasian pig assemblies retrieved 72.5 Mb novel insertions that are absent in the Duroc-based reference assembly [57, 58]. Additionally, this study also discovered a non-reference insertion that segregates at high frequency in Chinese breeds (but not in European breeds) encompassing the *TIG3* gene region, which is important for fatty acid metabolism.

1.10 Main knowledge gaps

A well-annotated reference genome is the foundation for genomic analyses. However, The current *Bos taurus* linear reference genome represents a mosaic haplotype from a single individual [16], which poorly represents the cattle diversity across the globe. This inadvertently introduces reference bias when DNA sequences that are diverged from the reference are compared to the reference sequences. A variation-aware, graph-based reference structure, is needed for accurate and unbiased genomic analysis in the cattle population.

Well-established graph genome methods, implemented in software like *Graphtyper* [78] and *vg* [28], construct graphs by augmenting linear reference backbones with known variants. These graph-based approaches are currently tailored towards human genomics applications and have never been applied to other gigabase-sized genomes. Large international initiatives, such as the 1000 Bull Genomes Project [11], have provided an exhaustive catalogue of variants segregating across global breeds of cattle, prompting the use of these resources to improve genomic analysis in livestock species. Thus, it is appealing to develop genome-graph techniques also in *Bos taurus* to create a variation-aware reference genome.

Variant prioritization is critical to develop informative graph genomes [27]. Thus, a thorough assessment of factors affecting variant prioritization is required to establish informative cattle genome graphs. Specifically, it is important to investigate whether a unified pangenome graph is as informative as the breed-specific graphs. Due to subdivision of cattle into multiple breeds with distinct genetic and phenotypic characteristics, cattle are an ideal species to address this research question.

Multiple studies [33, 35, 57] have reported that reference genomes may lack of millions of nucleotides with unknown functional relevance. However, this quantity has never been determined in cattle populations. Multi-assembly graphs now provide a framework to establish the full pangenome of a species. Specifically, with the availability of reference-quality genome assemblies across numerous cattle breeds [16, 17, 18], a multi-assembly graph may integrate these resources into a unified variation-aware reference structure. Moreover, it is desirable to develop an efficient end-to-end computational pipeline to build comprehensive pangenome graphs and characterize sequence variations in the graphs, which is currently lacking for any livestock population.

References

[1] K Zhang, JA Lenstra, S Zhang, W Liu, and J Liu. Evolution and domestication of the Bovini species. *Animal Genetics*, 51(5):637–657, 2020.

[2] Michel Georges, Carole Charlier, and Ben Hayes. Harnessing genomic information for livestock improvement. *Nature Reviews Genetics*, 20(3):135–156, 2019.

[3] Theo HE Meuwissen, Ben J Hayes, and Michael E Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, 2001.

[4] Michael E Goddard and Ben J Hayes. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics*, 10(6):381–391, 2009.

[5] George R Wiggans, John B Cole, Suzanne M Hubbard, and Tad S Sonstegard. Genomic selection in dairy cattle: the USDA experience. *Annual review of animal biosciences*, 5:309–327, 2017.

[6] Hubert Pausch, Iona M MacLeod, Ruedi Fries, Reiner Emmerling, Phil J Bowman, Hans D Daetwyler, and Michael E Goddard. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. *Genetics Selection Evolution*, 49(1):1–14, 2017.

[7] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43 (5):491, 2011.

[8] Antonio Regalado. China BGI says it can sequence a genome for just 100 USD, 2020. URL https://www.technologyreview.com/2020/02/26/905658/china-bgi-100-dollar-genome/. Accessed 10 April 2021.

[9] Kris A. Wetterstrand. Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP), 2020. URL www.genome.gov/sequencingcostsdata. Accessed 10 April 2021.

[10] Hans D Daetwyler, Aurélien Capitan, Hubert Pausch, Paul Stothard, Rianne Van Binsbergen, Rasmus F Brøndum, Xiaoping Liao, Anis Djari, Sabrina C Rodriguez, Cécile 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.

[11] Ben J Hayes and Hans D Daetwyler. 1000 bull genomes project to map simple and complex genetic traits in cattle: applications and outcomes. *Annual review of animal biosciences*, 7:89–102, 2019.

[12] Warren M Snelling, Jesse L Hoff, Jeremiah H Li, Larry A Kuehn, Brittney N Keel, Amanda K Lindholm-Perry, and Joseph K Pickrell. Assessment of Imputation from Low-Pass Sequencing to Predict Merit of Beef Steers. *Genes*, 11(11):1312, 2020.

[13] DM Bickhart, JC McClure, RD Schnabel, BD Rosen, JF Medrano, and TPL Smith. Symposium review: advances in sequencing technology herald a new frontier in cattle genomics and genome-enabled selection. *Journal of dairy science*, 2020.

[14] Christine G Elsik, Ross L Tellam, Kim C Worley, et al. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*, 324(5926):522–528, 2009.

[15] Aleksey V Zimin, Arthur L Delcher, Liliana Florea, David R Kelley, Michael C Schatz, Daniela Puiu, Finnian Hanrahan, Geo Pertea, Curtis P Van Tassell, Tad S Sonstegard, et al. A whole-genome assembly of the domestic cow, Bos taurus. *Genome biology*, 10(4):1–10, 2009.

[16] Benjamin D Rosen, Derek M Bickhart, Robert D Schnabel, Sergey Koren, Christine G Elsik, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience*, 9(3): giaa021, 2020.

[17] Edward S Rice, Sergey Koren, Arang Rhie, Michael P Heaton, Theodore S Kalbfleisch, et al. Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. *Gigascience*, 9(4):giaa029, 2020.

[18] Sergey Koren, Arang Rhie, Brian P Walenz, Alexander T Dilthey, Derek M Bickhart, Sarah B Kingan, Stefan Hiendleder, John L Williams, Timothy PL Smith, and Adam M Phillippy. De novo assembly of haplotype-resolved genomes with trio binning. *Nature biotechnology*, 36(12):1174–1182, 2018.

[19] Jonas Oppenheimer, Benjamin D Rosen, Michael P Heaton, Brian L Vander Ley, Wade R Shafer, Fred T Schuetze, Brad Stroud, Larry A Kuehn, Jennifer C McClure, Jennifer P Barfield, et al. A reference genome assembly of American bison, Bison bison bison. *Journal of Heredity*, 112(2):174–183, 2021.

[20] Michael P Heaton, Timothy PL Smith, Derek M Bickhart, Brian L Vander Ley, Larry A Kuehn, Jonas Oppenheimer, Wade R Shafer, Fred T Schuetze, Brad Stroud, Jennifer C McClure, et al. A reference genome assembly of Simmental cattle, Bos taurus taurus. *Journal of Heredity*, 2021.

[21] Amanda Warr, Nabeel Affara, Bronwen Aken, Hamid Beiki, Derek M Bickhart, et al. An improved pig reference genome sequence to enable pig genetics and genomics research. *GigaScience*, 9(6):giaa051, 2020.

[22] Derek M Bickhart, Benjamin D Rosen, Sergey Koren, Brian L Sayre, Alex R Hastie, Saki Chan, Joyce Lee, Ernest T Lam, Ivan Liachko, Shawn T Sullivan, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature genetics*, 49(4):643–650, 2017.

[23] Kerstin Lindblad-Toh, Claire M Wade, Tarjei S Mikkelsen, Elinor K Karlsson, David B Jaffe, Michael Kamal, Michele Clamp, Jean L Chang, Edward J Kulbokas, Michael C Zody, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069):803–819, 2005.

[24] Sara Ballouz, Alexander Dobin, and Jesse A Gillis. Is it time to change the reference genome? *Genome biology*, 20(1):1–9, 2019.

[25] Wai Yee Low, Rick Tearle, Ruijie Liu, Sergey Koren, Arang Rhie, Derek M. Bickhart, Benjamin D. Rosen, et al. Haplotype-Resolved Cattle Genomes Provide Insights Into Structural Variation and Adaptation. *Nature Communications*, 11(1), aug 2020.

[26] Harsh G Shukla, Pushpinder Singh Bawa, and Subhashini Srinivasan. hg19KIndel: ethnicity normalized human reference genome. *BMC genomics*, 20(1):1–17, 2019.

[27] Jacob Pritt, Nae-Chyun Chen, and Ben Langmead. FORGe: prioritizing variants for graph genomes. *Genome biology*, 19(1):1–16, 2018.

[28] Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature biotechnology*, 36(9):875–879, 2018.

[29] Mazdak Salavati, Stephen J Bush, Sergio Palma-Vera, Mary EB McCulloch, David A Hume, and Emily L Clark. Elimination of reference mapping bias reveals robust immune related allele-specific expression in crossbred sheep. *Frontiers in genetics*, 10:863, 2019.

[30] Alexander Dilthey, Charles Cox, Zamin Iqbal, Matthew R Nelson, and Gil McVean. Improved genome inference in the MHC using a population reference graph. *Nature genetics*, 47(6):682–688, 2015.

[31] Dong-Dong Wu, Xiang-Dong Ding, Sheng Wang, Jan M Wójcik, YI Zhang, Małgorzata Tokarska, Yan Li, Ming-Shan Wang, Omar Faruque, Rasmus Nielsen, et al. Pervasive introgression facilitated domestication and adaptation in the Bos species complex. *Nature ecology & evolution*, 2(7):1139–1145, 2018.

[32] Rachel M Colquhoun, Michael B Hall, Leandro Lima, Leah W Roberts, Kerri M Malone, Martin Hunt, Brice Letcher, Jane Hawkey, Sophie George, Louise Pankhurst, et al. Nucleotide-resolution bacterial pan-genomics with reference graphs. *bioRxiv*, 2020.

[33] Adam Ameur, Huiwen Che, Marcel Martin, Ignas Bunikis, Johan Dahlberg, Ida Höijer, Susana Häggqvist, Francesco Vezzi, Jessica Nordlund, Pall 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.

[34] Peter A Audano, Arvis Sulovari, Tina A Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, AnneMarie E Welch, Max L Dougherty, Bradley J Nelson, Ankeeta Shah, Susan K Dutcher, et al. Characterizing the major structural variant alleles of the human genome. *Cell*, 176(3):663–675, 2019.

[35] Rachel M Sherman, Juliet Forman, Valentin Antonescu, Daniela Puiu, Michelle Daya, Nicholas Rafaels, Meher Preethi Boorgula, Sameer Chavan, Candelaria Vergara, Victor 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.

[36] Carole Charlier, Wanbo Li, Chad Harland, Mathew Littlejohn, Wouter Coppieters, Frances Creagh, Steve Davis, Tom Druet, Pierre Faux, François Guillaume, et al. NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. *Genome research*, 26(10): 1333–1341, 2016.

[37] Daniel Pitt, Natalia Sevane, Ezequiel L Nicolazzi, David E MacHugh, Stephen DE Park, Licia Colli, Rodrigo Martinez, Michael W Bruford, and Pablo Orozco-terWengel. Domestication of cattle: Two or three events? *Evolutionary applications*, 12(1):123–136, 2019.

[38] Heidi Signer-Hasler, Alexander Burren, Markus Neuditschko, Mirjam Frischknecht, Dorian Garrick, Christian Stricker, Birgit Gredler, Beat Bapst, and Christine Flury. Population structure and genomic inbreeding in nine Swiss dairy cattle populations. *Genetics Selection Evolution*, 49(1):1–13, 2017.

[39] Maulik Upadhyay, Susanne Eriksson, Sofia Mikko, Erling Strandberg, Hans Stålhammar, Martien AM Groenen, Richard PMA Crooijmans, Göran Andersson, and Anna M Johansson. Genomic relatedness and diversity of Swedish native cattle breeds. *Genetics Selection Evolution*, 51(1):1–11, 2019.

[40] L Koufariotis, BJ Hayes, M Kelly, BM Burns, R Lyons, P Stothard, AJ 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.

[41] Kwondo Kim, Taehyung Kwon, Tadelle Dessie, DongAhn Yoo, Okeyo Ally Mwai, Jisung Jang, Samsun Sung, SaetByeol Lee, Bashir Salim, Jaehoon 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.

[42] Peter H Sudmant, Tobias Rausch, Eugene J Gardner, Robert E Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, 2015.

[43] Jaemin Kim, Olivier Hanotte, Okeyo Ally Mwai, Tadelle Dessie, Salim Bashir, Boubacar Diallo, Morris Agaba, Kwondo Kim, Woori Kwak, Samsun Sung, et al. The genome landscape of indigenous African cattle. *Genome biology*, 18(1):1–14, 2017.

[44] Medhat Mahmoud, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J Sedlazeck. Structural variant calling: the long and the short of it. *Genome biology*, 20 (1):1–14, 2019.

[45] Mark JP Chaisson, Ashley D Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J Gardner, Oscar L Rodriguez, Li Guo, Ryan L Collins, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature communications*, 10(1): 1–16, 2019.

[46] Benjamin Kaminow, Sara Ballouz, Jesse Gillis, and Alexander Dobin. Virtue as the mean: Pan-human consensus genome significantly improves the accuracy of RNA-seq analyses. *bioRxiv*, 2020.

[47] Nae-Chyun Chen, Brad Solomon, Taher Mun, Sheila Iyer, and Ben Langmead. Reference flow: reducing reference bias using multiple population genomes. *Genome biology*, 22(1):1–17, 2021.

[48] Rachel M Sherman and Steven L Salzberg. Pan-genomics in the human genome era. *Nature Reviews Genetics*, 21(4):243–254, 2020.

[49] Karen H Miga, Sergey Koren, Arang Rhie, Mitchell R Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, Edmund Howe, David Porubsky, Glennis A Logsdon, et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, 585(7823):79–84, 2020.

[50] Glennis A Logsdon, Mitchell R Vollger, PingHsun Hsieh, Yafei Mao, Mikhail A Liskovykh, Sergey Koren, Sergey Nurk, Ludovica Mercuri, Philip C Dishuck, Arang Rhie, et al. The structure, function and evolution of a complete human chromosome 8. *Nature*, pages 1–7, 2021.

[51] Arang Rhie, Shane A McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, William Chow, Arkarachai Fungtammasan, Juwan Kim, Lee, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, 592(7856):737–746, 2021.

[52] Harris A Lewin, Gene E Robinson, W John Kress, William J Baker, Jonathan Coddington, Keith A Crandall, Richard Durbin, Scott V Edwards, Félix Forest, M Thomas P Gilbert, et al. Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, 115(17): 4325–4333, 2018.

[53] Karen H. Miga and Ting Wang. The Need for a Human Pangenome Reference Sequence. *Annual Review of Genomics and Human Genetics*, 22(1), 2021.

[54] Philipp E Bayer, Agnieszka A Golicz, Armin Scheben, Jacqueline Batley, and David Edwards. Plant pan-genomes are the new reference. *Nat. Plants*, 6:914–920, 2020.

[55] Rafael Della Coletta, Yinjie Qiu, Shujun Ou, Matthew B Hufford, and Candice N Hirsch. How the pan-genome is changing crop genomics and improvement. *Genome biology*, 22(1):1–19, 2021.

[56] Hervé Tettelin, Vega Masignani, Michael J Cieslewicz, Claudio Donati, Duccio Medini, Naomi L Ward, Samuel V Angiuoli, Jonathan Crabtree, Amanda L Jones, A Scott Durkin, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, 102(39):13950–13955, 2005.

[57] Mingzhou Li, Lei Chen, Shilin Tian, Yu Lin, Qianzi Tang, Xuming Zhou, Diyan Li, Carol KL Yeung, Tiandong Che, Long 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.

[58] Xiaomeng Tian, Ran Li, Weiwei Fu, Yan Li, Xihong Wang, Ming Li, Duo Du, Qianzi Tang, Yudong Cai, Yiming 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.

[59] Ran Li, Weiwei Fu, Rui Su, Xiaomeng Tian, Duo Du, Yue Zhao, Zhuqing Zheng, Qiuming Chen, Shan Gao, Yudong Cai, et al. Towards the complete goat pan-genome by recovering missing genomic segments from the reference genome. *Frontiers in genetics*, 10, 2019.

[60] Zhongqu Duan, Yuyang Qiao, Jinyuan Lu, Huimin Lu, Wenmin Zhang, Fazhe Yan, Chen Sun, Zhiqiang Hu, Zhen Zhang, Guichao Li, et al. HUPAN: a pan-genome analysis pipeline for human genomes. *Genome biology*, 20(1):1–11, 2019.

[61] Agnieszka A Golicz, Philipp E Bayer, Prem L Bhalla, Jacqueline Batley, and David Edwards. Pangenomics comes of age: from bacteria to plant and animal applications. *Trends in Genetics*, 36(2):132–145, 2020.

[62] Marco Gerdol, Rebeca Moreira, Fernando Cruz, Jessica Gómez-Garrido, Anna Vlasova, Umberto Rosani, Paola Venier, Miguel A Naranjo-Ortiz, Maria Murgarella, Samuele Greco, et al. Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. *Genome biology*, 21(1):1–21, 2020.

[63] Qiang Zhao, Qi Feng, Hengyun Lu, Yan Li, Ahong Wang, Qilin Tian, Qilin Zhan, Yiqi Lu, Lei Zhang, Tao Huang, et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature genetics*, 50(2):278–284, 2018.

[64] Lei Gao, Itay Gonda, Honghe Sun, Qiyue Ma, Kan Bao, Denise M Tieman, Elizabeth A Burzynski-Chang, Tara L Fish, Kaitlin A Stromberg, Gavin L Sacks, et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature genetics*, 51(6):1044–1051, 2019.

[65] Sean Walkowiak, Liangliang Gao, Cecile Monat, Georg Haberer, et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature*, pages 1–7, 2020.

[66] Yongfu Tao, Xianrong Zhao, Emma Mace, Robert Henry, and David Jordan. Exploring and exploiting pan-genomics for crop improvement. *Molecular plant*, 12(2):156–169, 2019.

[67] Christine Tranchant-Dubreuil, Mathieu Rouard, and Francois Sabot. Plant pangenome: impacts on phenotypes and evolution. *Annual Plant Reviews Online*, pages 453–478, 2018.

[68] Shannon M Soucy, Jinling Huang, and Johann Peter Gogarten. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16(8):472–482, 2015.

[69] Jesper Eisfeldt, Gustaf Mårtensson, Adam Ameur, Daniel Nilsson, and Anna Lindstrand. Discovery of novel sequences in 1,000 Swedish genomes. *Molecular biology and evolution*, 37(1):18–30, 2020.

[70] Lindsay A Holden, Meharji Arumilli, Marjo K Hytönen, Sruthi Hundi, Jarkko Salojärvi, Kim H Brown, and Hannes Lohi. Assembly and analysis of unmapped genome sequence reads reveal novel sequence and variation in dogs. *Scientific reports*, 8(1):1–11, 2018.

[71] Veronika N Laine, Toni I Gossmann, Kees van Oers, Marcel E Visser, and Martien AM Groenen. Exploring the unmapped DNA and RNA reads in a songbird genome. *BMC genomics*, 20(1):1–12, 2019.

[72] Benedict Paten, Adam M Novak, Jordan M Eizenga, and Erik Garrison. Genome graphs and the evolution of genome inference. *Genome research*, 27(5):665–676, 2017.

[73] The Computational Pangenomics Consortium. Computational pan-genomics: status, promises and challenges. *Briefings in bioinformatics*, 19(1):118–135, 2018.

[74] Jordan M Eizenga, Adam M Novak, Jonas A Sibbesen, Simon Heumos, Ali Ghaffaari, Glenn Hickey, Xian Chang, Josiah D Seaman, Robin Rounthwaite, Jana Ebler, et al. Pangenome graphs. *Annual Review of Genomics and Human Genetics*, 21:139–162, 2020.

[75] Benedict Paten, Jordan M Eizenga, Yohei M Rosen, Adam M Novak, Erik Garrison, and Glenn Hickey. Superbubbles, ultrabubbles, and cacti. *Journal of Computational Biology*, 25(7):649–663, 2018.

[76] Zamin Iqbal, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature genetics*, 44(2):226–232, 2012.

[77] Isaac Turner, Kiran V Garimella, Zamin Iqbal, and Gil McVean. Integrating long-range connectivity information into de Bruijn graphs. *Bioinformatics*, 34(15):2556–2565, 2018.

[78] Hannes P Eggertsson, Hakon Jonsson, Snaedis Kristmundsdottir, Eirikur Hjartarson, Birte Kehr, Gisli Masson, Florian Zink, Kristjan E Hjorleifsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, et al. Graphtyper enables population-scale genotyping using pangenome graphs. *Nature genetics*, 49(11): 1654, 2017.

[79] Jonas Andreas Sibbesen, Lasse Maretty, and Anders Krogh. Accurate genotyping across variant classes and lengths using variant graphs. *Nature genetics*, 50(7):1054–1059, 2018.

[80] Goran Rakocevic, Vladimir Semenyuk, Wan-Ping Lee, James Spencer, John Browning, Ivan J Johnson, Vladan Arsenijevic, Jelena Nadj, Kaushik Ghose, Maria C Suciu, et al. Fast and accurate genomic analyses using genome graphs. *Nature genetics*, 51(2):354–362, 2019.

[81] Daehwan Kim, Joseph M Paggi, Chanhee Park, Christopher Bennett, and Steven L Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology*, 37(8):907–915, 2019.

[82] Hannes P Eggertsson, Snaedis Kristmundsdottir, Doruk Beyter, Hakon Jonsson, Astros Skuladottir, Marteinn T Hardarson, Daniel F Gudbjartsson, Kari Stefansson, Bjarni V Halldorsson, and Pall Melsted. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature communications*, 10(1):1–8, 2019.

[83] Jouni Sirén, Erik Garrison, Adam M Novak, Benedict Paten, and Richard Durbin. Haplotype-aware graph indexes. *Bioinformatics*, 36(2):400–407, 2020.

[84] Jouni Sirén. Indexing variation graphs. In *2017 Proceedings of the ninteenth workshop on algorithm engineering and experiments (ALENEX)*, pages 13–27. SIAM, 2017.

[85] Jouni Sirén, Jean Monlong, Xian Chang, Adam M Novak, Jordan M Eizenga, Charles Markello, Jonas Sibbesen, Glenn Hickey, Pi-Chuan Chang, Andrew Carroll, et al. Genotyping common, large structural variations in 5,202 genomes using pangenomes, the Giraffe mapper, and the vg toolkit. *Biorxiv*, 2020.

[86] Xiaowen Feng and Heng Li. Higher rates of processed pseudogene acquisition in humans and three great apes revealed by long read assemblies. *bioRxiv*, 2020.

[87] Brice Letcher, Martin Hunt, and Zamin Iqbal. Enabling multiscale variation analysis with genome graphs. *bioRxiv*, 2021.

[88] Joel Armstrong, Glenn Hickey, Mark Diekhans, Ian T Fiddes, Adam M Novak, Alden Deran, Qi Fang, Duo Xie, Shaohong Feng, Josefin Stiller, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, 587(7833):246–251, 2020.

[89] Shaohong Feng, Josefin Stiller, Yuan Deng, Joel Armstrong, Qi Fang, Andrew Hart Reeve, Duo Xie, Guangji Chen, Chunxue Guo, Brant C Faircloth, et al. Dense sampling of bird diversity increases power of comparative genomics. *Nature*, 587(7833):252–257, 2020.

[90] Glenn Hickey, David Heller, Jean Monlong, Jonas A Sibbesen, Jouni Sirén, Jordan Eizenga, Eric T Dawson, Erik Garrison, Adam M Novak, and Benedict Paten. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome biology*, 21(1):1–17, 2020.

[91] Heng Li, Xiaowen Feng, and Chong Chu. The design and construction of reference pangenome graphs with minigraph. *Genome biology*, 21(1):1–19, 2020.

[92] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[93] Li Lei, Eugene Goltsman, David Goodstein, Guohong Albert Wu, Daniel S Rokhsar, and John P Vogel. Plant Pan-Genomics Comes of Age. *Annual Review of Plant Biology*, 72(1), 2021.

[94] Heewook Lee and Carl Kingsford. Kourami: graph-guided assembly for novel human leukocyte antigen allele discovery. *Genome biology*, 19(1):1–16, 2018.

[95] Rui Martiniano, Erik Garrison, Eppie R Jones, Andrea Manica, and Richard Durbin. Removing reference bias in ancient DNA data analysis by mapping to a sequence variation graph. *bioRxiv*, 2019.

[96] Ivar Grytten, Knut D Rand, Alexander J Nederbragt, Geir O Storvik, Ingrid K Glad, and Geir K Sandve. Graph Peak Caller: Calling ChIP-seq peaks on graph-based reference genomes. *PLoS computational biology*, 15(2):e1006731, 2019.

[97] Manuel Tognon, Vincenzo Bonnici, Erik Garrison, Rosalba Giugno, and Luca Pinello. GRAFIMO: variant and haplotype aware motif scanning on pangenome graphs. *bioRxiv*, 2021.

[98] Yucheng Liu, Huilong Du, Pengcheng Li, Yanting Shen, Hua Peng, Shulin Liu, Guo-An Zhou, Haikuan Zhang, Zhi Liu, Miao Shi, et al. Pan-genome of wild and cultivated soybeans. *Cell*, 182 (1):162–176, 2020.

[99] Wensheng Wang, Ramil Mauleon, Zhiqiang Hu, Dmytro Chebotarov, Shuaishuai Tai, Zhichao Wu, Min Li, Tianqing Zheng, Roven Rommel Fuentes, Fan Zhang, et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, 557(7703):43–49, 2018.

[100] Jia-Ming Song, Zhilin Guan, Jianlin Hu, Chaocheng Guo, Zhiquan Yang, Shuo Wang, Dongxu Liu, Bo Wang, Shaoping Lu, Run Zhou, et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nature Plants*, 6(1):34–45, 2020.

[101] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526 (7571):68, 2015.

[102] Birte Kehr, Anna Helgadottir, Pall Melsted, Hakon Jonsson, Hannes Helgason, Adalbjörg Jonasdottir, Aslaug Jonasdottir, Asgeir Sigurdsson, Arnaldur Gylfason, Gisli H Halldorsson, et al. Diversity in non-repetitive human sequences not found in the reference genome. *Nature Genetics*, 49(4):588–593, 2017.

[103] Doruk Beyter, Helga Ingimundardottir, Asmundur Oddsson, Hannes P Eggertsson, et al. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nature Genetics*, 2021.

[104] Lei Chen, Qiang Qiu, Yu Jiang, Kun Wang, Zeshan Lin, Zhipeng Li, Faysal Bibi, Yongzhi Yang, Jinhuan Wang, Wenhui Nie, et al. Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science*, 364(6446), 2019.

[105] Lynsey K Whitacre, Polyana C Tizioto, Kim, et al. What's in your next-generation sequence data? An exploration of unmapped sequence reads from the bovine reference individual. *BMC genomics*, 16(1): 1–7, 2015.

Chapter 2

Accurate sequence variant genotyping in cattle using variation-aware genome graphs

Danang Crysnanto[1], Christine Wurmser[2], Hubert Pausch[1]

[1] Animal Genomics, ETH Zurich, Zurich, Switzerland.
[2] Chair of Animal Breeding, TU München, Freising, Germany.

Published in *Genetics Selection Evolution (2019) 51:21*

Contribution: I participated in conceiving the study, analysing the results and writing the manuscript. I wrote the graph genotyping pipelines.

Abstract

Background: The genotyping of sequence variants typically involves as a first step the alignment of sequencing reads to a linear reference genome. Because a linear reference genome represents only a small fraction of sequence variation within a species, reference allele bias may occur at highly polymorphic or diverged regions of the genome. Graph-based methods facilitate to compare sequencing reads to a variation-aware genome graph that incorporates a collection of non-redundant DNA sequences that segregate within a species. We compared accuracy and sensitivity of graph-based sequence variant genotyping using the *Graphtyper* software to two widely used methods, i.e., *GATK* and *SAMtools*, that rely on linear reference genomes using whole-genomes sequencing data of 49 Original Braunvieh cattle.

Results: We discovered 21,140,196, 20,262,913 and 20,668,459 polymorphic sites using *GATK*, *Graphtyper*, and *SAMtools*, respectively. Comparisons between sequence variant and microarray-derived genotypes showed that *Graphtyper* outperformed both *GATK* and *SAMtools* in terms of genotype concordance, non-reference sensitivity, and non-reference discrepancy. The sequence variant genotypes that were obtained using *Graphtyper* had the lowest number of mendelian inconsistencies for both SNPs and indels in nine sire-son pairs with sequence data. Genotype phasing and imputation using the *Beagle* software improved the quality of the sequence variant genotypes for all tools evaluated particularly for animals that have been sequenced at low coverage. Following imputation, the concordance between sequence- and microarray-derived genotypes was almost identical for the three methods evaluated, i.e., 99.32, 99.46, and 99.24 % for *GATK*, *Graphtyper*, and *SAMtools*, respectively. Variant filtration based on commonly used criteria improved the genotype concordance slightly but it also decreased sensitivity. *Graphtyper* required considerably more computing resources than *SAMtools* but it required less than *GATK*.

Conclusions: Sequence variant genotyping using *Graphtyper* is accurate, sensitive and computationally feasible in cattle. Graph-based methods enable sequence variant genotyping from variation-aware reference genomes that may incorporate cohort-specific sequence variants which is not possible with the current implementations of state-of-the-art methods that rely on linear reference genomes.

Keywords: Sequence variant genotyping, Genome graph, Variation-aware graph, cattle, Whole-genome sequencing

2.1 Introduction

The sequencing of important ancestors of many cattle breeds revealed millions of sequence variants that are polymorphic in dairy and beef populations [1, 2, 3, 4]. In order to compile an exhaustive catalog of polymorphic sites that segregate in Bos taurus, the 1000 Bull Genomes consortium was established [5, 6]. The 1000 Bull Genomes Project imputation reference panel facilitates to infer sequence variant genotypes for large cohorts of genotyped animals thus enabling genomic investigations at nucleotide resolution [5, 7, 8, 9].

Sequence variant discovery and genotyping typically involves two steps that are carried out successively [10, 11, 12, 13]: first, raw sequencing data are generated, trimmed and filtered to remove adapter sequences and bases with low sequencing quality, respectively, and aligned towards a linear reference genome using, e.g., *Bowtie* [14] or the Burrows-Wheeler Alignment (*BWA*) software [15]. The aligned reads are subsequently compared to the nucleotide sequence of a reference genome in order to discover and genotype polymorphic sites using, e.g., *SAMtools* [16] or the Genome Analysis Toolkit (*GATK*) [17, 18, 19]. Variant discovery may be performed either in single- or multi-sample mode. The accuracy (i.e., ability to correctly genotype sequence variants) and sensitivity (i.e., ability to detect true sequence variants) of sequence variant discovery is higher using multi-sample than single-sample approaches particularly when the sequencing depth is low [20, 21, 22, 23, 24]. However, the genotyping of sequence variants from multiple samples simultaneously is a computationally intensive task, particularly when the sequenced cohort is large and diverse and had been sequenced at high coverage [19]. The multi-sample sequence variant genotyping approach that has been implemented in the *SAMtools* software has to be restarted for the entire cohort once new samples are added.*GATK* implements two different approaches to multi-sample variant discovery, i.e., the *UnifiedGenotyper* and *HaplotypeCaller* modules, with the latter relying on intermediate files in *gVCF* format that include probabilistic data on variant and non-variant sites for each sequenced sample. Applying the *HaplotypeCaller* module allows for separating variant discovery within samples from the estimation of genotype likelihoods across samples. Once new samples are added to an existing cohort, only the latter needs to be performed for the entire cohort, thus enabling computationally efficient parallelization of sequence variant genotyping in a large number of samples.

Sequence variant genotyping approaches that rely on alignments to a linear reference genome have limitations for variant discovery, because a haploid reference sequence does not reflect variation within a species. As a result, read alignments may

be erroneous particularly at genomic regions that differ substantially between the sequenced individual and the reference sequence, thus introducing reference allele bias, flawed genotypes, and false-positive variant discovery around indels [25, 26, 27]. Aligning reads to population- or breed-specific reference genomes may overcome most of these limitations [28, 29, 30]. However, considering multiple (population-specific) linear reference genomes with distinct genomic coordinates complicates the biological interpretation and annotation of sequence variant genotypes across populations [31].

Genome graph-based methods consider non-linear reference sequences for variant discovery [31, 32, 33, 34, 35]. A variation-aware genome graph may incorporate distinct (population-specific) reference sequences and known sequence variants. Recently, the *Graphtyper* software has been developed in order to facilitate sequence variant discovery from a genome graph that has been constructed and iteratively augmented using variation of the sequenced cohort [32]. So far, sequence variant genotyping using variation-aware genome graphs has not been evaluated in cattle.

An unbiased evaluation of the accuracy and sensitivity of sequence variant genotyping is possible when high confidence sequence variants and genotypes are accessible that were detected using genotyping technologies and algorithms different from the ones to be evaluated [36]. For species where such a resource is not available, the accuracy of sequence variant genotyping may be evaluated by comparing sequence variant to microarray-derived genotypes (e.g., [4, 24]). Due to the ascertainment bias in SNP chip data, this comparison may overestimate the accuracy of sequence variant discovery particularly at variants that are either rare or located in less-accessible genomic regions [37, 38].

In this study, we compared sequence variant discovery and genotyping from a variation-aware genome graph using *Graphtyper* to two state-of-the-art methods (*GATK*, *SAMtools*) that rely on linear reference genomes in 49 Original Braunvieh cattle. We compared sequence variant to microarray-derived genotypes in order to assess accuracy and sensitivity of sequence variant genotyping for each of the three methods evaluated.

2.2 Methods

Selection of animals We selected 49 Original Braunvieh (OB) bulls that were either frequently used in artificial insemination or explained a large fraction of the genetic diversity of the active breeding population. Semen straws of the bulls were purchased

from an artificial insemination center and DNA was prepared following standard DNA extraction protocols.

Sequencing data pre-processing All samples were sequenced on either an Illumina HiSeq 2500 (30 animals) or an Illumina HiSeq 4000 (19 animals) sequencer using 150 bp paired-end sequencing libraries with insert sizes ranging from 400 to 450 bp. Quality control (removal of adapter sequences and bases with low quality) of the raw sequencing data was caried out using the *fastp* software (version 0.19.4) with default parameters [39]. The filtered reads were mapped to the UMD3.1 version of the bovine reference genome [40] using *BWA mem* (version 0.7.12) [15] with option-M to mark shorter split hits as secondary alignments, default parameters were applied in all other steps. Optical and PCR duplicates were marked using *Samblaster* (version 0.1.24) [41]. The output of *Samblaster* was converted into *BAM* format using *SAMtools view* (version 1.3) [16], and subsequently coordinate-sorted using *Sambamba* (version 0.6.6) [42]. We used the *GATK* (version 3.8) *RealignerTargetCreator* and *IndelRealigner* modules to realign reads around indels. The realigned BAM files served as input for *GATK* base quality score recalibration using 102,092,638 unique positions from the Illumina BovineHD SNP chip and Bovine dbSNP version 150, as known variants. The *mosdepth* software (version 0.2.2) [43] was used to extract the number of reads that covered a genomic position.

Sequence variant discovery We followed the best practice guidelines recommended for variant discovery and genotyping using *GATK* (version 4.0.6) with default parameters for all commands [17, 24, 44]. First, genotype likelihoods were calculated separately for each sequenced animal using *GATK HaplotypeCaller* [44], which resulted in files in *gVCF* (genomic Variant Call Format) format for each sample [45]. The gVCF files from the 49 samples were consolidated using *GATK* GenomicsDBImport. Subsequently, *GATK GenotypeGVCFs* was applied to genotype polymorphic sequence variants for all samples simultaneously.

Graphtyper (version 1.3) was run in a multi-sample mode as recommended in Eggertsson et al. Eggertsson et al. [32]. Because the original implementation of *Graphtyper* is limited to the analysis of the human chromosome complement, we cloned the *Graphtyper GitHub* repository (https://github.com/DecodeGenetics/graphtyper), modified the source code to allow analysis of the cattle chromosome complement, and compiled the program from the modified source code (see Additional file 2.1). The *Graphtyper* workflow consisted of four steps that were executed successively. First, sequence variants were identified from the read alignments that were produced using *BWA mem* (see

above). Second, these cohort-specific variants were used to augment the UMD3.1 reference genome and construct the variation-aware genome graph. Third, the sequencing reads were locally realigned against the variation-aware graph. A clean variation graph was produced by removing unobserved haplotypes paths from the raw graph. In the final step, genotypes were identified from the realigned reads in the clean graph. The *Graphtyper* pipeline was run in segments of 1 million bp and whenever the program failed to genotype variants for a particular segment either because it ran out of memory or exceeded the allocated runtime of 12 h, the interval was subdivided into smaller segments (10 kb).

Our implementation of *SAMtools mpileup* (version 1.8) [46] was run in a multi-sample mode to calculate genotype likelihoods from the aligned reads for all samples simultaneously. The parameters -E and -t were used to recalculate (and apply) base alignment quality and produce per-sample genotype annotations, respectively. Next, the estimated genotype likelihoods were converted into genotypes using *BCFtools call* using the -v and -m flags to output variable sites only, and permit sites to have more than two alternative alleles, respectively.

We implemented all pipelines using Snakemake (version 5.2.0) [47]. The scripts for the pipelines are available via *Github* repository
https://github.com/danangcrysnanto/Graph-genotyping-paper-pipelines

Sequence variant filtering and genotype refinement The *GATK VariantFiltration* module was used to parse and filter the raw VCF files. Quality control on the raw sequencing variants and genotypes was applied according to guidelines that were recommended for each variant caller. Variants that were identified using *GATK* were retained if they met the following criteria: QualByDepth (QD) $>$ 2.0, FisherStrand $>$ 60.0, RMSMappingQuality (MQ) $>$ 40.0, MappingQualityRankSumTest (MQRankSum) $>$ 12.5, ReadPosRankSumTest (ReadPosRankSum) $>$ -8.0, SOR $<$ 3.0 (SNPs) and QD $>$ 2.0, FS $<$ 200.0, ReadPosRankSum $>$ 20.0, SOR $<$ 10.0 (indels). For the variants identified using *SAMtools*, the thresholds that have been applied in the 1000 Bull Genomes project [5] were considered to remove variants with indication of low quality. Variants were retained if they met the following criteria: QUAL $>$ 20, MQ $>$ 30, ReadDepth (DP) $>$ 10, DP $<$ median(DP) $+ 3 *$ mean(DP). Moreover, SNPs were removed from the data if they had the same positions as the starting position of an indel. The output of *Graphtyper* was filtered so that it included only variants that met criteria recommeded by Eggertsson et al. Eggertsson et al. [32]: ABHet $<$ 0.0 | ABHet $>$ 0.33, ABHom $<$ 0.0 | ABHom $>$ 0.97, MaxAASR $>$ 0.4, and MQ $>$ 30.

We used *Beagle* (version 4.1) [48] to improve the raw sequence variant genotype quality and impute missing genotypes. The genotype likelihood (*gl*) mode of *Beagle* was applied to infer missing and modify existing genotypes based on the phred-scaled likelihoods (*PL*) of all other non-missing genotypes of the 49 Original Braunvieh animals in our study.

Evaluation of sequence variant genotyping To ensure consistent variant representation across the different sequence variant genotyping methods evaluated, we applied the *vt normalize* software (version 0.5) [49]. Normalized variants are parsimonious (i.e., represented by as few nucleotides as possible) and left aligned [49]. The number of variants detected and transition to transversion (Ti/Tv) ratios were calculated using *vt peek* [49] and *BCFtools stats* [46]. The intersection of variants that were common to the evaluated tools was calculated and visualized using *BCFtools isec* [46] and the UpSet R package [50], respectively.

Mendelian inconsistencies were calculated as the proportion of variants showing opposing homozygous genotypes in nine parent–offspring pairs that were included in the 49 sequenced animals. For this comparison, we considered only the sites for which the genotypes of both sire and son were not missing.

All 49 sequenced cattle were also genotyped using either the Illumina BovineHD (N = 29) or the BovineSNP50 (N = 20) Bead chip that comprise 777,962 and 54,001 SNPs, respectively. The average genotyping rate at autosomal SNPs was 98.91%. In order to assess the quality of sequence variant genotyping, the genotypes detected by the different variant calling methods were compared to the array-called genotypes in terms of genotype concordance, non-reference sensitivity and non-reference discrepancy [24, 51], and for more details on the metrics (see Additional file 2.2). Non-parametric Kruskal–Wallis tests followed by pairwise Wilcoxon signed-rank tests were applied to determine if any of the three metrics differed significantly between the three tools evaluated.

Computing environment and statistical analysis All computations were performed on the ETH Zurich Leonhard Open Cluster with access to multiple nodes equipped with 18 cores Intel Xeon E5-2697v4 processors (base frequency rated at 2.3 GHz) and 128 GB of random-access memory. Unless otherwise stated, the R (version 3.3.3) software environment [52] was used for statistical analyses and ggplot2 (version 3.0.0) [53] was used for data visualisation.

2.3 Results

Following quality control (removal of adapter sequences and low-quality bases), we aligned more than 13 billion paired-end reads (2 × 125 and 2 × 150 bp) from 49 Original Braunvieh cattle to the UMD3.1 assembly of the bovine genome. On average, 98.44% (91.06–99.59%) of the reads mapped to the reference genome and 4.26% (2.0–10.91%) of these were flagged as duplicates and not considered for further analyses. Sequencing depth ranged from 6.00 to 37.78 with an average depth per animal of 12.75 and was above 12-fold for 31 samples. Although the realignment of sequencing reads around indels is no longer required when sequence variants are genotyped using the latest version of *GATK* (v 4), it is still recommended to improve the genotyping of indels by using *SAMtools*. To ensure a fair comparison of the three tools evaluated, we realigned the reads around indels on all BAM files and used the re-aligned files as a starting point for our comparisons (Fig. 2.1). The sequencing read data of 49 cattle were deposited at European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena) under primary accession PRJEB28191.

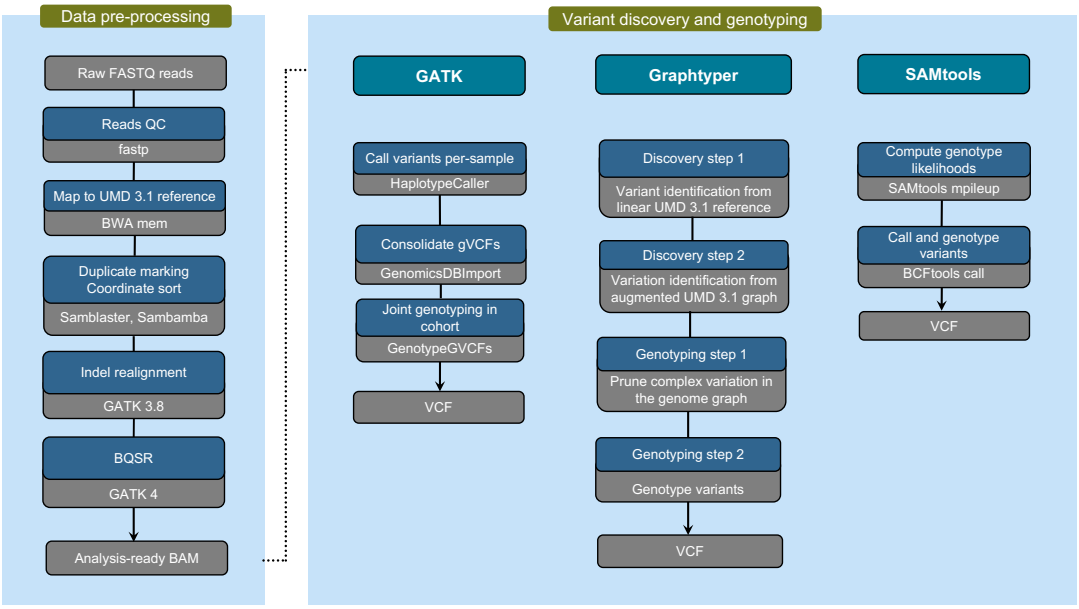


Figure 2.1: **Schematic representation of the three sequence variant discovery and genotyping methods evaluated.**
According to the best practice recommendations for sequence variant discovery using *GATK*, the VQSR module should be applied to distinguish between true and false positive variants. Because this approach requires a truth set of variants, which is not (publicly) available for cattle, the VQSR module was not considered in our evaluation

Sequence variant discovery and genotyping

Polymorphic sites (SNPs, indels) were discovered and genotyped in the 49 animals using either *GATK* (version 4), *Graphtyper* (version 1.3) or *SAMtools* (version 1.8). All software programs were run using default parameters and workflow descriptions for variant discovery (Fig. 2.1 and also see Methods). Only autosomal sequence variants were considered to evaluate the accuracy and sensitivity of sequence variant genotyping. Because variant filtering has a strong impact on the accuracy and sensitivity of sequence variant genotyping [54, 55], we evaluated both the raw variants that were detected using default parameters for variant discovery (Fig. 2.1) and variants that remained after applying filtering criteria that are commonly used but may differ slightly between different software tools. Note that *GATK* was run by using the suggested filtering parameters, when application of Variant Quality Score Recalibration (*VQSR*) is not possible.

Using default parameters for variant discovery for each of the software programs evaluated, 21,140,196, 20,262,913, and 20,668,459 polymorphic sites were discovered using *GATK*, *Graphtyper* and *SAMtools*, respectively (Table 2.1). The vast majority (86.79, 89.42 and 85.11%) of the detected variants were biallelic SNPs. Of the 18,594,182, 18,120,724 and 17,592,038 SNPs detected using *GATK*, *Graphtyper* and *SAMtools*, respectively, 7.46, 8.31 and 5.02% were novel, i.e., they were not among the 102,091,847 polymorphic sites of the most recent version (150) of the Bovine dbSNP database. The Ti/Tv ratio of the detected SNPs was equal to 2.09, 2.07 and 2.05 using *GATK*, *Graphtyper* and *SAMtool*s, respectively. Using *GATK* revealed four times more multiallelic SNPs (246,220) than either *SAMtools* or *Graphtyper*.

Table 2.1: **Number of different types of autosomal sequence variants** detected in 49 Original Braunvieh cattle using three sequence variant genotyping methods (Full) and subsequent variant filtration based on commonly used criteria (Filtered).

	Full			Filtered		
	GATK	*Graphtyper*	*SAMtools*	*GATK*	*Graphtyper*	*SAMtools*
Variants	21,140,196	20,262,913	20,668,459	19,761,679	17,679,155	18,871,549
SNPs	18,594,182	18,120,724	17,592,038	17,248,593	15,777,446	16,272,917
Not in dbSNP	1,387,781	1,505,586	882,575	867,838	564,326	570,901
Biallelic	18,347,962	18,053,396	17,528,249	17,111,806	15,730,153	16,218,714
Multi-allelic	246,220	67,328	63,789	136,787	47,293	54,203
Ti/Tv ratio	2.09	2.07	2.05	2.17	2.18	2.16
SNP array (%)						
BovineHD	99.46	99.61	99.32	99.21	98.79	98.85
Bovine SNP50	99.14	99.26	99.12	98.91	98.87	98.90
Indels	2,478,489	2,044,585	3,076,421	2,445,766	1,826,808	2,598,632
Not in dbSNP	663,831	596,137	1,279,162	639,219	456,752	979,291
Biallelic	2,166,352	1,753,391	2,704,413	2,133,840	1,571,195	2,310,386
Multi-allelic	312,137	291,194	372,008	311,926	255,613	288,246
Insertion/Deletion	0.88	0.88	1	0.88	0.88	0.99
Complex variation	67,525	97,604	0	67,320	74,901	0

We identified 2,478,489, 2,044,585, and 3,076,421 indels using *GATK*, *Graphtyper*, and *SAMtools*, respectively, and 26.78%, 29.15%, and 41.75% of them were novel. *SAMtools* revealed the largest number and highest proportion (14.9%) of indels. Between 12 and 14% of the detected indels were multiallelic. While *Graphtyper* and *GATK* identified more (12%) deletions than insertions, the proportions were almost the same using *SAMtools*.

On average, each Original Braunvieh cattle carried between 7 and 8 million variants that differed from the UMD3.1 reference genome. Of these, between 2.4 and 2.6 million SNPs were homozygous for the alternate allele, between 3.8 and 4.7 million SNPs were heterozygous and between 0.7 and 1 million were indels (Table 2.2). An intersection of 15,901,526 biallelic SNPs was common to all sequence-variant discovery tools evaluated Fig 2.2a, i.e., between 85.51 and 90.39% of the detected SNPs of each tool, and 466,029 (2.93%, Ti/Tv: 1.81) of them were novel, i.e., they were not present in dbSNP 150. The Ti/Tv-ratio of the common SNPs was 2.22. *SAMtools* had the largest number of SNPs in common with the other two tools (90.39%). The number of private SNPs, i.e., SNPs that were detected by one but not the other tools was largest for *GATK* and smallest for *Graphtyper*.

An intersection of 15,901,526 biallelic SNPs was common to all sequence-variant discovery tools evaluated (Fig. 2.2), i.e., between 85.51 and 90.39% of the detected SNPs of each tool, and 466,029 (2.93%, Ti/Tv: 1.81) of them were novel, i.e., they were not present in dbSNP 150. The Ti/Tv-ratio of the common SNPs was 2.22. *SAMtools* had the largest number of SNPs in common with the other two tools (90.39%). The number of private SNPs, i.e., SNPs that were detected by one but not the other tools was largest for *GATK* and smallest for *Graphtyper*.

In total, 1,299,467 biallelic indels Fig. 2.2b were common to all evaluated tools and 98,931 (13.13%) of these were novel, i.e., they were not present in dbSNP 150. The intersection among the three tools was considerably smaller for indels than for SNPs. *Graphtyper* had the highest proportion of indels in common with the other tools (74.11%). *SAMtools* discovered the largest number (2,704,413) of biallelic indels and most of them (41.26%) were not detected using either *GATK* or *Graphtyper*. *GATK* (21.2%) and *Graphtyper* (12.38%) discovered fewer private indels than *SAMtools*.

Table 2.2: **Average number of autosomal variants** identified per animal using three sequence variant genotyping methods

	Full			Filtered		
	GATK	*Graphtyper*	*SAMtools*	*GATK*	*Graphtyper*	*SAMtools*
Total biallelic SNPs	6,324,455	7,384,058	6,617,948	6,105,674	6,533,711	6,564,229
Heterozygous	3,890,351	4,758,297	4,187,882	3,744,336	4,074,011	4,147,033
Homozygous ALT	2,434,104	2,625,761	2,430,066	2,361,338	2,459,700	2,417,196
Ti/Tv	2.17	2.13	2.11	2.2	2.14	2.13
Total biallelic indels	693,697	767,261	1,007,420	691,765	697,637	960,218
Heterozygous	390,495 s	441,172	616,981	388,622	391,856	593,417
Homozygous ALT	303,202	326,089	390,439	303,143	305,781	366,801
Singletons	49,166	23,406	32,810	41,408	17,999	32,398

The number of variants is presented for the three tools evaluated before (Full) and after (Filtered) applying recommended filters to identify and exclude low quality variants

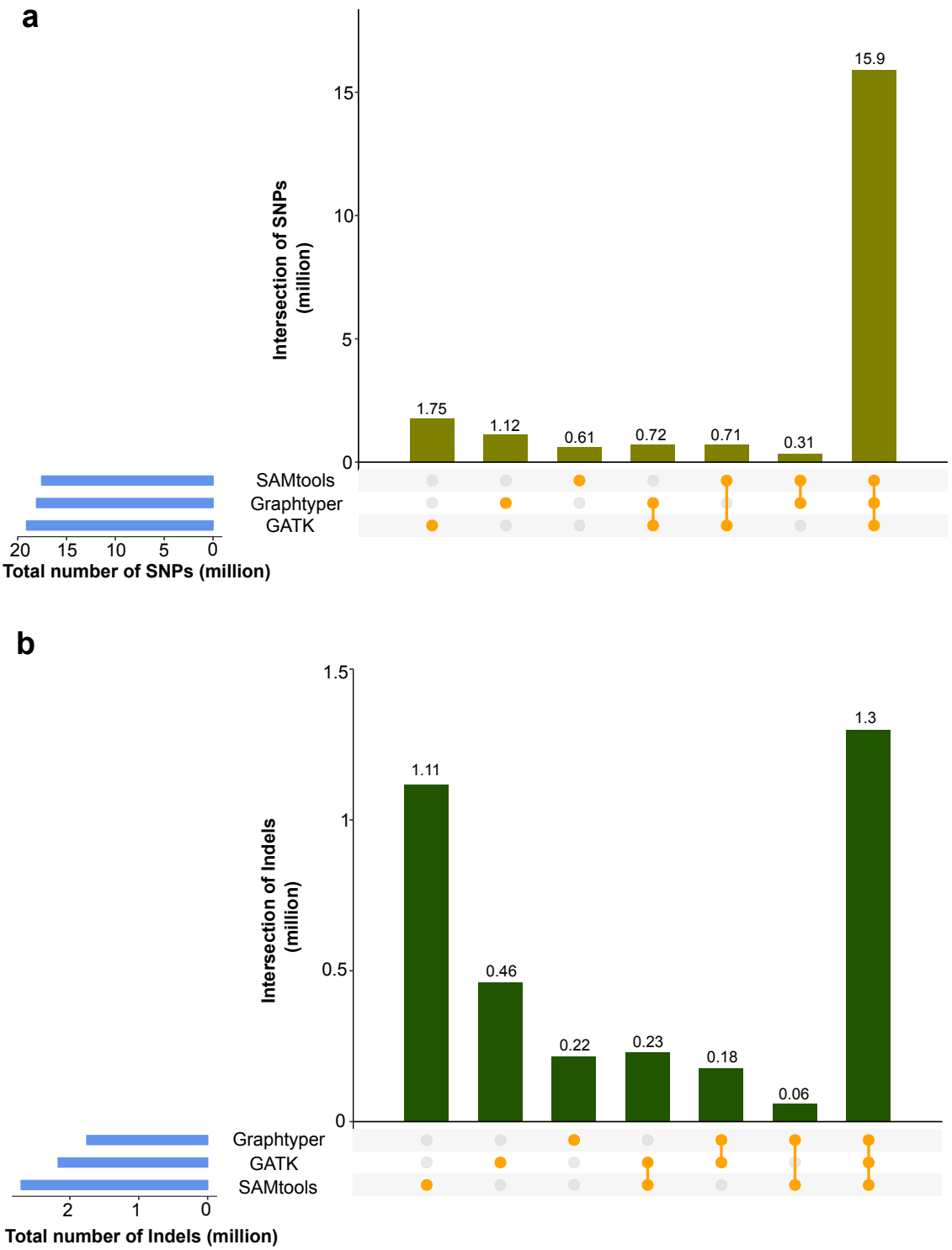


Figure 2.2: **Number of biallelic SNPs (a) and indels (b)** identified in 49 Original Braunvieh cattle using three sequence variant genotyping methods. Blue horizontal bars represent the total number of sites discovered for each method. Vertical bars indicate private and common variants detected by the methods evaluated

Sequence variant genotyping using *Graphtyper* is accurate

The 49 sequenced animals were also genotyped using either the Illumina BovineHD or the Illumina BovineSNP50 Bead chip. Genotype concordance, non-reference sensitivity and non-reference discrepancy were calculated using array-called and sequence variant genotypes at corresponding positions. Genotype concordance is a measure of the proportion of variants that have identical genotypes on the microarray and in whole-genome sequencing data. Non-reference sensitivity is the proportion of microarray-derived variants that were also detected in the sequencing data. Non-reference discrepancy reflects the proportion of sequence variants that have genotypes that differ from the microarray-derived genotypes [for more details on how the different metrics were calculated (see Additional file 2.2)]. All metrics were calculated both for raw and filtered variants either before or after applying the algorithm implemented in the *Beagle* software for haplotype phasing and imputation.

In the raw data, the proportion of missing non-reference sites was 1.90%, 0.56%, and 0.47% using *GATK*, *Graphtyper*, and *SAMtools*, respectively. The genotype concordance between the sequence- and microarray-derived genotypes was higher ($P < 0.005$) when *Graphtyper* (97.72%) was used than when either *SAMtools* (97.68%) or *GATK* (95.99%) was used (Table 2.3). For the three tools evaluated, the genotype concordance was higher at homozygous than heterozygous sites, particularly in animals that were sequenced at low depth (see Additional file 2.3). In order to take the variable proportions of missing genotypes in the sequence variants into account, we calculated non-reference sensitivity and non-reference discrepancy. Non-reference sensitivity was almost identical using *Graphtyper* (98.26%) and *SAMtools* (98.21%). However, non-reference sensitivity was clearly lower using *GATK* (93.81%, $P < 0.001$). Non-reference discrepancy was lower using *Graphtyper* (3.53%) than using either *SAMtools* (3.6%, $P = 0.003$) or *GATK* (6.35%, $P < 0.001$).

Table 2.3: **Comparisons between array-called and sequence variant genotypes.**

	Genotype concordance				Non-reference sensitivity				Non-reference discrepancy			
	full		filtered		full		filtered		full		filtered	
	raw	imp	raw	imp	raw	imp	raw	imp	raw	imp	raw	imp
GATK	95.99***	99.32***	96.02***	99.39***	93.81***	*99.36*	93.67***	*99.15*	6.35***	1.05***	6.3***	0.95***
Graphtyper	*97.71*	*99.46*	*97.75*	*99.52*	*98.26*	99.35	*97.91*	99.00***	*3.53*	*0.83*	*3.47*	*0.73*
SAMtools	97.68***	99.24***	97.7*	99.29***	98.21	99.35	97.53***	98.67***	3.6**	1.17***	3.56**	1.09***

Genotype concordance, non-reference sensitivity and non-reference discrepancy (in percentage) was calculated between the genotypes from the Bovine SNP Bead chip and sequence–derived genotypes for 49 Original Braunvieh cattle considering either the raw or imputed (imp) sequence variant genotypes before (full) and after (filtered) variants were filtered based on commonly used criteria. Asterisks denote a significant difference (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$) with the best value (italic) for a respective parameter.

Table 2.4: **Proportions of opposing homozygous genotypes observed in nine sire-son pairs**

	SNPs				indels			
	full		filtered		full		filtered	
	raw	imp	raw	imp	raw	imp	raw	imp
Bovine HD SNP array	*0.001*							
GATK	0.73*	0.15*	0.72*	0.13*	0.98*	0.24*	0.99*	0.21*
Graphtyper	0.36	*0.11*	0.36	*0.11*	*0.54*	*0.13*	*0.54*	*0.13*
SAMtools	*0.33*	0.28*	*0.32*	0.25*	0.67	0.54*	0.61	0.57*

The ratio (in percentage) was calculated using autosomal sequence variants considering either the raw or imputed (imp) sequence variant genotypes before (full) and after (filtered) variants were filtered based on commonly used criteria. Asterisks denote significant differences (* $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$) with the best value (italic) for a respective parameter.

Next, we analysed the proportion of opposing homozygous genotypes for SNPs and indels in nine sire-son pairs that were included among the sequenced animals (Table 2.4). We observed that SNPs that were discovered using either *Graphtyper* or *SAMtools* had almost a similar proportion of genotypes with Mendelian inconsistencies in the full and filtered datasets, whereas the values were two times higher using *GATK*. The proportion of opposing homozygous genotypes was higher for indels than SNPs for all the tools evaluated. However, in the full and filtered datasets, it was lower when *Graphtyper* was used than when either *GATK* or *SAMtools* was used. Using filtering parameters that are commonly applied for the three evaluated tools (see Methods), we excluded 1,378,517 (6.52%, Ti/Tv 1.24), 2,583,758 (12.75%, Ti/Tv 1.47) and 1,796,910 (8.69%, Ti/Tv 1.36) variants due to low mapping or genotyping quality from the *GATK*, *Graphtyper*, and *SAMtools* datasets, respectively. The genotype concordance between sequence- and microarray-derived genotypes was slightly higher for the filtered than the raw genotypes, but the non-reference sensitivity was lower for the filtered than the raw genotypes, which indicates that the filtering step also removed some true variant sites from the raw data (Table 2.3). The filtering step had almost no effect on the proportion of Mendelian inconsistencies detected in the nine sire-son pairs (Table 2.4).

Beagle genotype refinement improved genotype quality

We used the *Beagle* software to refine the primary genotype calls and infer missing genotypes in the raw and filtered datasets. Following imputation, the quality of the sequence variant genotypes increased for all evaluated tools particularly for the individuals that had a sequencing coverage less than 12-fold (Fig. 2.3). The largest increase in the concordance metrics was observed for the sequence variants that were obtained using *GATK*

(Tables 2.3 and 2.4). Following imputation, the variants identified using *Graphtyper* had a significantly higher quality ($P < 0.05$) for eight out of the ten metrics evaluated.

The quality of the sequence variant genotypes, particularly before *Beagle* genotype phasing and imputation, was influenced by the variable depth of coverage for the 49 sequenced samples of our study (Fig. 2.3). When we restricted the evaluations to 31 samples that had an average sequencing depth above 12-fold, the three tools performed almost identically (see Additional file 2.4). However, the performance of *Graphtyper* was significantly ($P < 0.05$) higher for 12 (out of the total 20) metrics than either that of *GATK* or *SAMtools*. When 18 samples with an average sequencing depth lower than 12-fold were considered, the differences observed in the three metrics were more pronounced between the three tools. In samples with a low sequencing coverage, *Graphtyper* performed significantly ($P < 0.05$) better than either *GATK* or *SAMtools* for all concordance metrics both before and after filtering and *Beagle* imputation, except for the non-reference sensitivity.

Computing requirements

The multi-sample sequence variant genotyping pipelines that were implemented using either *GATK* or *SAMtools* were run separately for each chromosome in a single-threading mode. The *SAMtools mpileup* module took between 3.07 and 11.4 CPU hours and required between 0.12 and 0.25 gigabytes (GB) peak random-access memory (RAM) per chromosome. To genotype 20,668,459 sequence variants in 49 animals, *SAMtools* mpileup required 192 CPU hours (Fig. 2.4).

For *GATK*, we submitted 1421 parallel jobs of the *HaplotypeCaller* module (i.e., one job for each animal and chromosome) that required between 3.9 and 12.3 GB RAM and between 0.36 and 11 CPU hours to complete. To process 29 chromosomes in 49 samples, the *HaploytpeCaller* module required 2428 CPU hours. Subsequently, we ran the *GATK GenomicsDBImport* module, which required between 7.98 and 20.88 GB RAM and between 2.81 and 19.31 CPU hours per chromosome. *GATK Joint Genotyping* required between 4.33 and 17.32 GB of RAM and between 1.81 and 14.01 h per chromosome. To genotype 21,140,196 polymorphic sequence variants in 49 animals, the *GATK* pipeline required 2792 CPU hours (Fig. 2.4).

The *Graphtyper* pipeline including construction of the variation graph and genotyping of sequence variants was run in parallel for 2538 non-overlapping segments of 1 million bp as recommended by Eggertsson et al. [32]. The peak RAM required by *Graphtyper* was between 1 and 3 GB per segment. Twelve segments, for which *Graphtyper* either ran

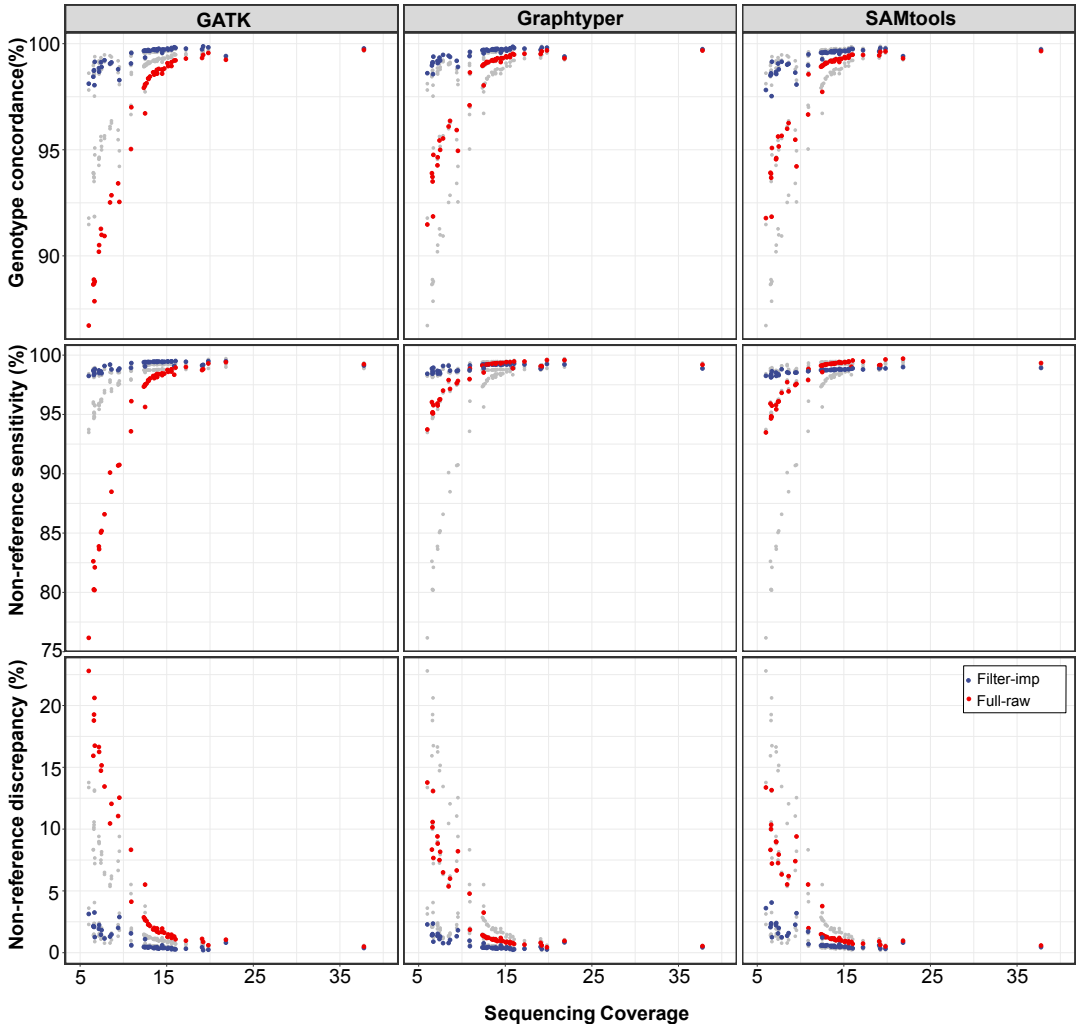


Figure 2.3: **Accuracy and sensitivity of sequence variant genotyping at different sequencing depths.** Genotype concordance, non-reference sensitivity and non-reference discrepancy were calculated for 49 Original Braunvieh cattle considering either raw (red) or filtered and imputed (blue) sequence variant genotypes. The grey points represent overlays of the results from the other methods

out of memory or did not finish within the allocated time, were subdivided into smaller segments of 10 kb and subsequently re-run (Additional file 2.5). The genotyping of 20,262,913 polymorphic sites in 49 animals using our implementation of the *Graphtyper* pipeline required 1066 CPU hours (Fig. 2.4).

The computing resources required by *SAMtools* and *GATK* increased linearly with chromosome length. The computing time required to genotype sequence variants was highly heterogeneous along the genome using *Graphtyper*. The CPU time for a 1-Mb segment ranged from 0.196 to 10.11 h, with an average CPU time of 0.42 h. We suspected

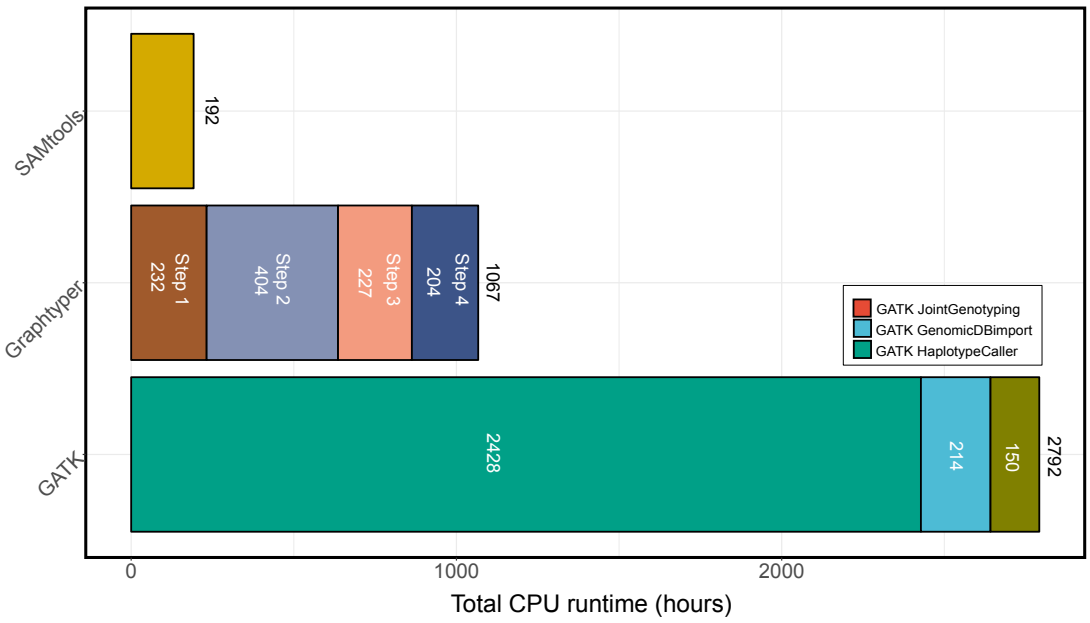


Figure 2.4: **Computing time required to genotype all autosomal sequence variants in 49 Original Braunvieh cattle.** The runtime of *GATK* and *Graphtyper* is shown for the different steps (see Fig. 2.1 for more details)

that flaws in the reference genome might increase the complexity of the variation-aware graph and that the construction of the graph might benefit from an improved assembly. To test this hypothesis, we re-mapped the sequencing reads to the recently released new bovine reference genome (ARS-UCD1.2, https://www.ncbi.nlm.nih.gov/assembly/GCF_002263795.1) and repeated the graph-based sequence variant discovery. Indeed, we did observe a decrease in the computing time required to genotype polymorphic sites (particularly at chromosomes 12, 27 and 29) and a more uniform runtime along the genome, which possibly indicates that graph-based variant discovery in cattle will be faster and more accurate with highly contiguous reference sequences (Fig. 2.5).

2.4 Discussion

We used either *GATK*, *Graphtyper*, or *SAMtools* to discover and genotype polymorphic sequence variants in whole-genome sequencing data of 49 Original Braunvieh cattle that were sequenced at between 6 and 38-fold genome coverage. Whereas *SAMtools* and *GATK* discover variants from a linear reference genome, *Graphtyper* locally realigns reads to a variation-aware reference graph that incorporates cohort-specific sequence variants [32]. Our graph-based variant discovery pipeline that is implemented by using

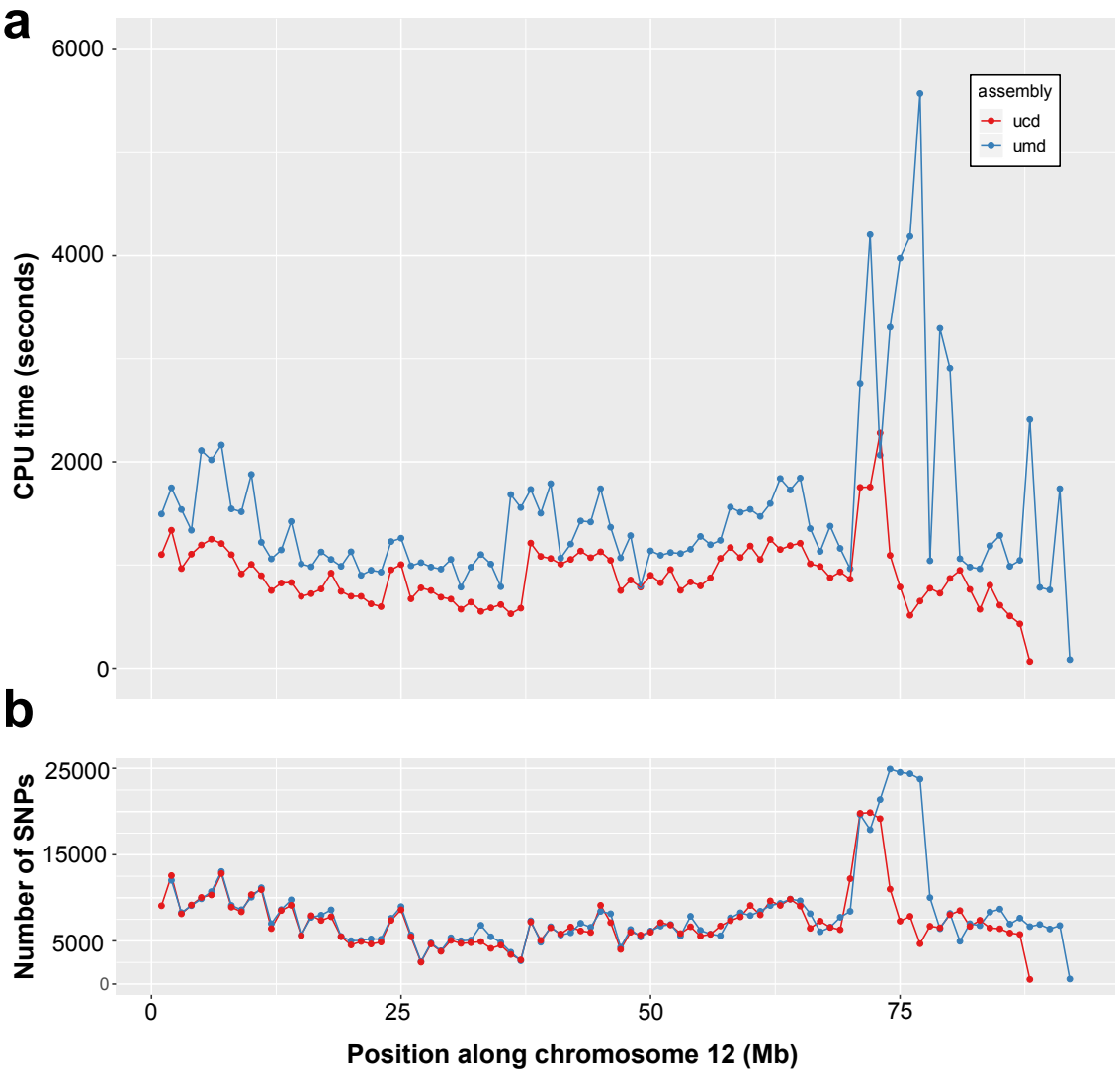


Figure 2.5: **Sequence variant genotyping on chromosome 12 using *Graphtyper*.** Computing time required (**a**) and number of variants discovered (**b**) for bovine chromosome 12 using *Graphtyper*. Each dot represents an interval of 1 million bp. Blue and red colours represent values for the UMD3.1 and ARS-UCD1.2 versions of the bovine assembly, respectively

the *Graphtyper* software used the existing bovine reference sequence to construct the genome graph. Subsequently, the graph was augmented with variants that were detected from linear alignments of the 49 Original Braunvieh cattle. The use of more sophisticated genome graph-based approaches that have been developed very recently facilitates the mapping of raw sequencing reads directly against a genome graph without the need to first align reads towards a linear reference genome [34]. Whereas genome graph-based variant discovery has been explored recently in mammalian-sized genomes [31, 34, 35**?**], our work is the first to apply graph-based sequence variant genotyping in cattle.

In order to evaluate graph-based variant discovery in cattle, we compared accuracy and sensitivity of *Graphtyper* to *GATK*, and *SAMtools* , i.e., two state-of-the-art methods on linear reference genomes that have been evaluated thoroughly in many species including cattle [4, 22]. We ran each tool with default parameters for variant discovery and applied commonly used or recommended filtration criteria. However, our evaluation of the software tools may suffer from ascertainment bias because we relied on SNPs that are included in bovine SNP arrays, i.e., they are located predominantly at genomic regions where variants are easy to identify [37, 38, 51]. Thus, the global accuracy and sensitivity of sequence variant discovery might be overestimated in our study. However, this ascertainment bias is unlikely to affect the relative performance of the methods evaluated.

In 49 Original Braunvieh cattle, sequence variant genotyping was more accurate using *Graphtyper* than either *GATK* or *SAMtools*. Differences in accuracy are small between the three tools, particularly when samples are sequenced at an average coverage higher than 12-fold (see Additional file 2.4). Yet, *Graphtyper* performed significantly better than *GATK* and *SAMtools* for samples sequenced at medium ($>$ 12-fold) or low($<$ 12-fold) coverage indicating that genome graph-based variant discovery in cattle is accurate across a wide range of sequencing depths. *GATK* might perform better than observed in our study, when the VQSR module is applied to train the variant filtration algorithm on true and false variants [56]. However, to the best of our knowledge, the required sets of true and false variants are not available in cattle. An intersection of variants detected by different sequence variant genotyping software may be considered as a truth set (e.g., Alberto et al. [57]) and compiling such a set is possible using the 49 samples from our study. However, a truth set that has been constructed from the data that are used for evaluation is likely to be depleted for variants that are difficult to discover in the target data set, thus preventing an unbiased evaluation of variant calling [36]. Variants from the 1000 Bull Genomes project [5, 6] could potentially serve as a truth/training set. However, variants from the 1000 Bull Genomes project were detected from short read sequencing data using either *GATK* or *SAMtools*, i.e., technologies and software that are part of our evaluation, thus precluding an unbiased comparison of variant discovery between *GATK*, *Graphtyper*, and *SAMtools* [36]. Vander Jagt et al. [44] showed in a subset of samples from the 1000 Bull Genomes project that *GATK VQSR* does not notably improve the concordance between sequence-derived and microarray-called genotypes compared to *GATK* hard filtering. Interestingly, the proportion of opposing homozygous genotypes in sire/offspring pairs was slightly higher in their study using *GATK VQSR* than *GATK* hard-filtering as used by the 1000 Bull Genomes project [44]. Applying *GATK* VQSR to the variants of our dataset corroborates the findings of Vander Jagt et al. [44] (see Additional file 2.6). Considering that the quality of the truth/training sets has a

strong impact on the capabilities of VQSR (Additional file 2.6) and that high-confidence variants are currently not publicly available for cattle, we report *GATK* results using the recommended filtering parameters when VQSR is not possible.

Regardless of the method evaluated, we observed heterozygous under-calling in animals that are sequenced at low coverage, i.e., heterozygous variants were erroneously genotyped as homozygous due to an insufficient number of sequencing reads supporting the heterozygous genotype [10, 58, 59, 60]. In agreement with previous studies [4, 5], *Beagle* imputation improved genotype concordance and reduced heterozygous under-calling particularly in individuals that are sequenced at low coverage. After the imputation step, the genotype concordance, non-reference sensitivity, and non-reference discrepancy of the three tools were almost identical, which indicates that genotyping sequence variants from samples with a medium genome coverage is possible at high accuracy (at least for common variants in more accessible regions of the genome) using any of the three tools evaluated and subsequent *Beagle* error correction. While such conclusions have been drawn previously for *SAMtools* and *GATK* [4, 22], our findings demonstrate that the genotype likelihoods estimated from the *Graphtyper* software are also compatible with and benefit from the imputation algorithm implemented in the *Beagle* software. Considering that sequence data are enriched for rare variants that are more difficult to impute than common variants from SNP microarrays [61], the benefits from *Beagle* error correction might be overestimated in our study. An integration of phasing and imputation of missing genotypes directly in a graph-based variant genotyping approach would simplify sequence variant genotyping from variation-aware graphs [31, 62, 63]. Using *Graphtyper* for variant genotyping and *Beagle* for genotype refinement enabled us to genotype sequence variants in 49 Original Braunvieh cattle at a genotypic concordance of 99.52%, i.e., higher than previously achieved using either *GATK* or *SAMtools* for variant calling in cattle that are sequenced at a similar genome coverage [2, 3, 4, 5, 22, 64]; this indicates that graph-based variant discovery might improve sequence variant genotyping. However, applying the filtering criteria that are recommended for *Graphtyper* [32] removed more variants from the *Graphtyper* (12.75%) than from either *GATK* (6.52%) or *SAMtools* (8.69%) datasets. It should be mentioned that *GATK* VQSR would remove considerably more variants from the *GATK* dataset than *GATK* hard filtering as applied in our study (see Additional file 2.6). Fine-tuning of the variant filtering parameters is necessary to further increase the accuracy and sensitivity of sequencing variant genotyping, particularly for *Graphtyper* [54, 55]. Moreover, the accuracy and sensitivity of graph-based variant discovery may be higher when known variants are considered for the initial construction of the variation graph [32]. Indeed, we observed a slight increase in genotype concordance (see Additional file 2.7) when we used *Graphtyper* to genotype sequence variants from a variation-aware genome-graph

that incorporated bovine variants listed in dbSNP 150. However, additional research is required to prioritize a set of variants to augment bovine genome graphs for different cattle breeds [65].

Using microarray-derived genotypes as a truth set may overestimate the accuracy of sequence variant discovery particularly at variants that are rare or located in less accessible regions of the genome. Moreover, it does not allow assessment of the accuracy and sensitivity of indel discovery because variants other than SNPs are currently not routinely genotyped with commercially available microarrays. Estimating the proportion of opposing homozygous genotypes between parent–offspring pairs may be a useful diagnostic metric to detect sequencing artefacts or flawed genotypes at indels [66]. Our results show that genotypes at indels are more accurate using *Graphtyper* than either *SAMtools* or *GATK* because *Graphtyper* produced less opposing homozygous genotypes at indels in nine sire-son pairs than the other methods both in the raw and filtered datasets. These findings are in line with those reported by Eggertsson et al. [32], who showed that the mapping of the sequencing reads to a variation-aware graph could improve read alignment nearby indels, thus enabling highly accurate sequence variant genotyping also for variants other than SNPs. Recently, Garrison et al. [34] showed that graph-based variant discovery may also mitigate reference allele bias. An assessment of reference allele bias was, however, not possible in our study because the sequencing depth was too low for most samples.

In our study, *Graphtyper* required less computing time than *GATK* to genotype sequence variants for 49 individuals. *SAMtools* required the least computing resources, probably because the implemented mpileup algorithm produces genotypes from the aligned reads without performing the computationally intensive local realignment of the reads. However, with an increasing number of samples, the multi-sample variant genotyping implementation of the *GATK HaplotypeCaller* module seems to be more efficient than *SAMtools mpileup* because variant discovery within samples can be separated from the joint genotyping across samples [19, 44]. A highly parallelized graph-based variant discovery pipeline also offers a computationally feasible and scalable framework for variant discovery in thousands of samples [32]. However, the computing time necessary for graph-based variant genotyping might be high in genomic regions where the nucleotide diversity is high or the assembly is flawed [35, 67]. In our study, the algorithm implemented in the *Graphtyper* software failed to finish within the allocated time for 12 1-Mb segments including a segment on chromosome 12 that contains a large segmental duplication [61, 68, 69] possibly because many mis-mapped reads increased graph complexity. The region on chromosome 12 contains an unusually large number of sequence variants and has been shown to suffer from low accuracy of imputation [61].

Graphtyper also failed to finish within the allocated time for a region on chromosome 23 that encompasses the bovine major histocompatibility complex, which is known to have a high level of diversity. Our results show that *Graphtyper* may also produce genotypes for problematic segments when they are split and processed in smaller parts. Moreover, most of these problems disappeared when we considered the latest assembly of the bovine genome, which possibly corroborates that more complete and contiguous genome assemblies may facilitate more reliable genotyping from variation-aware graphs [37, 70].

2.5 Conclusions

Genome graphs facilitate sequence variant discovery from non-linear reference genomes. Sequence variant genotyping from a variation-aware graph is possible in cattle using *Graphtyper*. Sequence variant genotyping at both SNPs and indels is more accurate and sensitive using *Graphtyper* than either *SAMtools* or *GATK*. The proportion of Mendelian inconsistencies at both SNPs and indels is low using *Graphtyper*, which indicates that sequence variant genotyping from a variation-aware genome graph facilitates accurate variant discovery at different types of genetic variation. Considering highly informative variation-aware genome graphs that have been constructed from multiple breed-specific de-novo assemblies and high-confidence sequence variants may facilitate more accurate, sensitive and unbiased sequence variant genotyping in cattle.

References

[1] Jesse L. Hoff, Jared E. Decker, Robert D. Schnabel, and Jeremy F. Taylor. Candidate lethal haplotypes and causal mutations in Angus cattle. *BMC Genomics*, 18(1), 2017.

[2] Paul Stothard, Xiaoping Liao, Adriano S Arantes, Mary De Pauw, Colin Coros, Graham S Plastow, Mehdi Sargolzaei, John J Crowley, John A Basarab, Flavio Schenkel, Stephen Moore, and Stephen P Miller. A large and diverse collection of bovine genome sequences from the Canadian Cattle Genome Project. *GigaScience*, 2015.

[3] Mekki Boussaha, Pauline Michot, Rabia Letaief, Chris Hozé, et al. Construction of a large collection of small genome variations in French dairy and beef breeds using whole-genome sequences. *Genetics Selection Evolution*, 48(1):87, dec 2016.

[4] Sandra Jansen, Bernhard Aigner, Hubert Pausch, Michal Wysocki, Sebastian Eck, Anna Benet-Pagès, Elisabeth Graf, Thomas Wieland, Tim M. Strom, Thomas Meitinger, and Ruedi Fries. Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. *BMC Genomics*, 14(1):446, jul 2013.

[5] Hans D Daetwyler, Aurélien Capitan, Hubert Pausch, Paul Stothard, Rianne Van Binsbergen, Rasmus F Brøndum, Xiaoping Liao, Anis Djari, Sabrina C Rodriguez, Cécile Grohs, Diane Esquerré, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genetics*, 46(8):858–865, aug 2014.

[6] Ben J. Hayes and Hans D. Daetwyler. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annual Review of Animal Biosciences*, 7(1):annurev–animal–020518–115024, feb 2019.

[7] Hubert Pausch, Reiner Emmerling, Birgit Gredler-Grandl, Ruedi Fries, Hans D. Daetwyler, and Michael E. Goddard. Meta-analysis of sequence-based association studies across three cattle breeds reveals 25 QTL for fat and protein percentages in milk at nucleotide resolution. *BMC Genomics*, 18(1): 853, dec 2017.

[8] Aniek C. Bouwman, Hans D. Daetwyler, Amanda J. Chamberlain, Carla Hurtado Ponce, Mehdi Sargolzaei, Flavio S. Schenkel, Goutam Sahana, Armelle Govignon-Gion, Simon Boitard, Marlies Dolezal, Hubert Pausch, Rasmus F. Brøndum, Phil J. Bowman, Bo Thomsen, Bernt Guldbrandtsen, Mogens S. Lund, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nature Genetics*, 50(3):362–367, feb 2018.

[9] Biaty Raymond, Aniek C. Bouwman, Chris Schrooten, Jeanine Houwing-Duistermaat, and Roel F. Veerkamp. Utility of whole-genome sequence data for across-breed genomic prediction. *Genetics Selection Evolution*, 50(1):27, dec 2018.

[10] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, 2011.

[11] Yan Guo, Fei Ye, Quanghu Sheng, Travis Clark, and David C Samuels. Three-stage quality control strategies for DNA re-sequencing data. *Briefings in bioinformatics*, 15(6):879–889, 2014.

[12] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333, 2016.

[13] SP Pfeifer. From next-generation resequencing reads to a high-quality variant data set. *Heredity*, 118 (2):111–124, 2017.

[14] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9 (4):357, 2012.

[15] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.

[16] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25 (16):2078–2079, 2009.

[17] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[18] Geraldine A Van der Auwera, Mauricio O Carneiro, Christopher Hartl, Ryan Poplin, Guillermo Del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, 43(1):11–10, 2013.

[19] Ryan Poplin, Valentin Ruano-Rubio, Mark A DePristo, Tim J Fennell, Mauricio O Carneiro, Geraldine A Van der Auwera, David E Kling, Laura D Gauthier, Ami Levy-Moonshine, David Roazen, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, page 201178, 2018.

[20] Xiangtao Liu, Shizhong Han, Zuoheng Wang, Joel Gelernter, and Bao-Zhu Yang. Variant callers for next-generation sequencing data: a comparison study. *PloS one*, 8(9), 2013.

[21] Anthony Youzhi Cheng, Yik-Ying Teo, and Rick Twee-Hee Ong. Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics*, 30(12):1707–1713, 2014.

[22] Christine F Baes, Marlies A Dolezal, James E Koltes, Beat Bapst, Eric Fritz-Waters, Sandra Jansen, Christine Flury, Heidi Signer-Hasler, Christian Stricker, Rohan Fernando, et al. Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. *BMC genomics*, 15(1):948, 2014.

[23] Pankaj Kumar, Mashael Al-Shafai, Wadha Ahmed Al Muftah, Nader Chalhoub, Mahmoud F Elsaid, Alice Abdel Aleem, and Karsten Suhre. Evaluation of SNP calling using single and multiple-sample calling algorithms by validation against array base genotyping and Mendelian inheritance. *BMC research notes*, 7(1):747, 2014.

[24] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43 (5):491, 2011.

[25] Jacob F Degner, John C Marioni, Athma A Pai, Joseph K Pickrell, Everlyne Nkadori, Yoav Gilad, and Jonathan K Pritchard. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–3212, 2009.

[26] Débora Y. C. Brandt, Vitor R. C. Aguiar, Bárbara D. Bitarello, Kelly Nunes, Jérôme Goudet, and Diogo Meyer. Mapping bias overestimates reference allele frequencies at the hla genes in the 1000 genomes project phase i data. *G3: Genes, Genomes, Genetics*, 5(5):931–941, 2015.

[27] Alexander Dilthey, Charles Cox, Zamin Iqbal, Matthew R Nelson, and Gil McVean. Improved genome inference in the MHC using a population reference graph. *Nature Genetics*, 47(6):682–688, 2015.

[28] Khalid A Fakhro, Michelle R Staudt, Monica Denise Ramstetter, Amal Robay, et al. The Qatar genome: a population-specific tool for precision medicine in the Middle East. *Human Genome Variation*, 3(1): 16016, 2016.

[29] Lingling Shi, Yunfei Guo, Chengliang Dong, John Huddleston, et al. Long-read sequencing and de novo assembly of a Chinese genome. *Nature Communications*, 7(1):12065, 2016.

[30] Adam Ameur, Huiwen Che, Marcel Martin, Ignas Bunikis, 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, oct 2018.

[31] Goran Rakocevic, Vladimir Semenyuk, Wan-Ping Lee, James Spencer, John Browning, Ivan J Johnson, Vladan Arsenijevic, Jelena Nadj, Kaushik Ghose, Maria C Suciu, et al. Fast and accurate genomic analyses using genome graphs. *Nature genetics*, 51(2):354–362, 2019.

[32] Hannes P Eggertsson, Hakon Jonsson, Snaedis Kristmundsdottir, Eirikur Hjartarson, Birte Kehr, Gisli Masson, Florian Zink, Kristjan E Hjorleifsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, et al. Graphtyper enables population-scale genotyping using pangenome graphs. *Nature genetics*, 49(11): 1654, 2017.

[33] Adam M Novak, Glenn Hickey, Erik Garrison, Sean Blum, Abram Connelly, Alexander Dilthey, Jordan Eizenga, MA Saleh Elmohamed, Sally Guthrie, André Kahles, et al. Genome graphs. *bioRxiv*, 2017.

[34] Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature biotechnology*, 36(9):875–879, 2018.

[35] Jonas Andreas Sibbesen, Lasse Maretty, and Anders Krogh. Accurate genotyping across variant classes and lengths using variant graphs. *Nature genetics*, 50(7):1054–1059, 2018.

[36] Heng Li, Jonathan M Bloom, Yossi Farjoun, Mark Fleharty, Laura Gauthier, Benjamin Neale, and Daniel MacArthur. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nature methods*, 15(8):595–597, 2018.

[37] Heng Li. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20):2843–2851, 2014.

[38] Dorcus Kholofelo Malomane, Christian Reimer, Steffen Weigend, Annett Weigend, Ahmad Reza Sharifi, and Henner Simianer. Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. *BMC genomics*, 19(1):22, 2018.

[39] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890, 2018.

[40] Aleksey V Zimin, Arthur L Delcher, Liliana Florea, David R Kelley, Michael C Schatz, Daniela Puiu, Finnian Hanrahan, Geo Pertea, Curtis P Van Tassell, Tad S Sonstegard, et al. A whole-genome assembly of the domestic cow, Bos taurus. *Genome biology*, 10(4):R42, 2009.

[41] Gregory G Faust and Ira M Hall. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*, 30(17):2503–2505, 2014.

[42] Artem Tarasov, Albert J Vilella, Edwin Cuppen, Isaac J Nijman, and Pjotr Prins. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, 31(12):2032–2034, 2015.

[43] Brent S Pedersen and Aaron R Quinlan. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*, 34(5):867–868, 2018.

[44] CJ Vander Jagt, AJ Chamberlain, RD Schnabel, BJ Hayes, and HD Daetwyler. Which is the best variant caller for large whole-genome sequencing datasets. In *Proceedings of the 11th world congress on genetics applied to livestock production*, pages 11–16, 2018.

[45] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011.

[46] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

[47] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[48] Brian L Browning and Sharon R Browning. Genotype imputation with millions of reference samples. *The American Journal of Human Genetics*, 98(1):116–126, 2016.

[49] Adrian Tan, Gonçalo R Abecasis, and Hyun Min Kang. Unified representation of genetic variants. *Bioinformatics*, 31(13):2202–2204, 2015.

[50] Jake R Conway, Alexander Lex, and Nils Gehlenborg. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18):2938–2940, 2017.

[51] Michael D Linderman, Tracy Brandt, Lisa Edelmann, Omar Jabado, Yumi Kasai, Ruth Kornreich, Milind Mahajan, Hardik Shah, Andrew Kasarskis, and Eric E Schadt. Analytical validation of whole exome and whole genome sequencing for clinical applications. *BMC medical genomics*, 7(1):20, 2014.

[52] Team R Core. R: A language and environment for statistical computing. 2013.

[53] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer, 2016.

[54] Andrew R Carson, Erin N Smith, Hiroko Matsui, Sigrid K Brækkan, Kristen Jepsen, John-Bjarne Hansen, and Kelly A Frazer. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC bioinformatics*, 15(1):125, 2014.

[55] Goo Jun, Mary Kate Wing, Gonçalo R Abecasis, and Hyun Min Kang. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Research*, 25(6):918–925, 2015.

[56] Mehdi Pirooznia, Melissa Kramer, Jennifer Parla, Fernando S Goes, James B Potash, W Richard McCombie, and Peter P Zandi. Validation and assessment of variant calling pipelines for next-generation sequencing. *Human genomics*, 8(1):14, 2014.

[57] Florian J Alberto, Frédéric Boyer, Pablo Orozco-terWengel, Ian Streeter, Bertrand Servin, Pierre De Villemereuil, Badr Benjelloun, Pablo Librado, Filippo Biscarini, Licia Colli, et al. Convergent genomic signatures of domestication in sheep and goats. *Nature Communications*, 9(1):1–9, 2018.

[58] David Sims, Ian Sudbery, Nicholas E Ilott, Andreas Heger, and Chris P Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121, 2014.

[59] Christopher A Fragoso, Christopher Heffelfinger, Hongyu Zhao, and Stephen L Dellaporta. Imputing genotypes in biallelic populations from low-coverage sequence data. *Genetics*, 202(2):487–495, 2016.

[60] Timothy P Bilton, John C McEwan, Shannon M Clarke, Rudiger Brauning, Tracey C van Stijn, Suzanne J Rowe, and Ken G Dodds. Linkage disequilibrium estimation in low coverage high-throughput sequencing data. *Genetics*, 209(2):389–400, 2018.

[61] Hubert Pausch, Iona M MacLeod, Ruedi Fries, Reiner Emmerling, Phil J Bowman, Hans D Daetwyler, and Michael E Goddard. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. *Genetics Selection Evolution*, 49(1):24, 2017.

[62] Jouni Sirén, Erik Garrison, Adam M Novak, Benedict Paten, and Richard Durbin. Haplotype-aware graph indexes. *Bioinformatics*, 36(2):400–407, 2020.

[63] Adam M Novak, Erik Garrison, and Benedict Paten. A graph extension of the positional Burrows–Wheeler transform and its applications. *Algorithms for Molecular Biology*, 12(1):18, 2017.

[64] Nedenia Bonvino Stafuzza, Adhemar Zerlotini, Francisco Pereira Lobo, Michel Eduardo Beleza Yamagishi, Tatiane Cristina Seleguim Chud, Alexandre Rodrigues Caetano, Danisio Prado Munari, Dorian J Garrick, Marco Antonio Machado, Marta Fonseca Martins, et al. Single nucleotide variants and InDels identified from whole-genome re-sequencing of Guzerat, Gyr, Girolando and Holstein cattle breeds. *PLoS One*, 12(3), 2017.

[65] Jacob Pritt, Nae-Chyun Chen, and Ben Langmead. FORGe: prioritizing variants for graph genomes. *Genome biology*, 19(1):1–16, 2018.

[66] Zubin Hasmukh Patel, Leah Claire Kottyan, Sara Lazaro, Marc S Williams, David H Ledbetter, Gerard Tromp, Andrew Rupert, Mojtaba Kohram, Michael Wagner, Ammar Husami, et al. The struggle to find reliable results in exome sequencing data: filtering out Mendelian errors. *Frontiers in genetics*, 5: 16, 2014.

[67] Sergey Koren, Gregory P Harhay, Timothy PL Smith, James L Bono, Dayna M Harhay, Scott D Mcvey, Diana Radune, Nicholas H Bergman, and Adam M Phillippy. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome biology*, 14(9):R101, 2013.

[68] George E Liu, Mario Ventura, Angelo Cellamare, Lin Chen, Ze Cheng, Bin Zhu, Congjun Li, Jiuzhou Song, and Evan E Eichler. Analysis of recent segmental duplications in the bovine genome. *BMC genomics*, 10(1):571, 2009.

[69] Derek M Bickhart, Yali Hou, Steven G Schroeder, Can Alkan, Maria Francesca Cardone, Lakshmi K Matukumalli, Jiuzhou Song, Robert D Schnabel, Mario Ventura, Jeremy F Taylor, et al. Copy number variation of individual cattle genomes using next-generation sequencing. *Genome research*, 22(4):778–790, 2012.

[70] Yan Guo, Yulin Dai, Hui Yu, Shilin Zhao, David C Samuels, and Yu Shyr. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, 109(2):83–90, 2017.

Chapter 3

Bovine breed-specific augmented reference graphs facilitate accurate sequence read mapping and unbiased variant discovery

Danang Crysnanto[1], Hubert Pausch[1]

[1] Animal Genomics, ETH Zurich, Zurich, Switzerland.

Published in *Genome Biology (2020) 21:184*

Contribution: I participated in conceiving the study, analysing the results and writing the manuscript. I wrote the whole-genome graph pipelines.

Abstract

Background: The current bovine genomic reference sequence was assembled from a Hereford cow. The resulting linear assembly lacks diversity because it does not contain allelic variation, a drawback of linear references that causes reference allele bias. High nucleotide diversity and the separation of individuals by hundreds of breeds make cattle ideally suited to investigate the optimal composition of variation-aware references.

Results: We augment the bovine linear reference sequence (ARS-UCD1.2) with variants filtered for allele frequency in dairy (Brown Swiss, Holstein) and dual-purpose (Fleckvieh, Original Braunvieh) cattle breeds to construct either breed-specific or pan-genome reference graphs using the *vg toolkit*. We find that read mapping is more accurate to variation-aware than linear references if pre-selected variants are used to construct the genome graphs. Graphs that contain random variants do not improve read mapping over the linear reference sequence. Breed-specific augmented and pan-genome graphs enable almost similar mapping accuracy improvements over the linear reference. We construct a whole-genome graph that contains the Hereford-based reference sequence and 14 million alleles that have alternate allele frequency greater than 0.03 in the Brown Swiss cattle breed. Our novel variation-aware reference facilitates accurate read mapping and unbiased sequence variant genotyping for SNPs and Indels.

Conclusions: We develop the first variation-aware reference graph for an agricultural animal https://doi.org/10.5281/zenodo.3759712. Our novel reference structure improves sequence read mapping and variant genotyping over the linear reference. Our work is a first step towards the transition from linear to variation-aware reference structures in species with high genetic diversity and many subpopulations.

Keywords: Variation-aware genome graph, Sequence variant genotyping, Reference allele bias

3.1 Introduction

A reference sequence is an assembly of digital nucleotides that are representative for a species' genetic constitution. Discovery and genotyping of polymorphic sites from whole-genome sequencing data typically involve reference-guided alignment and genotyping steps that are carried out successively [1]. Variants are discovered at positions where aligned sequencing reads differ from corresponding reference nucleotides. Long-read sequencing and sophisticated genome assembly methods enabled spectacular improvements in the quality of linear reference sequences particularly for species with gigabase-sized genomes [2]. Recently generated de novo assemblies exceed in quality and continuity all current reference sequences [3, 4]. However, modifications and amendments to existing linear reference sequences causes shifts in their coordinates that require large efforts from the genomics community to make data compatible with updated reference sequences [5].

Domestication and selection for beef and milk production under various environmental conditions have led to the formation of more than thousand breeds of cattle (*Bos taurus*) with distinct genetic characteristics and high allelic variation within and between breeds [6]. The 1000 Bull Genomes Project discovered almost 100 million sequence variants that are polymorphic in 2700 cattle from worldwide cattle breeds [7, 8]. Nucleotide diversity is higher in cattle than human populations [7, 9]. Yet, all bovine DNA sequences are aligned to the linear consensus reference sequence of a highly inbred Hereford cow to facilitate reference-guided variant discovery and genotyping [10, 11]. A genome-wide alignment of DNA fragments from a *B. taurus* individual differs from the Hereford-based reference sequence at between 7 and 8 million single-nucleotide polymorphisms (SNPs) and small ($<$ 50 bp) insertions and deletions (Indels) [12, 13]. The number of differences is higher for DNA samples with greater divergence from the reference [14, 15].

The bovine linear reference sequence lacks allelic variation and nucleotides that might segregate at high frequency in animals from breeds other than Hereford. Lack of allelic diversity is an inherent drawback of linear reference sequences because it causes reference allele bias. DNA sequencing reads that contain only alleles that match corresponding reference nucleotides are more likely to align correctly than DNA fragments that also contain non-reference alleles [16, 17]. Reads originating from DNA fragments that are highly diverged from corresponding reference nucleotides will either obtain low alignment scores, or align at incorrect locations, or remain un-mapped [18]. Reference bias compromises analyses that are sensitive to accurately mapped reads and prevents

the precise estimation of allele frequencies [16, 19, 20, 21].

Graph-based [17, 22] and personalized reference genomes [5, 23] mitigate reference allele bias. Existing linear reference coordinates can serve as backbones for variation-aware genome graphs. Nodes in the graph represent alleles at sites of variation and edges connect adjacent alleles. Once a variation-aware genome graph contains all alleles at known polymorphic sites, every haplotype can be represented as a walk through the graph [24]. However, an optimal balance between graph density and computational complexity is key to efficient whole-genome graph-based variant analysis because adding sites of variation to the graph incurs computational costs. Recently, Pritt et al. [18] developed the FORGe software tool to prioritize variants for graph genomes. Their results provide a framework to build genome graphs that enable read mapping accuracy improvements over linear references at tractable computational complexity. A genome graph-based sequence analysis workflow is implemented in the variation graph toolkit (*vg*, https://github.com/vgteam/vg) [22]. The *vg toolkit* enables the mapping of sequence reads to variation-aware graphs that incorporate linear reference coordinates as a backbone. It also facilitates to augment genome graphs with genetic variants that have more complex topology (e.g., duplications, inversions, and translocations) [25]. Graph-based references have been investigated primarily in humans and species with small genome sizes [17]. High nucleotide diversity and the separation of individuals by breeds make cattle an ideally suited species to investigate the optimal composition of reference graphs for gigabase-sized genomes.

Here, we investigate sequence read mapping and variant genotyping accuracy using variation-aware reference structures in cattle. Using sequence variant genotypes of 288 cattle from four dairy and dual-purpose breeds, we construct breed-specific augmented and pan-genome reference graphs using the *vg toolkit* [22]. We prioritize sequence variants to be added to the graphs and assess accuracy of read mapping for variation-aware and linear references (Fig. 3.1). We show that breed-specific augmented and pan-genome graphs allow for significant read mapping accuracy improvements over linear reference sequences. We also construct a bovine whole-genome reference graph and show that unbiased and accurate sequence variant genotyping is possible from this novel reference structure. Together, we hope that our study can serve as a first step towards the transition from linear to variation-aware references in species with high genetic diversity and many sub-populations.

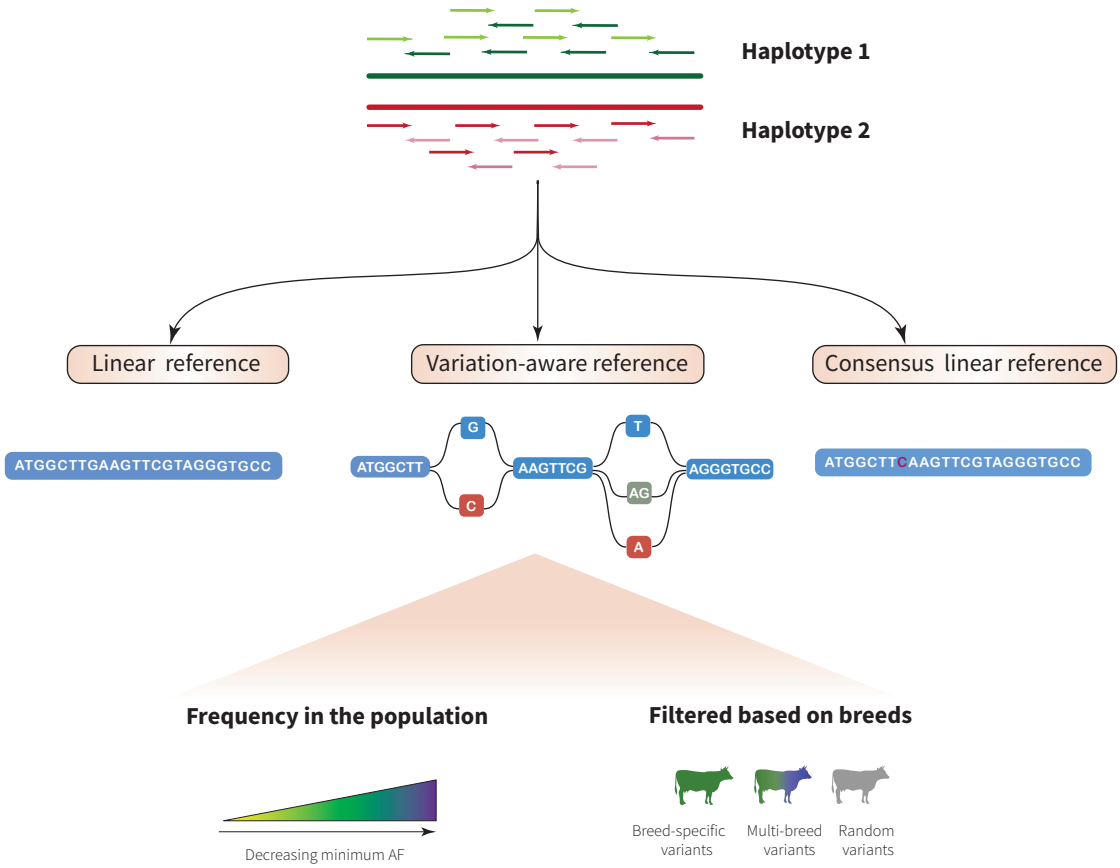


Figure 3.1: **Schematic overview of the construction of breed-specific augmented genome graphs.** We used the *vg toolkit* to augment the bovine linear reference sequence (ARS-UCD1.2) with alleles at SNPs and Indels that were discovered in 288 cattle from four breeds. Alleles that were added to the linear reference were prioritized based on their alternate allele frequency (AF). Reads simulated from true haplotypes were aligned to variation-aware, linear and consensus reference sequences to assess read mapping accuracy on cattle chromosome 25. Short-read sequencing data of Brown Swiss cattle were used to investigate sequence variant genotyping accuracy and reference allele bias using a bovine whole-genome graph as a novel reference.

3.2 Results

Construction of bovine breed-specific augmented genome graphs

Breed-specific augmented reference graphs were constructed for four genetically distinct dairy (Brown Swiss (BSW), Holstein (HOL)) and dual-purpose (Fleckvieh (FV), Original Braunvieh (OBV)) cattle breeds using the Hereford-based linear reference sequence (ARS-UCD1.2) of chromosome 25 as a backbone (Fig. 3.2a). Average nucleotide diversity (π) estimated using 295,801 (HOL), 336,390 (FV), 347,402 (BSW), and 387,855

(OBV) biallelic variants of chromosome 25 ranged from 0.00177 (BSW) to 0.0019 (OBV) for the four breeds (Fig. 3.2b). To determine the optimal composition of bovine variation-aware references, we augmented the linear reference of chromosome 25 with an increasing number of variants (SNPs and Indels) that were filtered for alternate allele frequency in 82 BSW, 49 FV, 49 HOL, and 108 OBV cattle. In total, we constructed 20 variation-aware graphs for each breed that contained between 2046 (variants had alternate allele frequency > 0.9) and 293,804 (no alternate allele frequency threshold) alleles.

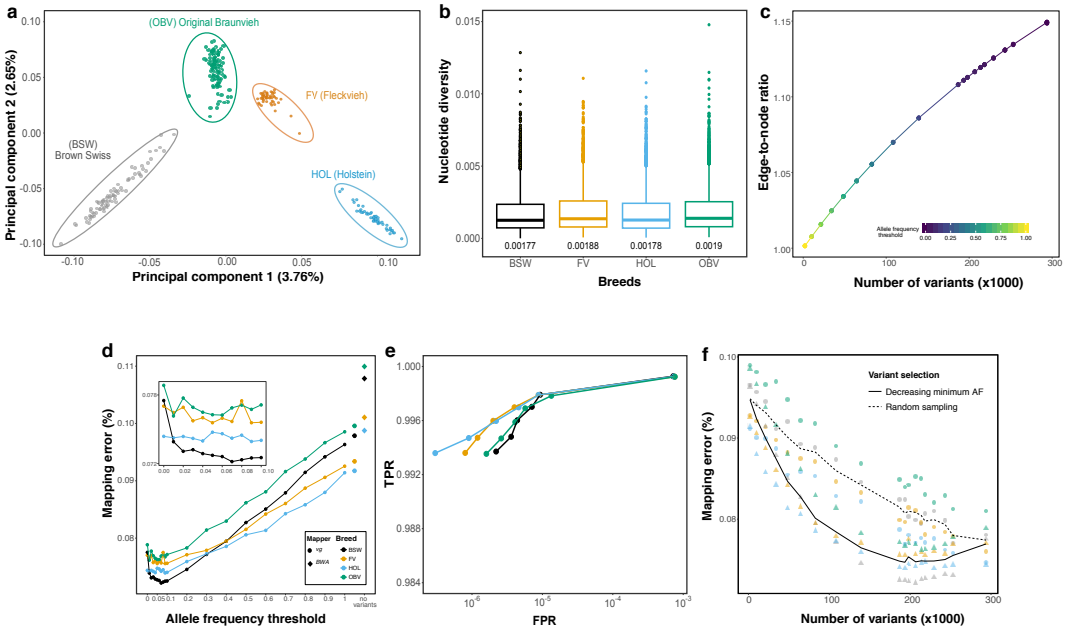


Figure 3.2: **Accuracy of mapping simulated paired-end reads to genome graphs that contained variants filtered for allele frequency at chromosome 25. a** The top principal components of a genomic relationship matrix constructed from whole-genome sequence variants reflect the genetic diversity of the four cattle breeds considered. **b** Nucleotide diversity of the four breeds calculated in non-overlapping 10-kb windows for variants of chromosome 25. The values below each boxplot indicate the nucleotide diversity for the four breeds averaged across all sliding-windows. **c** Edge-to-node ratio of graphs that contained between 2046 and 293,804 variants filtered for allele frequency. **d** Proportion of incorrectly mapped reads for four breed-specific augmented genome graphs. Diamonds and large dots represent values from linear mapping using *BWA mem* and *vg*, respectively. The inset represents a larger resolution of the mapping accuracy for alternate allele frequency thresholds less than 0.1. **e** True-positive (sensitivity) and false-positive mapping rate (specificity) parameterized on mapping quality of the best performing graph from each breed. **f** Read mapping accuracy for breed-specific augmented graphs that contained variants that were either filtered for alternate allele frequency (triangles) or sampled randomly (circles) from all variants detected within a breed. The dashed and solid line represents the average proportion of mapping errors across four breeds using random sampling and variant prioritization, respectively. Colors indicate values obtained for different breeds. Results for single-end mapping are presented in Fig. S3.2.

The graph-based representation of bovine chromosome 25 (42,350,435 nucleotides) had 1,323,451 nodes and 1,323,450 edges. The number of nodes increased proportionally

with the number of variants added to the reference. When we added a maximum number of 293,804 variants to the linear reference sequence of chromosome 25, the variation-aware graph contained 2.02 million nodes. The number of edges increased faster than the number of nodes, ranging from 1.32 (empty) to 2.33 (293,804 variants included) million. Consequently, the edge-to-node ratio increased when variants were added to the graph (Fig. 3.2c). The number of paths through a graph grows rapidly with the number of variants being added to the graph. The index for the chromosome 25 reference graph contained 84.69 and 118.82 million *k*-mers (k = 256) when 2046 and 293,804 variants, respectively, were added to the graphs (Fig. S3.1).

Variant prioritization based on allele frequency

We simulated 10 million paired-end reads (2 x 150 bp) corresponding to approximately 35-fold coverage of bovine chromosome 25 from haplotypes of BSW, FV, HOL, and OBV cattle. Using either *BWA mem* or *vg*, we mapped the simulated reads to the respective breed-specific augmented reference graphs and the linear reference sequence. Variants that were only detected in animals used for read simulation were not added to the breed-specific augmented genome graphs. We observed fewer mapping errors using *vg* than *BWA mem* when simulated reads were aligned to a linear reference sequence. This finding was consistent for the four breeds investigated (Fig. 3.2d). Variation-aware references that contained variants filtered for allele frequency in the respective breed reduced the mapping errors for all breeds. The proportion of reads with mapping errors decreased significantly with the number of variants added to the genome graph (Fig. 3.2d, Pearson R = 0.94, $P < 10^{-16}$).

Read mapping accuracy increased almost linearly between alternate allele frequency threshold 1 and 0.1, i.e., until 186,680 variants with allele frequency greater than 0.1 were added to the graph (Pearson R = 0.94, $P < 10^{-16}$). Adding additional alleles that had alternate allele frequency between 0.1 and 0.01 to the graphs did not further improve read mapping accuracy over the scenario with an alternate allele frequency threshold of 0.1 ($P = 0.13$, Fig. 3.2d inset). Read mapping accuracy declined (particularly in BSW) when the graphs contained rare alleles (alternate allele frequency < 0.01) likely because such alleles are not observed in most animals of a population. Maximum read mapping accuracy was achieved at allele frequency thresholds between 0.2 and 0.01, when the graphs contained between 139,322 and 293,628 variants filtered for allele frequency. The number of erroneously mapped reads was clearly higher for graphs that contained randomly sampled than prioritized variants (Fig. 3.2f). This finding corroborates that variant prioritization based on alternate allele frequency is important to achieve high mapping accuracy with graph-based reference structures.

We also applied the methods implemented in the FORGe software [18] to prioritize variants for the breed-specific augmented graphs (Note S3.1). It turned out that genome graphs that were constructed with variants selected by the *Pop Cov* strategy, which relies solely on variant frequency information, enabled the highest mapping accuracy improvements over the linear reference. For example, we achieved the highest paired-end read mapping accuracy for the Brown Swiss reference graph (0.0722% erroneously mapped reads) using the *Pop Cov* method when 208,288 variants were added to the chromosome 25 reference (i.e., the top 60% of the ranked variants). The prioritized variants correspond to an alternate allele frequency threshold of 0.06. Variant prioritization approaches that also take into account factors other than allele frequency, e.g., the proximity of a variant to an already added variant in the graph or the repetitiveness of the resulting genome graph, did not lead to additional accuracy improvements.

Read mapping accuracy was highly correlated (Pearson R = 0.94, $P < 10^{-16}$) for single- and paired-end reads (Fig. S3.2). However, the accuracy improvement of variation-aware over linear mapping was higher for single- than paired-end reads, possibly because distance and sequence information from paired reads facilitate linear read alignment.

Read mapping accuracy differed significantly among the four breeds analyzed ($P = 10^{-15}$, linear model with allele frequency as covariate) although all breed-specific augmented graphs contained the same number of variants at each allele frequency threshold (Fig. 3.2d). Linear mapping accuracy also differed among the breeds. We observed the highest error rate for reads aligned to the OBV-specific augmented reference graph. In 500 randomly sampled subsets of 35 sequenced cattle per breed, we discovered more sequence variants on chromosome 25 in OBV (N = 305 $\pm$ 5K) than either FV (N = 291 $\pm$ 3K), BSW (N = 276 $\pm$ 6K) or HOL (N = 259 $\pm$ 2K), reflecting that nucleotide diversity is higher in OBV than the other three breeds, which agrees with a recent study [26]. Across all alternate allele frequency thresholds considered, read mapping was more accurate for HOL than FV and OBV cattle, possibly because both genetic diversity and effective population size is less in HOL than the other breeds considered [27]. At allele frequency thresholds between 0.02 and 0.3, read mapping was more accurate for BSW than the other breeds. The proportion of variants with alternate allele frequency larger than 0.02 was lower for BSW(84.1%) than other breeds (86.3$-$89.2%). We detected more rare variants (allele frequency less than 0.05) in BSW and OBV than FV and HOL, likely reflecting differences in sample size (Fig. S3.3). An excess of singletons and rare variants in BSW and OBV cattle may have contributed to the decline in mapping accuracy at low alternate allele frequency thresholds (Fig. 3.2d inset, Table S3.2). Our findings indicate that differences in nucleotide diversity and allele frequency distributions across popu-

lations may affect read mapping accuracy to both linear and breed-specific augmented reference structures.

Comparison between bovine and human genome graphs

We used publicly available whole-genome sequence variant data from phase 3 of the 1000 Genomes Project [28] to construct genome graphs for four genetically distinct human populations (Fig. 3.3a, GBR (British, European), YRI (Yoruba Nigeria, African), STU (Sri Lankan Tamil, South Asia), and JPT (Japanese, East Asia)). The effective population size is more than 20-fold higher for the human than cattle populations (e.g., $\sim$ 3100 for JPT and $\sim$ 7500 for YRI [29] vs. $\sim$ 80 for OBV and $\sim$ 160 for FV [30, 31]). While the average number of sequence variants detected per sample was lower for the human than cattle populations (4,248,082 vs. 6,973,036), the proportion of singletons is higher in the human than cattle samples (23.00% in human vs. 14.01% in cattle) Table (S3.1). The proportion of sequence variants that had minor allele frequency less than 0.05 was between 44.88 and 55.45% in the four human and between 23.65 and 38.70% in the four cattle populations (Fig. S3.4). Nucleotide diversity ranged from 0.00098 (JPT) to 0.00141 (YRI) (Fig. 3.3b).

We considered the linear reference sequence of human chromosome 19 (g1k_v37 ref) as a backbone for the human genome graphs because its length (59,128,893 bp) and the number of variants detected per sample was similar to the values for bovine chromosome 25. Genetic diversity and allele frequency distributions were similar using either chromosome 19 or whole-genome variants indicating that the results obtained using chromosome 19 are representative for the human genome (Figs. S3.4, S3.5, Table S3.2). To construct population-specific augmented graphs, we used phased genotypes at 291,303, 306,304, 355,107, and 521,021 variants of chromosome 19 that were available for 104 JPT, 91 GBR, 102 STU, and 108 YRI individuals. Once the variants that were only detected in individuals used for simulating reads were removed from the graphs, the population-specific augmented graphs for the GBR, YRI, STU, and JPT populations contained between 3153 (alternate allele frequency > 0.9) and 290,593 (no alternate allele frequency threshold) variants. We subsequently simulated 10 million reads from haplotypes of one individual per population and mapped the reads to the respective population-specific augmented genome graphs.

As observed for the bovine breed-specific augmented genome graphs, read mapping accuracy increased almost linearly between alternate allele frequency threshold 1 (no variants included) and 0.1 (133,891 variants added to the graph) (Fig. 3.3c). Adding low-frequency variants (alternate allele frequency between 0.01 and 0.1) did not further

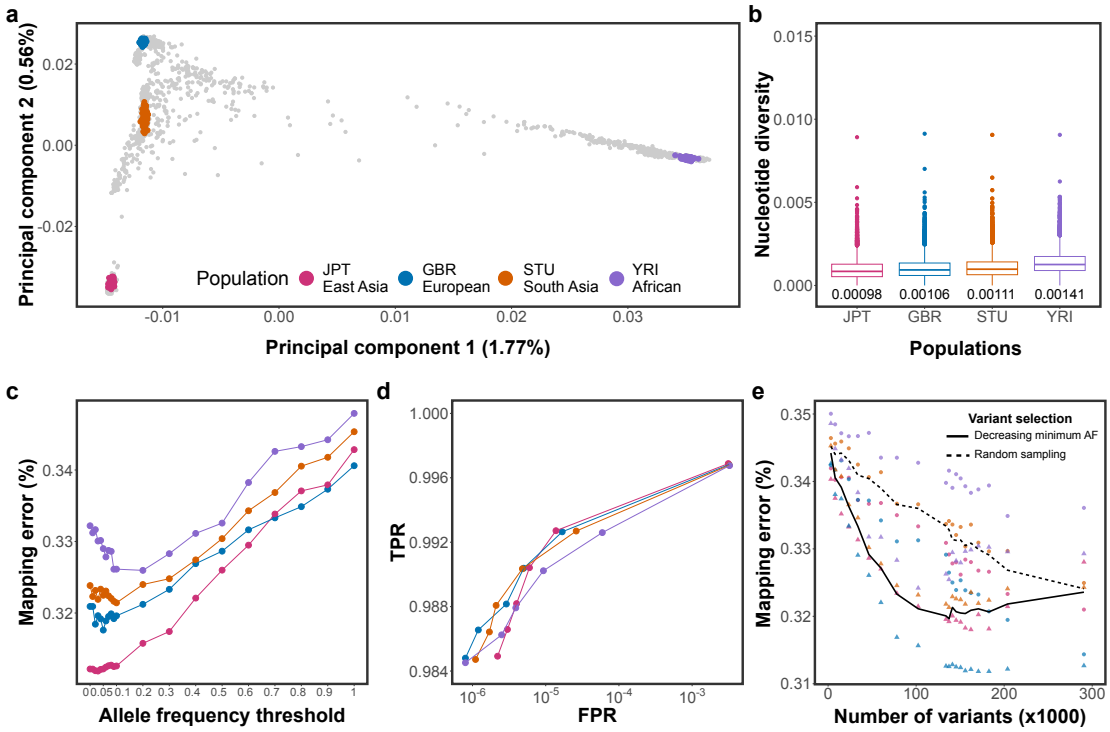


Figure 3.3: **Accuracy of mapping simulated paired-end reads to human population-specific augmented genome graphs.** **a** The top principal components of a genomic relationship matrix constructed from autosomal variants detected in 2504 individuals that were included in phase 3 of the 1000 Genomes Project. The colored points indicate 405 samples from the GBR (European), YRI (African), STU (South Asia), and JPT (East Asia) populations. **b** Nucleotide diversity of the four populations calculated in non−overlapping 10 kb windows for variants of chromosome 19. The values below each boxplot indicate the nucleotide diversity for the four populations averaged across all sliding-windows. **c** Proportion of incorrectly mapped reads for four population-specific augmented genome graphs. **d** Truepositive (sensitivity) and falsepositive mapping rate (specificity) parameterized on mapping quality of the best performing graph from each population. **e** Read mapping accuracy for population-specific augmented graphs that contained variants that were either filtered for alternate allele frequency (triangles) or sampled randomly (circles) from all variants detected within a population. The dashed and solid line represents the average proportion of mapping errors across four populations using variant prioritization and random sampling, respectively. Results for single-end mapping are presented in Fig S3.6

improve the mapping accuracy. Mapping accuracy decreased for all graphs when we added very rare variants and singletons to the graphs. This pattern was most apparent for YRI which had the highest proportion of rare variants and nucleotide diversity among the four populations considered. Read mapping accuracy differed among the four populations analyzed. We observed the lowest number of mismapped reads when reads simulated from a JPT individual were aligned to a JPT-specific augmented genome graph. The highest number of mis-mapped reads was observed when reads simulated from a YRI individual were aligned to a YRI-specific augmented genome graph. Mapping accuracy was higher for GBR than STU. These findings indicate that the mapping

accuracy is negatively correlated with nucleotide diversity. Mapping accuracy improvements over the linear reference sequence were less when randomly sampled variants were added to the graphs (Fig. 3.3e).

While the overall pattern of the mapping accuracy improvements over the linear reference was similar for human and bovine genome graphs across all allele frequency thresholds considered, the proportion of mis-mapped paired-end reads was approximately four-fold higher in the human than bovine alignments (two-fold for single-end reads; S3.6). This finding was also apparent when the population-specific augmented graphs were parameterized on mapping quality to obtain sensitivity and specificity (Fig. 3.2e and Fig. 3.3d).

Mapping to breed-specific augmented genome graphs

Next, we compared read mapping accuracy between bovine breed-specific augmented and pan-genome graphs (i.e., graphs that contained variants filtered for allele frequency across multiple populations) using reads simulated from phased variants of bovine chromosome 25. We constructed four breed-specific augmented genome graphs that contained variants that had alternate allele frequency > 0.03 in either the BSW, FV, HOL, or OBV breeds. HOL had the lowest number of variants (N $= 243{,}145$) with alternate allele frequency > 0.03, reflecting that sample size was lower in HOL than the other breeds. To ensure that the density of information was comparable across all breed-specific augmented graphs, we randomly sampled 243,145 variants with alternate allele frequency > 0.03 from the BSW, FV, and OBV populations and added them to the respective graphs. The pan-genome graph contained variants that had alternate allele frequency > 0.03 in 288 individuals from the four populations. The random graph contained 243,145 randomly sampled variants for which haplotype phase and the allele frequency in the BSW, FV, HOL, or OBV breeds was unknown (see the Methods section). To investigate read mapping accuracy, we simulated 10 million sequencing reads (150 bp) from BSW haplotypes and mapped them to the variation-aware and linear reference sequences. Variants that were only detected in the BSW animal used for simulating reads were excluded from the graphs. However, in order to determine an upper bound for graph-based read mapping accuracy, we also constructed a "personalized" genome graph, i.e., a graph that contains only haplotypes of the animal used for simulating the reads. We repeated the selection of variants, construction of variation-aware graphs and subsequent read mapping ten times.

The average length, number of nodes, number of edges, and edge-to-node ratio of the variation-aware graphs were 42.60 Mb, 1,907,248, 2,155,799, and 1.13, respectively.

Most variants of the random graph (87.81%) were not detected at alternate allele frequency > 0.03 in BSW, FV, OBV, and HOL indicating that they were either very rare or did not segregate in the four breeds considered in our study. Of 243,145 variants, an intersection of 48.13% had alternate allele frequency greater than 0.03 in the four breeds considered (Fig. 3.4a). The average number of variants that were specific to the breed-specific augmented graphs ranged from 8010 in BSW to 20,392 in FV.

Personalized genome graphs, i.e., graphs that are tailored to a specific individual, enable the largest read mapping accuracy improvements over linear references. The proportion of mis-mapped reads was 0.0694% when a personalized BSW graph was used as a reference. Apart from the personalized graph, the highest mapping accuracy, sensitivity, and specificity was achieved when the simulated BSW reads were aligned to a BSW-specific augmented graph (Fig. 3.4b–d). The proportion of erroneously mapped paired-end reads was 0.073% for the BSW-specific augmented graph. Sensitivity and specificity were slightly lower and the number of reads with mapping errors was slightly higher when the same reads were aligned to a pan-genome graph. The read mapping accuracy differed only slightly between the breed-specific augmented and pan-genome graph because the overlap of variants that were included in both variation-aware references was high (Fig. S3.8). The number of mapping errors was higher (adjusted $P < ^{-16}$, *pairwise t test*, S3.9) when BSW reads were aligned to genome graphs that contained variants filtered for allele frequency in either the FV, HOL, or OBV populations.

We also simulated reads from haplotypes of FV, HOL, and OBV cattle. Similar to our findings using reads simulated from BSW cattle, mapping was more accurate to breed-specific than either pan-genome graphs or graphs that were augmented with variants filtered for allele frequency in other breeds (Fig. S3.10).

Mapping reads to a linear reference sequence using *BWA mem* with default parameter settings was the least sensitive and least specific mapping approach tested. Linear mapping using *vg* was also less accurate than variation-aware mapping. This finding indicates that accuracy improvements of variation-aware over linear mapping are attributable to differences in the reference structure rather than mapping algorithms. All graphs that contained pre-selected variants that had alternate allele frequency greater than 0.03 enabled significantly ($P = 10^{-16}$, *two-sided t test*) higher mapping accuracy than linear references. This was also true when reads were mapped to graphs that contained variants that were filtered for allele frequency in a different breed, likely because many common variants segregated across the four breeds considered (Fig. 3.4a).

Recently, Grytten et al. [32] showed that an adjusted linear alignment approach that

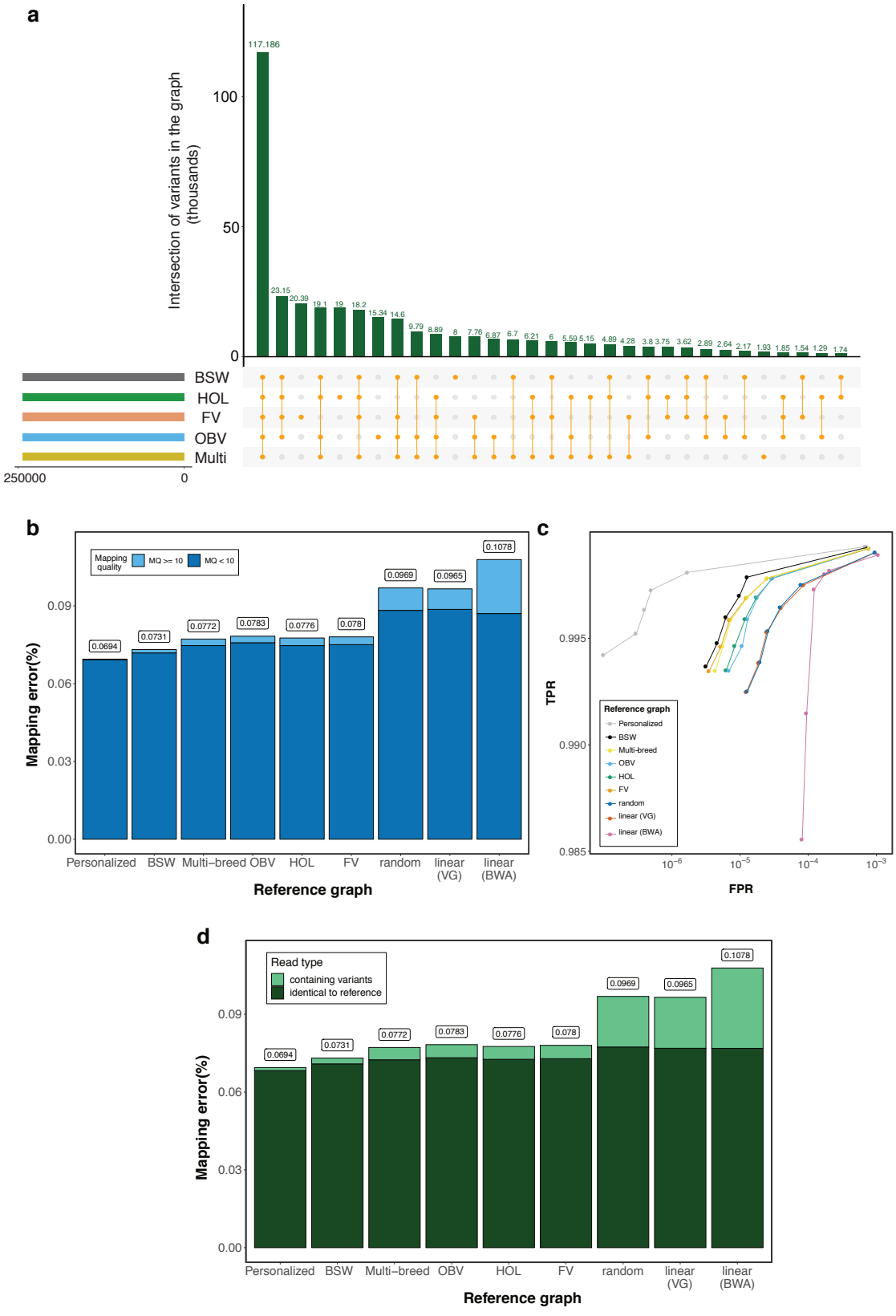

a
Intersection of variants in the graph (thousands)
117.186
23.15
20.39
19.1
19
18.2
15.34
14.6
9.79
8.89
8
7.76
6.87
6.7
6.21
6
5.59
5.15
4.89
4.28
3.8
3.75
3.62
2.89
2.64
2.17
1.93
1.85
1.54
1.29
1.74
BSW
HOL
FV
OBV
Multi
250000
0
b
Mapping quality
MQ >= 10
MQ < 10
Mapping error(%)
0.0694
0.0731
0.0772
0.0783
0.0776
0.078
0.0969
0.0965
0.1078
Personalized
BSW
Multi−breed
OBV
HOL
FV
random
linear (VG)
linear (BWA)
Reference graph
c
TPR
FPR
Reference graph
Personalized
BSW
Multi-breed
OBV
HOL
FV
random
linear (VG)
linear (BWA)
d
Read type
containing variants
identical to reference
Mapping error(%)
Reference graph

Figure 3.4: **The accuracy of mapping simulated BSW paired-end reads to variation-aware and linear reference structures.** **a** We added 243,145 chromosome 25 variants to the Hereford-based reference sequence that were filtered for alternate allele frequency > 0.03 in either the BSW, FV, HOL, or OBV populations. The pan-genome graph (Multi) contained 243,145 variants that had alternate allele frequency threshold > 0.03 across 288 cattle from the four breeds considered. The bars indicate the overlap of variants (averaged across ten replications) that were added to different graphs. **b** Proportion of simulated BSW reads that mapped erroneously against personalized graphs, breed-specific augmented graphs, pan-genome graphs (Multi-breed), random graphs, or linear reference sequences. We used *vg* and *BWA mem* for linear mapping. Dark and light blue colors represent the proportion of incorrectly mapped reads that had phred-scaled mapping quality (MQ) < 10 and MQ > 10, respectively. **c** True-positive (sensitivity) and false-positive mapping rate (specificity) parameterized on mapping quality. **d** Proportion of BSW reads that mapped incorrectly against breed-specific augmented graphs, pan-genome graphs (Multi-breed), random graphs, or linear reference sequences. Dark and light green colors represent the proportion of incorrectly mapped reads that matched corresponding reference nucleotides and contained non-reference alleles, respectively. Results for single-end mapping are presented in Fig. S3.7

relies on a combination of *BWA mem* and *Minimap2* [33] may improve linear mapping accuracy because the default setting of *BWA mem* might miss sub-optimal alignments and overestimate mapping quality for multi-mapping reads [32, 34]. We found that this approach enables to reduce the proportion of mis-mapped from 0.1078 to 0.0983 in cattle (Note S3.2). Improved mapping accuracy from the combination of *BWA mem* and *Minimap2* primarily results from less incorrectly mapped reads that had mapping quality > 10, indicating a better mapping quality assignment. The mapping accuracy from the adjusted linear alignment approach is similar to the linear mapping accuracy obtained using *vg* but considerably lower than using breed-specific augmented graphs (Note S3.2). The number of paired-end reads with mapping errors is 26% higher using the adjusted linear alignment approach than breed-specific augmented reference graphs.

Reference graphs that contained random variants, i.e., variants that were neither phased, nor filtered for allele frequency in the breeds of interest, did not improve mapping accuracy, sensitivity and specificity over linear references (adjusted $P = 0.74$ and 0.35 for single- and paired-end, *pairwise t test*, Fig. S3.9).

Compared to linear mapping using BWA mem with default parameter settings, the number of mapping errors decreased by 39 and 31% for single- and paired-end reads, respectively, using a breed-specific augmented reference graph. Extrapolated to whole-genome sequencing data required for a 35-fold genome coverage, the use of breed-specific augmented reference graphs could reduce the number of incorrectly mapped single- and paired-end reads by 1,300,000 and 220,000, respectively.

Using the BSW-specific augmented graph as a reference, only 1.76% of the incor-

rectly mapped reads had mapping quality (MQ) greater than 10. The MQ of the vast majority (98.24%) of incorrectly mapped reads was less than 10, i.e., they would not qualify for sequence variant discovery and genotyping using *GATK* with default parameter settings. The proportion of incorrectly mapped reads with MQ $>$ 10 was twice as high using either the pan-genome or an across-breed augmented reference graph (3.21–3.85%). The proportion of incorrectly mapped reads with MQ $>$ 10 was higher using either the random graph (8.92%) or linear reference sequence (*vg*: 8.19%, *BWA mem*: 19.3%).

Of 10 million simulated reads, 19.16% contained at least one nucleotide that differed from corresponding Hereford-based reference alleles. Using *BWA mem*, 47.44% and 28.72% of the erroneously mapped single- (SE) and paired-end (PE) reads, respectively, contained alleles that differed from corresponding reference nucleotides indicating that incorrectly mapped reads were enriched for reads that contained non-reference alleles (Fig. 3.4d, Figs. S3.7, S3.11). The proportion of erroneously mapped reads that contained non-reference alleles was similar for reads that were aligned to either random (47.62% and 20.13%) or empty graphs (48.20% and 20.35%) using *vg*. However, the proportion of incorrectly mapped reads that contained non-reference alleles was clearly lower for the breed-specific augmented (SE: 1.37%, PE: 3.08%) and pan-genome graphs (SE: 2.12%, PE: 6.14%). The proportion of incorrectly mapped reads that matched corresponding reference nucleotides was almost identical across all mapping scenarios tested (Figs. 3.4d, S3.7, S3.11).

Using data from the Ensembl bovine gene annotation (version 99) and RepeatMasker, we determined if the simulated reads originate from either genic regions, interspersed duplications, or low-complexity and simple repetitive regions Fig (S3.12). Regardless of the reference structure used, the mapping accuracy was low for reads originating from repetitive regions. Mapping accuracy was higher for reads originating from either genic or exonic regions. Graph-based references enabled more accurate mapping of reads originating from either genic regions or interspersed duplications (including SINEs, LINEs, LTR, and transposable elements) than linear reference sequences. However, graph-based references did not improve the mapping accuracy over linear references for reads that originate from low-complexity or simple repetitive regions.

We further augmented the BSW-specific genome graph with 157 insertion and deletion polymorphisms of bovine chromosome 25 that were detected from short paired-end reads (2×150 bp) of 82 BSW animals using *Delly*. Adding these variants to the graph either alone or in addition to 243,145 variants that were detected using *GATK* did not improve the mapping accuracy over the corresponding scenarios that did not include

these variants (Note S3.3).

Linear mapping accuracy using a consensus reference sequence

Previous studies reported that linear mapping may be more accurate using population-specific than universal linear reference sequences [5, 35, 36]. In order to construct bovine linear consensus reference sequences, we replaced the alleles of the chromosome 25 ARS-UCD1.2 reference sequence with corresponding major alleles at 67,142 and 73,011 variants that were detected in 82 BSW and 288 cattle from four breeds, respectively. Subsequently, we aligned 10 million simulated BSW reads to the linear adjusted sequences using either *vg* or *BWA mem*. Read mapping was more accurate to the consensus than original linear reference sequence (Figs. 3.5, S3.13). The accuracy of mapping was higher when reference nucleotides were replaced by corresponding major alleles that were detected in the target than multi-breed population. However, the mapping of reads was less accurate, sensitive, and specific using either of the consensus linear reference sequences than the breed-specific augmented graphs (Fig. 3.5b).

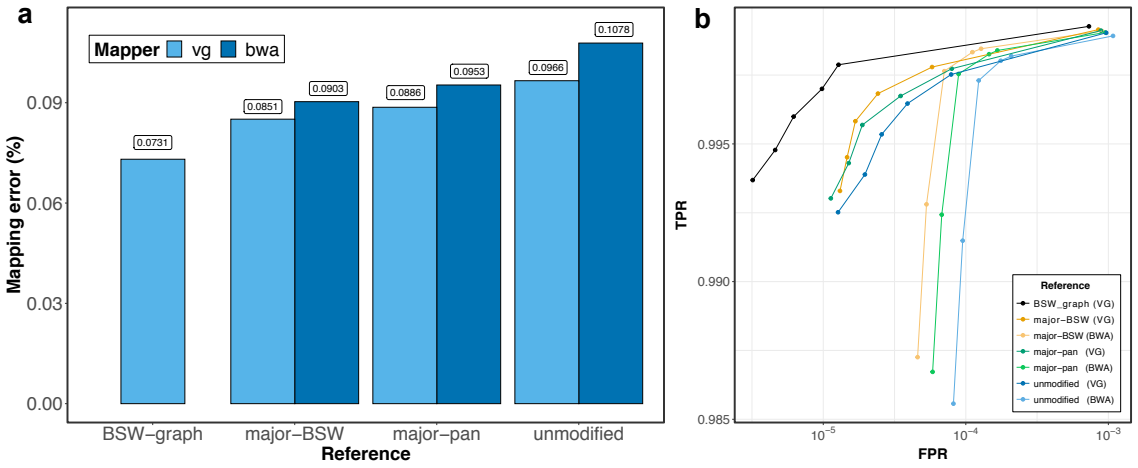


Figure 3.5: **Paired-end read mapping accuracy using breed-specific augmented genome graphs and consensus linear reference sequences. a** Dark and light blue represent the proportion of reads that mapped incorrectly using *BWA mem* and *vg*, respectively, to the BSW-specific augmented reference graph (BSW-graph), the BSW-specific (major-BSW) and the multi-breed linear consensus sequence (major-pan) and the bovine linear reference sequence (unmodified). **b** True-positive (sensitivity) and false-positive mapping rate (specificity) parameterized based on the mapping quality. The results of an analysis where reference nucleotides were only replaced at SNPs is available in Fig. S3.13

Read mapping and variant genotyping using whole genome graphs

In order to develop a breed-specific augmented reference structure for whole-genome applications, we constructed a BSW-specific augmented whole-genome variation-aware reference graph using 14,163,824 autosomal biallelic variants (12,765,895 SNPs and 1,397,929 Indels) that had alternate allele frequency greater than 0.03 in 82 BSW cattle. The result-

ing graph contained 111,511,367 nodes and 126,058,052 edges (an edge-to-node ratio of 1.13) and 6.32 x 10^9 256-mer paths. We also constructed a linear (empty) whole-genome graph that did not contain allelic variation. Subsequently, we mapped paired-end (2×150 bp) sequencing reads of 10 BSW cattle that had been sequenced at between 6- and 40-fold coverage (Table S3.4) to the variation-aware and linear reference sequence using either *vg map* or *BWA mem*. The 10 BSW cattle used for sequence read mapping were different to the 82 animals used for variant discovery, graph construction, and haplotype indexing.

62.19, 51.35 and 49.16% of the reads aligned perfectly (i.e., reads that aligned with full length (no clipping) and without any mismatches or Indels) to the BSW-specific augmented graph, the empty graph, and the linear reference sequence, respectively (Fig. 3.6a). We observed slightly less uniquely mapped reads using either the whole-genome (82.46%) or empty graph (82.18%) than the linear reference sequence (83.18%) indicating that variation-aware references can increase mapping ambiguity due to providing alternative paths for read alignment.

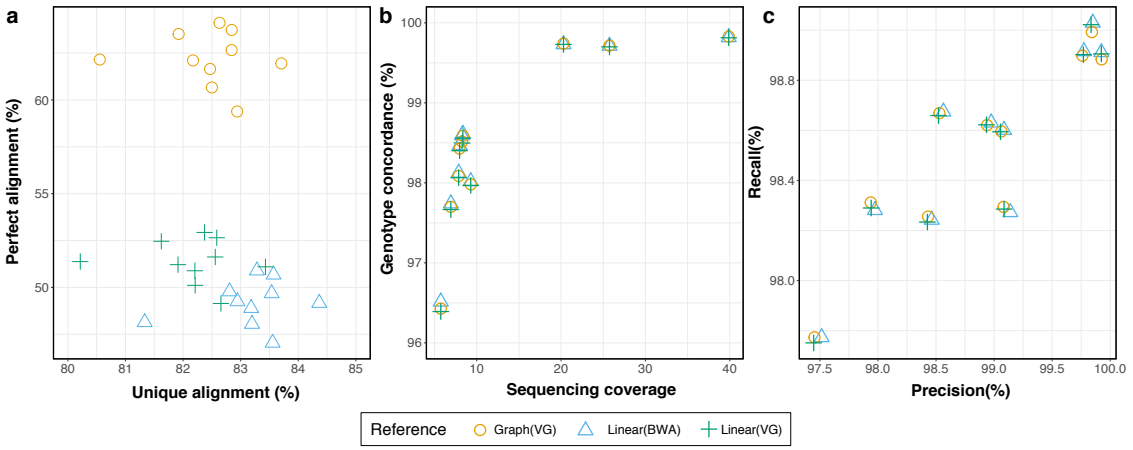


Figure 3.6: **Sequence read mapping and variant genotyping using a breed-specific augmented whole-genome graph. a** Proportion of sequencing reads that mapped perfectly and uniquely to the BSW-specific augmented (circle) and Hereford-based linear (triangle, cross) reference. **b** Concordance between sequence variant and corresponding microarray-derived genotypes as a function of sequencing depth. Sequence variant genotypes were obtained using the multi-sample variant calling approach implemented in *SAMtools*. **c** Corresponding precision-recall statistic. Each symbol represents one BSW animal

We converted (surjected) the graph-based read alignments of 10 BSW cattle to corresponding linear reference coordinates and genotyped polymorphic sites using *SAMtools mpileup*. In order to assess genotyping accuracy, we compared the sequence variant genotypes with array-called genotypes at corresponding positions. Sequence variant genotyping accuracy was correlated with sequencing coverage (Fig. 3.6b). Genotype concordance, non-reference sensitivity, non-reference discrepancy, and precision did not

differ between the graph-based and linear alignments for both raw and hard-filtered genotypes (Fig. 3.6b, c, Table S3.3). The average concordance, precision and recall from the graph-based alignments was 99.76, 99.84, and 98.93, respectively, for three samples (SAMEA6163185, SAMEA6163188, SAMEA6163187) with sequencing coverage greater than 20-fold. We observed similar values for genotypes called using either *GATK* or *Graphtyper* (Table S3.3). In agreement with our previous findings [12], genotype concordance was slightly higher using *Graphtyper*, than either *SAMtools* or *GATK*.

Variation-aware alignment mitigates reference allele bias

To investigate reference allele bias in genotypes called from linear and graph-based alignments, we aligned sequencing reads of a BSW animal that was sequenced at 40-fold coverage (SAMEA6163185) to either the BSW-specific augmented whole-genome graph or linear reference sequence (Table S3.4). We called genotypes using either *SAMtools mpileup* or *GATK*. The genotypes were filtered stringently to obtain a high-confidence set of 2,507,955 heterozygous genotypes (2,217,069 SNPs and 290,886 Indels, see the Methods section) for reference allele bias evaluation. The BSW-specific augmented whole-genome reference graph contained the alternate alleles at 2,194,422 heterozygous sites (87.49%).

Using *SAMtools* to genotype sequence variants from variation-aware and linear alignments, the support for reference and alternate alleles was almost equal at heterozygous SNPs (Fig. 3.7a), indicating that SNPs are not notably affected by reference allele bias regardless of the reference structure. Alternate allele support decreased with variant length for the linear alignments. As expected, bias towards the reference allele was more pronounced at insertion than deletion polymorphisms. For instance, for 456 insertions that were longer than 30 bp, only 26% of the mapped reads supported the alternate alleles. The allelic ratio of Indel genotypes was closer to 0.5 using graph-based than linear alignments indicating that variation-aware alignment mitigates reference allele bias. However, slight bias towards the reference allele was evident at insertions with length > 12 bp, particularly if the alternate alleles were not included in the graph (Fig. 7a). Inspection of the read alignments using the Sequence Tube Map graph visualization tool [37] corroborated that the support for alternate alleles is better using graph-based than linear references (Fig. S3.14).

Both the number of reads mapped and the number of mapped reads supporting alternate alleles was higher at Indels using graph-based than linear alignments (Fig. S3.15).

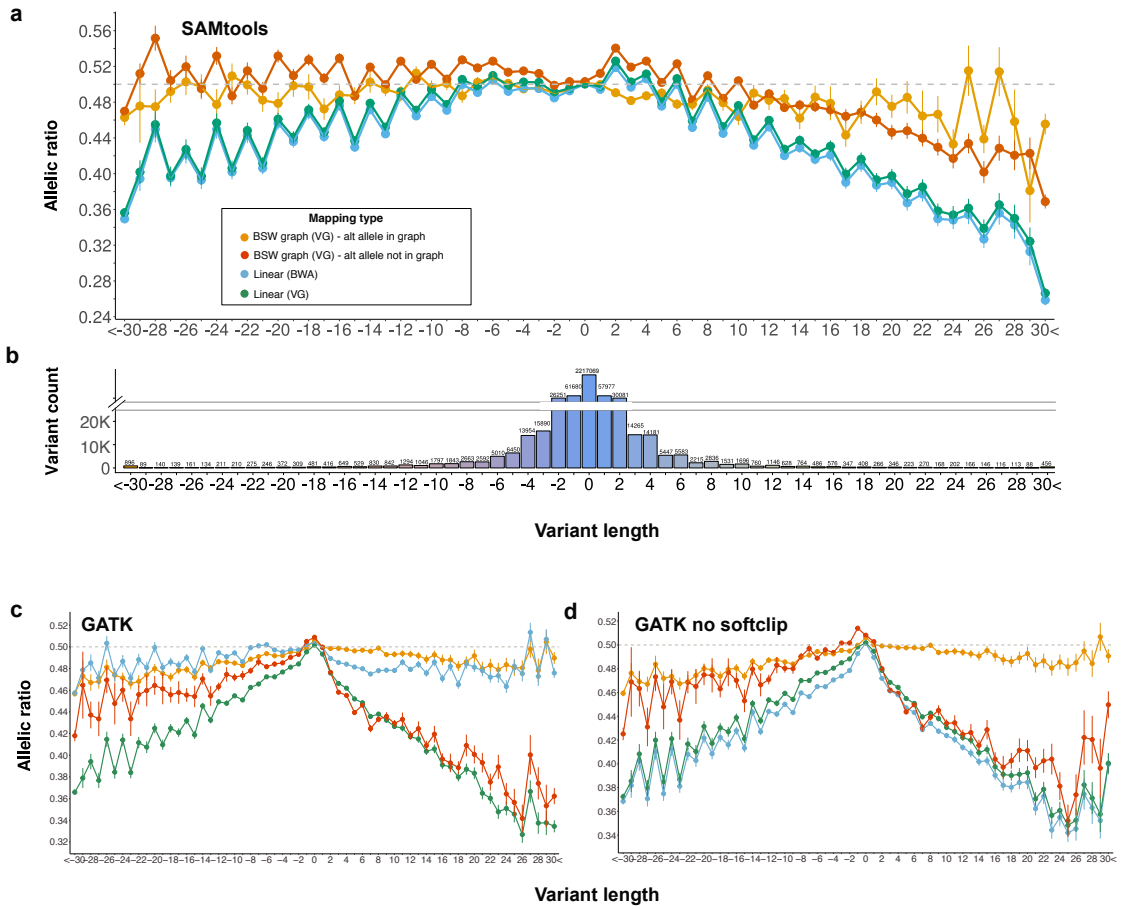


Figure 3.7: **Reference allele bias from graph-based and linear alignments.** Reference allele bias from graph-based and linear alignments using **a** *SAMtools*, **c** *GATK*, or **d** *GATK* without soft-clip for variant genotyping and either *BWA mem* or *vg* for alignment. Allelic ratio reflects the proportion of mapped reads supporting the alternate allele. The gray dashed line indicates equal support (0.5) for both alleles. Negative values, zero, and positive values along the *x*-axis represent deletions, SNPs, and insertions respectively. Each dot represents the mean ($\pm$ s.e.m.) allelic ratio for a given variant length. b Number of variants with a given length. To improve the readability, the values above the breakpoint of the y-axis do not scale proportionately with the height of the bars

The difference in the number of mapped reads between graph-based and linear alignments increased with variant length. However, the number of mapped reads supporting the reference alleles did not differ between the graph-based and linear alignments. This finding indicates that reduced reference allele bias at Indel genotypes called from graph-based alignments is due to the improved mapping of reads that contain non-reference alleles.

We next investigated if these conclusions also hold for genotypes called by *GATK*. While *SAMtools mpileup* detects variants directly from the aligned reads [38], *GATK HaplotypeCaller* locally realigns the reads and calls variants from the refined alignments

[39]. Using *GATK*, the allelic ratio was close to 0.5 for genotypes called from graph-based alignments across different lengths of variants that were included in the reference graph (Fig. 3.7c). However, reference allele bias was evident at insertions that were not included in the reference graph. We also observed an almost equal number of reference and alternate alleles at variants genotyped from linear alignments using *GATK*. These findings confirm that the local realignment and haplotype-based genotyping approach of *GATK* might also mitigate reference alleles from linear alignments.

The percentage of soft-clipped reads increased with Indel length in the linear alignments (Fig. S3.16). However, the graph-based alignments contained almost no soft-clipped reads across all Indel lengths. In order to investigate the impact of soft-clipping on variant genotyping, we repeated *GATK* variant discovery and genotyping for the graph-based and linear alignments after all soft-clipped reads were removed (Fig. 3.7d). As expected, the allelic ratio of genotypes called from the graph-based alignments was not affected by the removal of (very few) soft-clipped reads. However, bias towards the reference allele became evident in genotypes called from linear alignments. This finding confirms that the local realignment of *GATK* rescues Indels that are initially soft-clipped, thus mitigating reference allele bias. This finding also implies that the original pileup information from graph-based alignments facilitates to confidently detect known Indels while avoiding local realignment as implemented in the *GATK HaplotypeCaller*.

3.3 Discussion

To the best of our knowledge, our study is the first to investigate the utility of a variation-aware reference for a species with a gigabase-sized genome other than human. We constructed bovine breed-specific consensus sequences and variation-aware reference graphs using a Hereford-based linear reference sequence as backbone and variants that were filtered for allele frequency in four cattle breeds other than Hereford to investigate read mapping accuracy and variant genotyping from different reference structures.

Using sequencing reads simulated from haplotypes of BSW, FV, OBV, and HOL cattle, our findings confirm that a breed-specific consensus sequence improves linear mapping [5, 35]. However, read mapping is less accurate using linear consensus than variation-aware references that contain pre-selected variants. Grytten et al. [32] reported that an adjusted parameter setting of *BWA mem* and subsequent application of *Minimap2* may further improve the linear mapping accuracy. However, the adjusted linear mapping approach still performs worse than graph-based mapping on reads that contain

variants. The accuracy improvements of the adjusted linear mapping approach were small in our study, because the number of sequence variants detected per sample and thus the proportion of reads with variants is almost twice as high in cattle than humans (Table S3.1).

Using a bovine variation-aware reference reduced the proportion of erroneously mapped reads by more than 30% compared to the most widely used linear mapping approach. A similar improvement in mapping accuracy over the linear reference was achieved for a human variation-aware reference genome [18]. The graph-based alignments using the most accurate breed-specific augmented reference graph contained 0.073% erroneously mapped reads. Incorrectly mapped reads that had high mapping quality (MQ >10) were less frequent in the graph-based than linear alignments. Thus, a variation-aware reference may reduce the number of flawed genotypes arising from mapping errors that would remain unnoticed due to high mapping quality. Similar to findings in human genome graphs [18, 25], bovine variation-aware references did not improve the mapping of short reads that originate from low-complexity regions.

Our findings demonstrate that variant prioritization is key to accurate variation-aware read mapping. Based on investigations in four genetically distinct cattle breeds and human populations, we make three important observations: first, variation-aware references that contain random variants for which the allele frequency and haplotype phase in the target populations is unknown do not improve read mapping accuracy over linear references. Our previous study also showed that adding many random variants does barely affect sequence variant genotyping from reference graphs [12]. Adding random unphased variants increases the number of alternative alignment paths that are not necessarily biologically plausible haplotypes, thus increasing mapping ambiguity. Second, read mapping accuracy increases approximately linearly with the number of randomly sampled breed-specific variants being added to the genome graph. Similar findings in the four human population-specific augmented graphs confirm that this observation also holds for populations that are strongly enriched for rare alleles and singletons. Third, the highest mapping accuracy at tractable graph complexity can be achieved when variants filtered for allele frequency are added to the graph. Using variant prioritization approaches that are based on allele frequency, we observed the highest mapping accuracy at allele frequency thresholds between 0.01 and 0.10 in four cattle breeds and four human populations. In order to reduce the computational complexity of variation-aware read mapping, previous studies used arbitrarily chosen allele frequency thresholds to prioritize variants to be included in the graphs (e.g., 1% [22, 40], 5% [41], 10% [42]). Using fine-grained allele frequency inclusion thresholds, we find that the read mapping accuracy does not notably differ between the 0.01 and 0.1% thresh-

olds in most populations. Yet, mapping accuracy declined rapidly for the YRI-specific augmented graph when variants with frequency less than 10% were added indicating that the optimal inclusion threshold may vary across populations. Variant prioritization approaches that also take into account factors other than allele frequency [18] did not lead to further accuracy improvements in our study. Considering that most cattle breeds have an effective population size between 50 and 200 [43, 44], the vast majority of variants with allele frequency greater than 0.1 can be detected from a few sequenced key ancestor animals [13]. As a matter of fact, key ancestor animals have been sequenced for many cattle breeds [7, 45]. Thus, the construction of variation-aware reference structures that are informative for many cattle breeds is readily possible using, e.g., the sequence variant catalog of the 1000 Bull Genomes Project [7, 8].

A pan-genome graph that contained variants filtered for allele frequency across the four cattle breeds enabled almost similar accuracy improvements over the linear reference than breed-specific augmented graphs (Fig. 3.4b). Although the principal component analysis confirmed that the breeds considered in our study are genetically distinct populations, they share many common alleles. Moreover, compared to human populations, the proportion of rare alleles and singletons is low in cattle. The bovine pan-genome graph constructed in our study contained between 75.28 and 80.82% of the variants that were also added to the breed-specific augmented graphs. Instead of building many breed-specific graphs, the construction of a universal pan-genome graph is likely possible without notably compromising the accuracy of read mapping. This conclusion may hold for many species with genetically distinct sub-populations that share common alleles. Compared to the linear reference, the mapping accuracy was also significantly higher when reads from one breed were mapped to a genome graph that contained variants filtered for allele frequency in another somewhat related breed. Thus, the BSW-specific augmented whole-genome graph constructed in our study will likely improve read mapping accuracy over the linear reference and mitigate reference allele bias also for breeds other than BSW, FV, HOL, and OBV. Our BSW-specific augmented whole-genome graph is available at https://doi.org/10.5281/zenodo.3759712 [46]. In order to facilitate the construction of variation-aware reference structures, the entire workflow to establish whole-genome graphs is also available at https://github.com/danangcrysnanto/bovine-graphs-mapping.

The number of sequencing reads that aligned to the BSW-specific whole-genome graph with full identity increased considerably ($+13\%$) over the linear reference sequence at the cost of a slightly reduced (-0.72%) number of unique alignments. A two-step graph alignment approach that exploits a refined search space might reduce the number of multiple mappings in dense variation-aware graphs [32]. Compared to

a human whole-genome graph, the improvement in perfect mapping over the linear reference was slightly larger in our bovine whole genome graph (9.2%) [22]. However, the proportion of reads with perfect alignments (62.19%) was lower in our BSW-specific whole-genome graph, likely because it contained only sequences that were assembled to the 29 autosomes. The graph did not contain 269.77 Mb of the sex chromosomes, mitochondrial DNA, and 2180 unplaced contigs. A more sophisticated assembly of the bovine genome with increased continuity particularly at the sex chromosomes [4, 47] might serve as a backbone for an improved variation-aware genome graph.

In order to detect SNPs and Indels from the variation-aware reference graph using widely used sequence variant genotyping methods, we had to make the graph-based alignments compatible with linear coordinates. Thus, our assessment of sequence variant genotyping from the bovine whole-genome graph is based on surjected graph-based alignments. It is possible that converting graph-based to linear alignments compromises variant discovery. However, the accuracy and sensitivity of genotyping did not differ between graph-based and linear alignments indicating that our whole-genome graph facilitates accurate sequence variant (SNPs and small Indels) genotyping. It is worth noting that our analysis considered only SNPs that are located in well-accessible regions of the genome, thus possibly overestimating genotyping accuracy [48, 49]. A benchmark dataset that enables unbiased evaluation of sequence variant genotyping [50] is not available for the four cattle breeds considered in our study. Because approximately 90% of the considered SNPs were already included in the BSW-specific whole-genome graph, they can be detected and genotyped easily from graph-based alignments [17]. These variants can also be detected and genotyped accurately from linear alignments [12, 51].

As expected, bias towards the reference allele was less in graph-based than linear alignments particularly at variants that were included in the graph. Unbiased genotyping of heterozygous variants from graph-based alignments is possible because reads supporting alternate alleles align better to variation-aware than linear references. Thus, our bovine whole-genome graph offers an appealing novel reference for investigations that either rely on low-coverage sequencing or are sensitive to unbiased allele frequencies [16, 19, 52]. Because a benchmark dataset for an unbiased evaluation of sequence variant genotyping performance [50] is not available in cattle, our assessment was restricted to heterozygous variants that were identified from both linear and graph-based alignments. This set of variants is possibly enriched for variants that can be called confidently from linear alignments, thus underestimating the graph-based genotyping performance (e.g., [22]).

Our study has three limitations. First, variants used to construct the breed-specific augmented genome graphs might be biased because they were detected from linear alignments of short sequencing reads. Variant discovery from an independent variation-aware reference structure might allow for a more complete assessment of genetic variation [50]. Second, we used the Hereford-based linear reference sequence as backbone to construct breed-specific augmented reference sequences. However, the Hereford-based reference sequence might lack millions of basepairs that segregate in the four breeds considered in our study [53, 54, 55, 56]. These nucleotides are likely missing in the breed-specific augmented reference graphs constructed in our study. Accurate and continuous genome assemblies from BSW, FV, HOL, and OBV cattle are not available. All bovine genome assemblies that are available to date had been compiled from individuals that are distantly related to the breeds in our study [2, 4, 57]. Haplotype-resolved genome assemblies of cattle from different breeds will facilitate the construction of more informative genome graphs and make non-reference sequences and their sites of variation amenable to genetic investigations [2, 4]. Third, we did not investigate the impact of large sequence variation on sequence read mapping and variant genotyping performance because neither a high-quality benchmark set of large structural variants (cf. [58]) nor long-read sequencing data is available for the four cattle breeds considered. Adding insertion and deletion polymorphisms detected from short-read sequencing data did not lead to accuracy improvements in our study likely because structural variants detected from short reads are notoriously biased and incomplete [59]. Recent studies indicated that large structural variants can be identified accurately from genome graphs [25, 60, 61, 62]. Eventually, a bovine genome graph that unifies multiple breed-specific haplotype-resolved genome assemblies and their sites of variation might provide access to sources of variation that are currently neglected when short sequencing reads are aligned to a linear reference sequence [63, 64, 65].

3.4 Conclusions

We constructed the first variation-aware reference graph for *Bos taurus* that improves read mapping accuracy over the linear reference sequence. The use of this novel reference structure facilitates accurate and unbiased sequence variant genotyping. Our results indicate that the construction of a widely applicable bovine pan-genome graph is possible that enables accurate genome analyses for many diverged breeds.

3.5 Methods

Whole-genome sequencing data

We used short paired-end sequencing reads of 288 cattle from dairy (n = 82 Brown Swiss (BSW), n = 49 Holstein (HOL)) and dual-purpose (n = 49 Fleckvieh (FV), n = 108 Original Braunvieh (OBV)) breeds to detect variants that segregate in these populations. The average sequencing depth of the 288 cattle was 12.71-fold, and it ranged from 3.49 to 70.04. Most of the sequencing data were generated previously [7, 12, 13, 66, 67]. Accession numbers for all animals are available in Table S3.4.

We trimmed adapter sequences from the raw data and discarded reads for which the phred-scaled quality was below 15 for more than 15% of the bases using fastp [68]. Subsequently, the sequencing reads were aligned to the linear reference assembly of the bovine genome (ARS-UCD1.2, GCF_002263795.1) using *BWA mem* [34]. Duplicates were marked and the aligned reads were coordinate sorted using the Picard tools software suite (http://broadinstitute.github.io/picard) and Sambamba [69], respectively. We discovered and genotyped polymorphic sites from the linear read alignments using the Best Practices Workflow descriptions for multi-sample variant calling with *GATK* (version 4.1.0) [1]. Because a truth set of variants required for variant quality score recalibration (VQSR) is not available for Bos taurus, we followed the recommendations for sequence variant discovery and filtration when applying VQSR is not possible. Genotypes of the hard-filtered variants were subsequently refined, and sporadically missing genotypes were imputed with *BEAGLE* v4 [70] using the genotype likelihoods from the *GATK HaplotypeCaller* model as input values. Additional information on the sequence variant genotyping workflow and the expected genotyping accuracy can be found in Crysnanto et al. [12]. Nucleotide diversity was calculated in non-overlapping 10 kb windows separately for each breed using the π (nucleotide diversity) module implemented in the *vcftools* software [71].

We discovered and genotyped large structural variants ($>$ 50 bp) including insertions, deletions, inversions, duplications, and translocations in 82 sequenced BSW animals using *Delly* v0.7.8 [72] with the default settings. We retained only insertion and deletion variants that had been refined using split-reads (PRECISE-flag in the vcf file).

The principal components of a genomic relationship matrix constructed from whole-genome sequence variant genotypes were calculated using PLINK v1.9 [73]. The top principal components separated the animals by breeds, corroborating that the four

breeds are genetically distinct (Fig. 3.2a). To take haplotype diversity and different linkage disequilibrium phases across breeds into account, the sequence variant genotypes were phased for each breed separately using *BEAGLE* v5 [74].

Unless stated otherwise, our analyses included 541,876 biallelic SNPs and Indels that were detected on bovine chromosome 25. The *vg toolkit* version 1.17.0 "Candida" [22] was used for all graph-based analyses.

Haplotype-aware simulation of short sequencing reads

We simulated 10 million reads (150 bp) from reference haplotypes of one animal per breed that had sequencing coverage greater than 20-fold (see Table S3.4). Therefore, we added the phased sequence variants of each of the four animals to the linear reference to construct individualized reference graphs using *vg construct*. The haplotype-aware indexes of the resulting graphs were built using *vg index xg* and *gbwt*. *vg paths* and *vg mod* were used to extract the haplotype paths from the individualized reference graphs. Subsequently, we simulated 2.5 million paired-end reads (2×150 nt) from each haplotype using *vg sim*, yielding 10 million 150 bp reads per breed corresponding to approximately 35-fold sequencing coverage of bovine chromosome 25. The simulation parameter setting for read and fragment length was 150 and 500 (± 50), respectively. The substitution and indel error rate was 0.01 and 0.002, respectively, according to the settings used in Garrison et al. [22].

Read mapping to graphs augmented with variants filtered for allele frequency

The alternate allele frequency of 541,876 variants of bovine chromosome 25 was calculated separately for the BSW, FV, HOL, and OBV breeds using sequence variant genotypes of 82, 49, 49, and 108 sequenced cattle, respectively. We added to each breed-specific genome graph 20 sets of variants that were filtered for alternate allele frequency using thresholds between 0 and 1 with increments of 0.01 and 0.1 for frequency below and above 0.1, respectively. For instance, at an alternate allele frequency threshold of 0.05, the graph was constructed with variants that had alternate allele frequency greater than 5%. Alleles that were only detected in the four animals used to simulate reads (see above) were not added to the breed-specific augmented genome graphs.

The four breed-specific augmented genome graphs contained the same number of variants at a given allele frequency threshold to ensure that their density of information was similar. The number of variants added to the graphs was determined according to the breed in which the fewest variants were detected at a given allele frequency threshold. For the other three breeds, we sampled randomly from all variants that were

detected at the respective alternate allele frequency threshold. We indexed the breed-specific augmented graphs using *vg* index to obtain the topological (*xg*), query (*gcsa*), and haplotype (*gbwt*) index. Eventually, the simulated reads were aligned to the breed-specific augmented reference graphs using *vg map* with default mapping parameter settings considering both graph (*xg*, *gcsa*) and haplotype (*gbwt*) indexes.

To compare the accuracy of read mapping between variation-aware and linear reference structures, the simulated reads were also aligned to the linear reference sequence of bovine chromosome 25 using either *BWA mem* with default parameter settings or *vg map*. To enable linear mapping with *vg map*, we constructed an empty graph (without adding any sequence variants) from the linear reference sequence.

Read mapping to human population-specific augmented genome graphs

We downloaded phased whole-genome variants of 2504 individuals from phase 3 of the 1000 Genomes Project [28] as well as the corresponding reference sequence (g1k_v37; https://www.internationalgenome.org/category/reference/). We selected four populations which we considered to be genetically distinct based on the results of a principal components analysis and for which the number of individuals was similar to the number of individuals for the four cattle breeds, i.e., GBR (British in England and Scotland, European), YRI (Yoruba in Ibadan Nigeria, African), JPT (Japanese in Tokyo, East Asia), and STU (Sri Lankan Tamil, South Asia). The principal components were calculated from a genomic relationship matrix constructed using 81.27 million autosomal variants using the *PLINK* (v1.9) software [73]. Alternate allele frequency was calculated separately for the four populations for all variants of human chromosome 19. Nucleotide diversity was calculated with the *vcftools* software as detailed above. In order to construct population-specific augmented genome graphs, we used the reference sequence (g1k_v37) of human chromosome 19 as a backbone and added variants filtered for alternate allele frequency in the four populations (following the approach explained above). For each population, we constructed 20 graphs that contained between 3153 and 290,593 variants. We simulated 10 million paired-end reads for each population from reference haplotypes (as detailed above) of four selected samples (GBR: HG00096, YRI: NA18486, JPT: NA18939, STU: HG03642). The simulated reads were then mapped to the population-specific augmented genome graphs using the *vg toolkit*.

Read mapping to bovine breed-specific augmented graphs

We simulated 10 million reads from the haplotypes of a BSW animal (SAMEA6272105) and mapped them to variation-aware reference graphs that were constructed using variants (SNPs and Indels) filtered for alternate allele frequency greater than 0.03. Alle-

les that were only detected in SAMEA6272105 were excluded from the graphs. All graphs contained 243,145 variants. The number of variants was determined according to the HOL cattle breed because the lowest number of variants segregated at an alternate allele frequency greater than 0.03 in that breed. To investigate the utility of targeted genome graphs, we mapped the simulated BSW reads to a graph that contained variants filtered for allele frequency in BSW cattle. To investigate across-breed mapping, we mapped the simulated BSW reads to graphs that contained variants filtered for allele frequency in either FV, HOL, or OBV cattle. We also mapped the BSW reads to a bovine pan-genome graph that contained variants that were filtered for allele frequencies across the four cattle breeds. Additionally, we investigated the accuracy of mapping reads to a graph that was built from randomly selected variants. To construct the random graph, we randomly sampled from 2,294,416 variants that were detected on bovine chromosome 25 from animals of various breeds of cattle (http://www.1000bullgenomes.com/doco/ARS1.2PlusY_BQSR_v2.vcf.gz). The allele frequencies and haplotype phases of the random variants were not known. We constructed personalized graphs that contained only variants and haplotypes that were detected in the animals used for read simulation. The variation-aware graphs were subsequently indexed using *vg index* (see above). The simulated BSW reads were mapped to the different graphs using *vg map* (see above). The construction and indexing of graphs as well as read simulation and mapping were repeated ten times. We report in the main part of the paper the average values of ten replicates. This entire procedure was repeated with reads that were simulated from the haplotypes of FV (SAMN02671626), HOL (SAMN02671584), and OBV animals (SAMEA5059743).

Read mapping to consensus reference sequences

We modified alleles of the ARS-UCD1.2 linear reference sequence using the vcf2diploid tool [52]. We created two adjusted linear reference sequences for bovine chromosome 25:

- *major-BSW*: 67,142 nucleotides of the linear reference sequence were replaced with the corresponding major alleles detected in 82 BSW cattle.
- *major-pan*: 73,011 nucleotides of the linear reference sequence were replaced with the corresponding major alleles detected in 288 cattle from four breeds.

Ten million BSW reads were simulated (see above) and mapped to the original and modified linear reference sequences, as well as the corresponding variation aware reference structures using either *BWA mem* or *vg map* (see above) with default parameter settings. Since the replacement of reference alleles with Indels causes a shift in the reference coordinate system, we converted the coordinates of simulated reads between the

original and modified reference using a local instance of the *UCSC liftOver* tool [75] that was guided using a chain file produced by *vcf2diploid*. In order to prevent possible errors arising from coordinate shifts when reference nucleotides are either deleted or inserted at Indels, we repeated the analysis when only the alleles at SNPs were replaced.

Assessment of the read mapping accuracy

We used *vg stats* to obtain the number of nodes and edges, biologically plausible paths and length for each variation-aware reference graph. To assess the accuracy of graph-based alignment, we converted the Graph Alignment Map (GAM)-files to JavaScript Object Notation (JSON)-files using *vg view*. Subsequently, we applied the command-line JSON processor jq (https://stedolan.github.io/jq/) to extract mapping information for each read. Mapping information from linear alignments were extracted from the Binary Alignment Map (BAM)-files using the Python module *pysam* (version 0.15.3) (https://github.com/pysam-developers/pysam).

Using *vg annotate*, we annotated the simulated reads with respect to the linear reference coordinates and determined if they contained non-reference alleles. Comparing the true and mapped positions of the simulated reads enabled us to differentiate between correctly and incorrectly mapped reads. Following the approach of Garrison et al. [22] and taking into account the possibility that aligned reads may be clipped at Indels, we considered reads as incorrectly mapped if their starting positions were more than $k = 150$ ($k =$ read length) bases distant from true positions. The functional relevance genomic regions where the simulated reads originated from were determined based on the *Ensembl* annotation (version 99, [76]) of the bovine ARS-UCD 1.2 reference sequence. The coordinates of repetitive elements were determined based on RepeatMasker [77] annotation tables of the *UCSC* Genome Browser.

In order to assess mapping sensitivity and specificity, we calculated the cumulative TPR (true−positive rate) and FPR (false−positive rate) at different mapping quality thresholds and visualized it as pseudo-ROC (receiver operating characteristic) curve [22] using:

$$TPR_i = \frac{\sum_i^{60} TP_k}{n}$$

$$FPR_i = \frac{\sum_i^{60} FP_k}{n}$$

where TP_i and FP_i represent the number of correctly and incorrectly mapped reads,

respectively, at a given phred-scaled mapping quality threshold i (60, 50, 40, 30, 20, 10, 0), and n is the total number of reads mapped.

Read mapping and sequence variant genotyping from bovine whole-genome graph

Using 14,163,824 autosomal biallelic variants (12,765,895 SNPs and 1,397,929 Indels) that had alternate allele frequency greater than 0.03 in 82 BSW cattle, we constructed a BSW-specific augmented whole-genome graph. The Hereford-based linear reference sequence (ARS-UCD1.2) was the backbone of the graph. Specifically, we constructed graphs for each of the 29 autosomes separately using *vg construct*. Subsequently, *vg ids* was run to ensure that the node identifiers were unique in the concatenated whole-genome graph. We removed complex regions from the whole-genome graph using *vg prune* with default parameter settings and built the topological (*xg*) and query (*gcsa*) index for the full and pruned graph, respectively, using *vg index*. The haplotype paths of the 82 BSW cattle obtained using *BEAGLE* v5 (see above) were provided using a *gbwt* index.

To evaluate sequence variant genotyping from the whole-genome graph, we used between 122,753,846 and 904,047,450 million paired-end (2×150 bp) sequencing reads from 10 BSW cattle (SAMEA6163185, SAMEA6163188, SAMEA6163187, SAMEA6163177, SAMEA6163178, SAMEA6163176, SAMEA6163179, SAMEA6163183, SAMEA6163181, SAMEA6163182, Table S3.4) that had been sequenced at between 5.74 and 39.88-fold genome coverage. These animals were not part of the 82 BSW animals that were used to detect the variants that were added to the graph. We trimmed adapter sequences and removed reads that had more than 20% bases with phred-scaled quality less than 20 using *fastp* [68]. Subsequently, we mapped the pruned reads to either the BSW-specific augmented whole-genome graph or the linear reference sequence using either *vg map* while supplying both graph (*xg*, *gcsa*) and haplotype (*gbwt*) index to produce GAM files for each sample or *BWA mem*. To make the coordinates of the graph-based alignments compatible with linear reference coordinates, we converted the GAM- to BAM-files using *vg surject*. Variants were detected and genotyped from the surjected files using the multi-sample variant calling approach of either *GATK* [39], *Graphtyper* [40], or *SAMtools* [38], as stated above and detailed in [12].

In order to assess the read mapping accuracy from real sequencing data, we calculated the proportion of reads that aligned (i) perfectly and (ii) uniquely [18, 35, 78]. A read was considered to map perfectly if the edit distance was zero along the entire read (NM:0 tag in *BWA mem*-aligned BAM files; identity 1 in *vg map*-aligned GAM-files),

and without hard clipping (H tag) or soft clipping (S tag) in CIGAR string. A read was considered to map uniquely if either a single primary alignment was reported for the respective read or reads that had secondary alignments (XA tag in *BWA mem*-aligned BAM files; secondary_score > 0 in *vg map*-aligned GAM-files) had one alignment with phred-scaled mapping quality score of 60.

The sequenced BSW animals also had Illumina SNP BeadChip-derived genotypes at between 24,512 and 683,752 positions. The sequence variant genotypes were compared to microarray-called genotypes at corresponding positions to calculate recall/non-reference sensitivity, genotype concordance, precision, and non-reference discrepancy [1, 79]. The concordance metrics are explained in Fig. S3.17.

Snakemake workflows [80] for whole-genome graph construction, read mapping, and variant discovery are available in the *Github* repository (https://github.com/danangcrysnanto/bovine-graphs-mapping).

Assessment of reference allele bias

Reference allele bias was assessed at the heterozygous genotypes that had been detected in a BSW animal (SAMEA6163185) that had been sequenced at high (40-fold) coverage. Raw sequencing data were filtered as stated above and aligned to either the linear reference sequence or BSW-specific augmented genome graph using *BWA mem* and *vg map*, respectively. Sequence variant genotypes were discovered and genotyped from either surjected graph-based or linear alignments using the single sample variant calling approaches implemented in either *GATK HaplotypeCaller* or *SAMtools mpileup*. Variants were filtered using quality by depth (QD) > 10, mapping quality (MQ) > 40, and minimum read depth (DP) greater than 25 to ensure confident genotype calls and sufficient support for reference and alternate alleles at heterozygous genotypes. We considered only variants that were detected from both graph-based and linear alignments. At each heterozygous genotype, we quantified the number of reads supporting alternate and reference alleles using allelic depth information from the vcf files.

Availability of data and materials

The scripts and data used in this study are available via *GitHub* repository (https://github.com/danangcrysnanto/bovine-graphs-mapping) and archived in Zenodo (data: https://doi.org/10.5281/zenodo.3759712 [46] and scripts: https://doi.org/10.5281/zenodo.3763286 [81]). Raw sequencing read data of 298 cattle used for graph construction, evaluation of variant genotyping accuracy, and assessment of reference allele bias are available at the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena)

with study accession of PRJNA238491 [7], PRJEB28191 [12], and PRJEB18113 [67]. Detailed accession numbers for each sample are provided in Table S3.4.

References

[1] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43 (5):491, 2011.

[2] Sergey Koren, Arang Rhie, Brian P Walenz, Alexander T Dilthey, Derek M Bickhart, Sarah B Kingan, Stefan Hiendleder, John L Williams, Timothy PL Smith, and Adam M Phillippy. De novo assembly of haplotype-resolved genomes with trio binning. *Nature biotechnology*, 36(12):1174–1182, 2018.

[3] Karen H Miga, Sergey Koren, Arang Rhie, Mitchell R Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, Edmund Howe, David Porubsky, Glennis A Logsdon, et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, 585(7823):79–84, 2020.

[4] Edward S Rice, Sergey Koren, Arang Rhie, Michael P Heaton, Theodore S Kalbfleisch, Timothy Hardy, Peter H Hackett, Derek M Bickhart, Benjamin D Rosen, Brian Vander Ley, et al. Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. *Gigascience*, 9(4):giaa029, 2020.

[5] Sara Ballouz, Alexander Dobin, and Jesse A Gillis. Is it time to change the reference genome? *Genome biology*, 20(1):1–9, 2019.

[6] Beate D Scherf, Dafydd Pilling, et al. The second report on the state of the world's animal genetic resources for food and agriculture. 2015.

[7] Hans D Daetwyler, Aurélien Capitan, Hubert Pausch, Paul Stothard, Rianne Van Binsbergen, Rasmus F Brøndum, Xiaoping Liao, Anis Djari, Sabrina C Rodriguez, Cécile 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.

[8] Ben J Hayes and Hans D Daetwyler. 1000 bull genomes project to map simple and complex genetic traits in cattle: applications and outcomes. *Annual review of animal biosciences*, 7:89–102, 2019.

[9] Carole Charlier, Wanbo Li, Chad Harland, Mathew Littlejohn, Wouter Coppieters, Frances Creagh, Steve Davis, Tom Druet, Pierre Faux, François Guillaume, et al. NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. *Genome research*, 26(10): 1333–1341, 2016.

[10] Kim C Worley and Richard A Gibbs. Sequencing the bovine genome. *Bovine Genomics*, page 109, 2012.

[11] Christine G Elsik, Ross L Tellam, Kim C Worley, et al. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*, 324(5926):522–528, 2009.

[12] Danang Crysnanto, Christine Wurmser, and Hubert Pausch. Accurate sequence variant genotyping in cattle using variation-aware genome graphs. *Genetics Selection Evolution*, 51(1):1–15, 2019.

[13] Sandra Jansen, Bernhard Aigner, Hubert Pausch, Michal Wysocki, Sebastian Eck, Anna Benet-Pagès, Elisabeth Graf, Thomas Wieland, Tim M Strom, Thomas Meitinger, et al. Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. *BMC genomics*, 14(1):1–9, 2013.

[14] Jaemin Kim, Olivier Hanotte, Okeyo Ally Mwai, Tadelle Dessie, Salim Bashir, Boubacar Diallo, Morris Agaba, Kwondo Kim, Woori Kwak, Samsun Sung, et al. The genome landscape of indigenous African cattle. *Genome biology*, 18(1):1–14, 2017.

[15] L Koufariotis, BJ Hayes, M Kelly, BM Burns, R Lyons, P Stothard, AJ 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.

[16] Bryce Van De Geijn, Graham McVicker, Yoav Gilad, and Jonathan K Pritchard. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature methods*, 12(11):1061–1063, 2015.

[17] Benedict Paten, Adam M Novak, Jordan M Eizenga, and Erik Garrison. Genome graphs and the evolution of genome inference. *Genome research*, 27(5):665–676, 2017.

[18] Jacob Pritt, Nae-Chyun Chen, and Ben Langmead. FORGe: prioritizing variants for graph genomes. *Genome biology*, 19(1):1–16, 2018.

[19] Torsten Günther and Carl Nettelblad. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS genetics*, 15(7):e1008302, 2019.

[20] Mazdak Salavati, Stephen J Bush, Sergio Palma-Vera, Mary EB McCulloch, David A Hume, and Emily L Clark. Elimination of reference mapping bias reveals robust immune related allele-specific expression in crossbred sheep. *Frontiers in genetics*, 10:863, 2019.

[21] Jacob F Degner, John C Marioni, Athma A Pai, Joseph K Pickrell, Everlyne Nkadori, Yoav Gilad, and Jonathan K Pritchard. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–3212, 2009.

[22] Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature biotechnology*, 36(9):875–879, 2018.

[23] Cristian Groza, Tony Kwan, Nicole Soranzo, Tomi Pastinen, and Guillaume Bourque. Personalized and graph genomes reveal missing signal in epigenomic data. *Genome biology*, 21:1–22, 2020.

[24] Jouni Sirén, Erik Garrison, Adam M Novak, Benedict Paten, and Richard Durbin. Haplotype-aware graph indexes. *Bioinformatics*, 36(2):400–407, 2020.

[25] Glenn Hickey, David Heller, Jean Monlong, Jonas A Sibbesen, Jouni Sirén, Jordan Eizenga, Eric T Dawson, Erik Garrison, Adam M Novak, and Benedict Paten. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome biology*, 21(1):1–17, 2020.

[26] Meenu Bhati, Naveen Kumar Kadri, Danang Crysnanto, and Hubert Pausch. Assessing genomic diversity and signatures of selection in Original Braunvieh cattle using whole-genome sequencing data. *BMC genomics*, 21(1):1–14, 2020.

[27] Heidi Signer-Hasler, Alexander Burren, Markus Neuditschko, Mirjam Frischknecht, Dorian Garrick, Christian Stricker, Birgit Gredler, Beat Bapst, and Christine Flury. Population structure and genomic inbreeding in nine Swiss dairy cattle populations. *Genetics Selection Evolution*, 49(1):1–13, 2017.

[28] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526 (7571):68, 2015.

[29] Albert Tenesa, Pau Navarro, Ben J Hayes, David L Duffy, Geraldine M Clarke, Mike E Goddard, and Peter M Visscher. Recent human effective population size estimated from linkage disequilibrium. *Genome research*, 17(4):520–526, 2007.

[30] Hubert Pausch, Bernhard Aigner, Reiner Emmerling, Christian Edel, Kay-Uwe Götz, and Ruedi Fries. Imputation of high-density genotypes in the Fleckvieh cattle population. *Genetics Selection Evolution*, 45(1):1–10, 2013.

[31] C Hagger. Estimates of genetic diversity in the brown cattle population of Switzerland obtained from pedigree information. *Journal of Animal Breeding and Genetics*, 122(6):405–413, 2005.

[32] Ivar Grytten, Knut D Rand, Alexander J Nederbragt, and Geir K Sandve. Assessing graph-based read mappers against a baseline approach highlights strengths and weaknesses of current methods. *BMC genomics*, 21:1–9, 2020.

[33] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[34] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*, 2013.

[35] Harsh G Shukla, Pushpinder Singh Bawa, and Subhashini Srinivasan. hg19KIndel: ethnicity normalized human reference genome. *BMC genomics*, 20(1):1–17, 2019.

[36] Frederick E Dewey, Rong Chen, Sergio P Cordero, Kelly E Ormond, Colleen Caleshu, Konrad J Karczewski, Michelle Whirl-Carrillo, Matthew T Wheeler, Joel T Dudley, Jake K Byrnes, et al. Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet*, 7 (9):e1002280, 2011.

[37] Wolfgang Beyer, Adam M Novak, Glenn Hickey, Jeffrey Chan, Vanessa Tan, Benedict Paten, and Daniel R Zerbino. Sequence tube maps: making graph genomes intuitive to commuters. *Bioinformatics*, 35(24):5318, 2019.

[38] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25 (16):2078–2079, 2009.

[39] Ryan Poplin, Valentin Ruano-Rubio, Mark A DePristo, Tim J Fennell, Mauricio O Carneiro, Geraldine A Van der Auwera, David E Kling, Laura D Gauthier, Ami Levy-Moonshine, David Roazen, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, page 201178, 2017.

[40] Hannes P Eggertsson, Hakon Jonsson, Snaedis Kristmundsdottir, Eirikur Hjartarson, Birte Kehr, Gisli Masson, Florian Zink, Kristjan E Hjorleifsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, et al. Graphtyper enables population-scale genotyping using pangenome graphs. *Nature genetics*, 49(11): 1654, 2017.

[41] Sorina Maciuca, Carlos del Ojo Elias, Gil McVean, and Zamin Iqbal. A natural encoding of genetic variation in a Burrows-Wheeler transform to enable mapping and genome inference. In *International Workshop on Algorithms in Bioinformatics*, pages 222–233. Springer, 2016.

[42] Daehwan Kim, Joseph M Paggi, Chanhee Park, Christopher Bennett, and Steven L Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology*, 37 (8):907–915, 2019.

[43] SJG Hall. Effective population sizes in cattle, sheep, horses, pigs and goats estimated from census and herdbook data. *Animal*, 10(11):1778–1785, 2016.

[44] Grégoire Leroy, Tristan Mary-Huard, Etienne Verrier, Sophie Danvy, Eleonore Charvolin, and Coralie Danchin-Burge. Methods to estimate effective population size using pedigree data: Examples in dog, sheep, cattle and horse. *Genetics Selection Evolution*, 45(1):1–10, 2013.

[45] Aniek C Bouwman, Hans D Daetwyler, Amanda J Chamberlain, Carla Hurtado Ponce, Mehdi Sargolzaei, Flavio S Schenkel, Goutam Sahana, Armelle Govignon-Gion, Simon Boitard, Marlies Dolezal, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nature genetics*, 50(3):362–367, 2018.

[46] Danang Crysnanto and Hubert Pausch. Data for bovine graphs experiments (Version 1.1) [Data set], 2020. URL https://doi.org/10.5281/zenodo.3759712. Accessed 13 July 2020.

[47] Ruijie Liu, Wai Yee Low, Rick Tearle, Sergey Koren, Jay Ghurye, Arang Rhie, Adam M Phillippy, Benjamin D Rosen, Derek M Bickhart, Timothy PL Smith, et al. New insights into mammalian sex chromosome structure and evolution using high-quality sequences from bovine X and Y chromosomes. *BMC genomics*, 20(1):1–11, 2019.

[48] Heng Li. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20):2843–2851, 2014.

[49] Dorcus Kholofelo Malomane, Christian Reimer, Steffen Weigend, Annett Weigend, Ahmad Reza Sharifi, and Henner Simianer. Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. *BMC genomics*, 19(1):1–16, 2018.

[50] Heng Li, Jonathan M Bloom, Yossi Farjoun, Mark Fleharty, Laura Gauthier, Benjamin Neale, and Daniel MacArthur. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nature methods*, 15(8):595–597, 2018.

[51] Justin M Zook, Jennifer McDaniel, Nathan D Olson, Justin Wagner, Hemang Parikh, Haynes Heaton, Sean A Irvine, Len Trigg, Rebecca Truty, Cory Y McLean, et al. An open resource for accurately benchmarking small variant and reference calls. *Nature biotechnology*, 37(5):561–566, 2019.

[52] Joel Rozowsky, Alexej Abyzov, Jing Wang, Pedro Alves, Debasish Raha, Arif Harmanci, Jing Leng, Robert Bjornson, Yong Kong, Naoki Kitabayashi, et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology*, 7(1):522, 2011.

[53] Rachel M Sherman, Juliet Forman, Valentin Antonescu, Daniela Puiu, Michelle Daya, Nicholas Rafaels, Meher Preethi Boorgula, Sameer Chavan, Candelaria Vergara, Victor 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.

[54] Jayne Y Hehir-Kwa, Tobias Marschall, Wigard P Kloosterman, Laurent C Francioli, Jasmijn A Baaijens, Louis J Dijkstra, Abdel Abdellaoui, Vyacheslav Koval, Djie Tjwan Thung, René Wardenaar, et al. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nature communications*, 7(1):1–10, 2016.

[55] Lindsay A Holden, Meharji Arumilli, Marjo K Hytönen, Sruthi Hundi, Jarkko Salojärvi, Kim H Brown, and Hannes Lohi. Assembly and analysis of unmapped genome sequence reads reveal novel sequence and variation in dogs. *Scientific reports*, 8(1):1–11, 2018.

[56] Ruiqiang Li, Yingrui Li, Hancheng Zheng, Ruibang Luo, Hongmei Zhu, Qibin Li, Wubin Qian, Yuanyuan Ren, Geng Tian, Jinxiang Li, et al. Building the sequence map of the human pan-genome. *Nature biotechnology*, 28(1):57–63, 2010.

[57] Benjamin D Rosen, Derek M Bickhart, Robert D Schnabel, Sergey Koren, Christine G Elsik, Elizabeth Tseng, Troy N Rowan, Wai Y Low, Aleksey Zimin, Christine Couldrey, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience*, 9(3):giaa021, 2020.

[58] Mark JP Chaisson, Ashley D Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J Gardner, Oscar L Rodriguez, Li Guo, Ryan L Collins, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature communications*, 10(1):1–16, 2019.

[59] Can Alkan, Bradley P Coe, and Evan E Eichler. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5):363–376, 2011.

[60] Hannes P Eggertsson, Snaedis Kristmundsdottir, Doruk Beyter, Hakon Jonsson, Astros Skuladottir, Marteinn T Hardarson, Daniel F Gudbjartsson, Kari Stefansson, Bjarni V Halldorsson, and Pall Melsted. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature communications*, 10(1):1–8, 2019.

[61] Sai Chen, Peter Krusche, Egor Dolzhenko, Rachel M Sherman, Roman Petrovski, Felix Schlesinger, Melanie Kirsche, David R Bentley, Michael C Schatz, Fritz J Sedlazeck, et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome biology*, 20(1):1–13, 2019.

[62] Goran Rakocevic, Vladimir Semenyuk, Wan-Ping Lee, James Spencer, John Browning, Ivan J Johnson, Vladan Arsenijevic, Jelena Nadj, Kaushik Ghose, Maria C Suciu, et al. Fast and accurate genomic analyses using genome graphs. *Nature genetics*, 51(2):354–362, 2019.

[63] Zhongqu Duan, Yuyang Qiao, Jinyuan Lu, Huimin Lu, Wenmin Zhang, Fazhe Yan, Chen Sun, Zhiqiang Hu, Zhen Zhang, Guichao Li, et al. HUPAN: a pan-genome analysis pipeline for human genomes. *Genome biology*, 20(1):1–11, 2019.

[64] Doruk Beyter, Helga Ingimundardottir, Asmundur Oddsson, Hannes P Eggertsson, Eythor Bjornsson, Hakon Jonsson, Bjarni A Atlason, Snaedis Kristmundsdottir, Svenja Mehringer, Marteinn T Hardarson, et al. Long read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *BioRxiv*, page 848366, 2020.

[65] Heng Li, Xiaowen Feng, and Chong Chu. The design and construction of reference pangenome graphs with minigraph. *Genome biology*, 21(1):1–19, 2020.

[66] Christine F Baes, Marlies A Dolezal, James E Koltes, Beat Bapst, Eric Fritz-Waters, Sandra Jansen, Christine Flury, Heidi Signer-Hasler, Christian Stricker, Rohan Fernando, et al. Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. *BMC genomics*, 15(1):1–18, 2014.

[67] Sonja Hofstetter, F Seefried, Irene Monika Häfliger, Vidya Jagannathan, Tosso Leeb, and Cord Drögemüller. A non-coding regulatory variant in the 5′-region of the MITF gene is associated with white-spotted coat in Brown Swiss cattle. *Animal genetics*, 50(1):27–32, 2019.

[68] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890, 2018.

[69] Artem Tarasov, Albert J Vilella, Edwin Cuppen, Isaac J Nijman, and Pjotr Prins. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, 31(12):2032–2034, 2015.

[70] Brian L Browning and Sharon R Browning. Genotype imputation with millions of reference samples. *The American Journal of Human Genetics*, 98(1):116–126, 2016.

[71] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011.

[72] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M Stütz, Vladimir Benes, and Jan O Korbel. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, 2012.

[73] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):s13742–015, 2015.

[74] Brian L Browning, Ying Zhou, and Sharon R Browning. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3):338–348, 2018.

[75] Maximilian Haeussler, Ann S Zweig, Cath Tyner, Matthew L Speir, Kate R Rosenbloom, Brian J Raney, Christopher M Lee, Brian T Lee, Angie S Hinrichs, Jairo Navarro Gonzalez, et al. The UCSC genome browser database: 2019 update. *Nucleic acids research*, 47(D1):D853–D858, 2019.

[76] Andrew D Yates, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, et al. Ensembl 2020. *Nucleic acids research*, 48(D1):D682–D688, 2020.

[77] A Smith, R Hubley, and P Green. RepeatMasker Open-4.0. *RepeatMasker Open-4.0*, 2013.

[78] Adam M Novak, Glenn Hickey, Erik Garrison, Sean Blum, Abram Connelly, Alexander Dilthey, Jordan Eizenga, MA Saleh Elmohamed, Sally Guthrie, André Kahles, et al. Genome graphs. *bioRxiv*, 2017.

[79] Michael D Linderman, Tracy Brandt, Lisa Edelmann, Omar Jabado, Yumi Kasai, Ruth Kornreich, Milind Mahajan, Hardik Shah, Andrew Kasarskis, and Eric E Schadt. Analytical validation of whole exome and whole genome sequencing for clinical applications. *BMC medical genomics*, 7(1):1–11, 2014.

[80] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[81] Danang Crysnanto and Hubert Pausch. Scripts for bovine graphs experiments (Version 1.1), 2020. URL https://doi.org/10.5281/zenodo.3763286. Accessed 13 July 2020.

Chapter 4

Novel functional sequences uncovered through a bovine multi-assembly graph

Danang Crysnanto[1], Alexander S. Leonard[1], Zih-Hua Fang[1], Hubert Pausch[1]

[1] Animal Genomics, ETH Zurich, Zurich, Switzerland.

Published in *PNAS (2021) 118:20*

Contribution: I participated in conceiving the study, analysing the results and writing the manuscript. I wrote the multi-assembly graph pipelines.

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 transcriptomes from 56 cattle, we show that these non-reference sequences encode transcripts that hitherto remained undetected from the *Bos taurus* reference genome. We uncover putative 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 *Bos taurus* reference genome now align accurately to the pangenome. 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 putative genes from non-reference sequences 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.

4.1 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* [1]. 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 [2]. Recent method and sequencing technology developments have facilitated the assembly of multiple reference-quality genomes. The application of trio-binning [3] resulted in chromosome-scale haplotype-resolved assemblies for three taurine (Hereford, Angus, Highland) and one indicine (Brahman) cattle breeds, as well as for yak (*Bos grunniens*), a closely related species to domestic cattle [4, 5].

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 [6, 7, 8]. 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 [9, 10]. Moreover, reference-guided methods are blind to variations in sequences that are not present

Table 4.1: **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.

in the reference genome [11]. Recent estimates suggest that millions of bases are missing in mammalian reference genomes [12, 13], 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 [14, 15]. 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 [16, 17]. 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 [18, 19], 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 [20, 21].

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 [21] (Table 4.1). 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 transcripts that are differentially expressed as well as polymorphic sites that segregate within and between breeds of cattle.

4.2 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) [3, 4, 5] and their close relatives Brahman (*Bos taurus indicus*) [4] and yak (*Bos grunniens*) [5]. All assemblies, except for the Original Braunvieh breed, 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*, Note S4.1). The contig and scaffold N50 values of the six assemblies ranged from 21 to 80 Mb and 86.2 to 108 Mb, respectively Table 4.1.

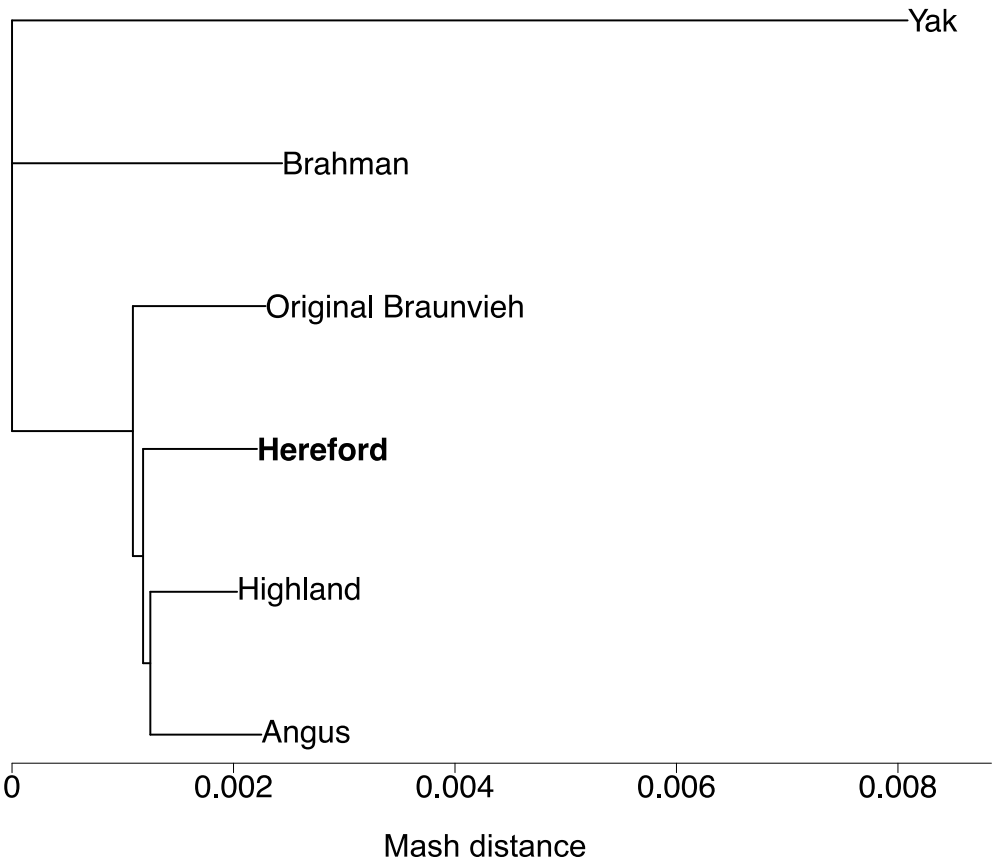


Figure 4.1: **Phylogenetic distance between six genome assemblies.**
A Mash-based phylogenetic tree derived from six bovine assemblies, including the current Hereford-based *Bos taurus* reference genome (**bold**). The yak assembly was used as the outgroup to root the tree during building.

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 [22] Fig. 4.1. 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.

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. S4.1 & S4.2). 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. S4.3). 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). 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*, Note S4.2).

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*, Note S4.3). However, the overall effect on the sequence content of the multi-assembly graph is relatively minor, with 68.72$\pm$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. S4.2). 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 were 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 [23] 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, $P = 6.56x10^{-3}$), microtubule cytoskeleton organization (10.48−fold, $P = 1.85x10^{-4}$), 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 S4.1). 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. 4.2). 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. 4.2, *SI Appendix*, Fig. S4.5). 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.

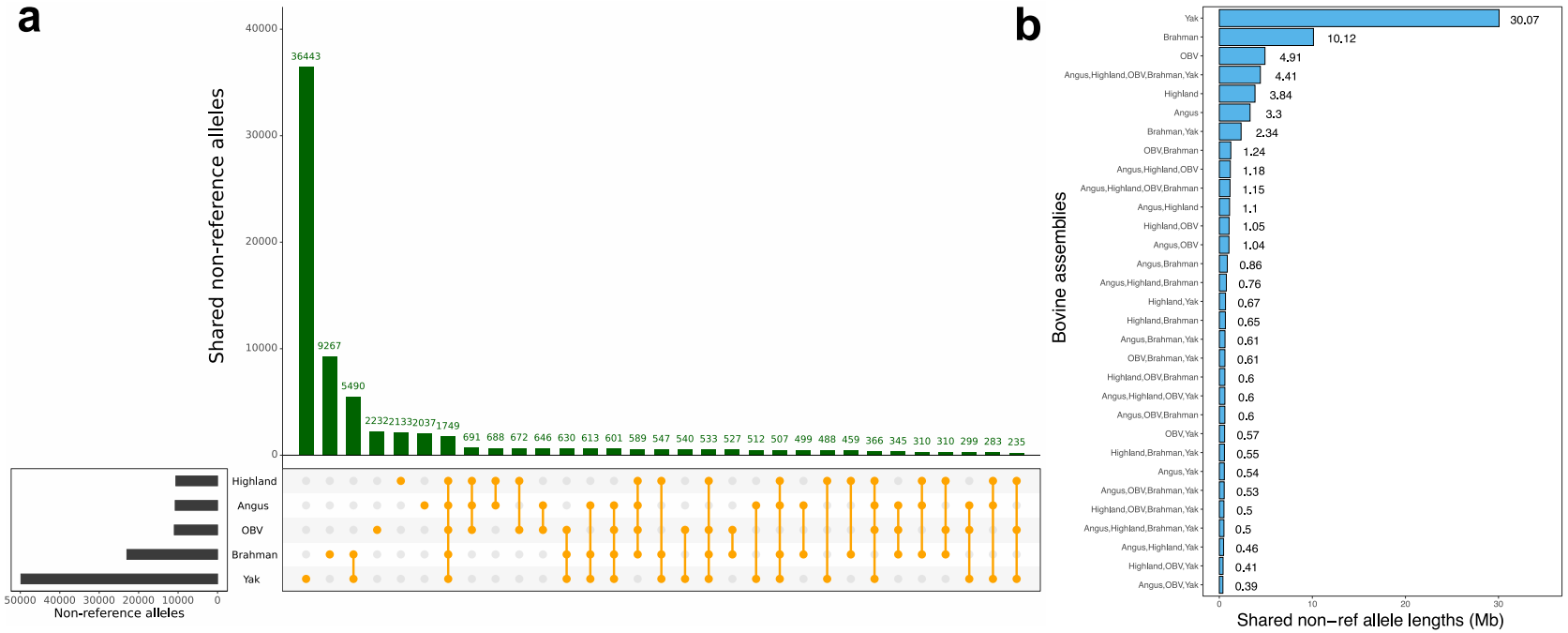


Figure 4.2: **Non-reference alleles detected across assemblies.**
Intersection of non-reference alleles **(a)** and cumulative length of the alleles **(b)** found in five assemblies when compared to ARS-UCD1.2. OBV : Original Braunvieh.

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. S4.6). 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. S4.7). 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. S4.8).

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 S4.2). On average, the transcript, coding sequence, and protein length of the complete gene models is respectively 4,742 bp, 794 bp, and 264 amino acids.

De novo transcript assembly from the non-reference sequences

As the two complementary gene prediction methods indicated that these non-reference 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. S4.9). Across the 40 samples, the overall mapping rate increased by 0.037%, which corresponds to approximately ∼18,000 reads for a paired-end RNA-seq dataset of 25 million reads. The mapping improvements were larger for samples with greater 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 [24], guided with gene models predicted by Augustus (see above), to assemble reads which aligned to non-reference sequences into 1,431 non-reference genes. Of these, 885 were expressed at TPM ≥ 1 in at least one breed, including 405 that were originally predicted by Augustus. We selected these 405 putative genes, supported by both *ab initio* prediction and *de novo* transcript assembly for further analyses.

Only 263 of the 405 putative genes were expressed at TPM ≥ 1 in Dominette, with BLASTP queries indicating they may mostly be divergent copies of ribosomal proteins or olfactory receptors. The remaining 142 genes were expressed at TPM ≥ 1 in Angus, Holstein, Jersey or Brahman cattle. Most were expressed in Brahman cattle (Fig. 4.3a), including 20 genes specific to this indicine breed. Among the taurine breeds, Angus contributed more genes than either Holstein or Jersey cattle. Approximately half of these genes, 68 of the 142, were common to all four nonreference breeds (Fig. 4.3b). 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 non-reference genes (*SI Appendix*, Fig. S4.10). The top hits suggest that these genes encode proteins related to: immune response (antigen-presenting glycoprotein, immunoglobulin, BOLA (Bovine Leukocyte Antigen), killer-T-cell, interferon, Ig-like lectin, CMRF35, MHC (Major Histocompatibility complex), 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).

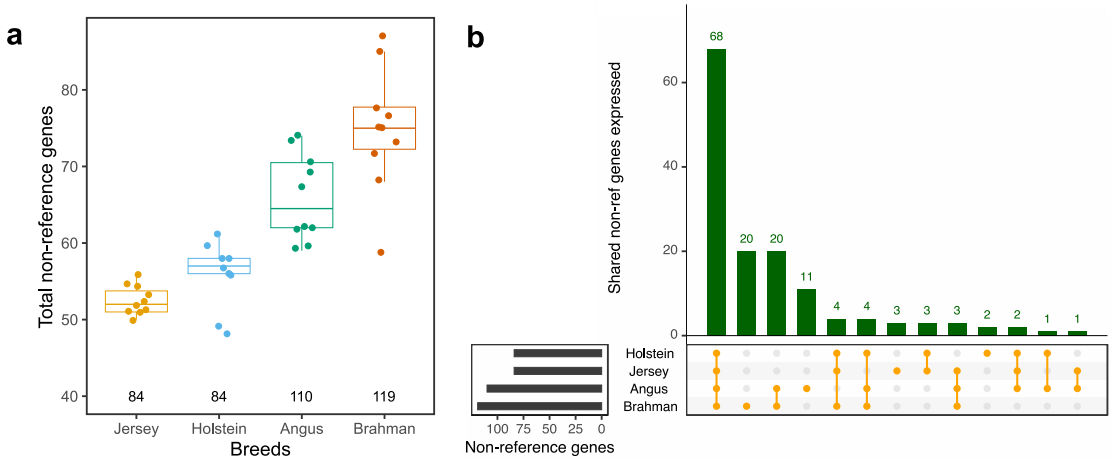


Figure 4.3: **Transcribed genes detected from non-reference sequences.**
(a) Number of non-reference genes expressed ≥1 TPM in liver tissue from taurine (Jersey, Holstein, Angus) and indicine (Brahman) cattle breeds. Each point represents the number of non-reference genes detected per animal. The number of distinct non-reference genes detected for each breed is indicated below the boxplots. **(b)** Expression of 142 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 [25]. 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 non-reference 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 ≥ 1 counts per million in at least eight samples for 13,085 genes, including 272 non-reference genes. We subsequently tested these genes for differential expression, finding 3,646 genes, including 36 non-reference genes, which were differentially expressed (FDR ≤ 0.05) between *Mycobacterium bovis*-infected and non-infected cattle (Fig. 4.4a). 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. [25] that were based on the previous UMD3.1 annotation (Pearson R $\log_2$ fold-change: 0.99) as well as with those from the standard ARS-UCD1.2 reference genome annotation (Pearson R $\log_2$ fold-change: 0.99, *SI Appendix*, Note S4.4).

Within the 36 differentially expressed non-reference 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. S4.11). Multidimensional scaling representations of transcript abundance estimates of the 36 differentially expressed genes separated *Mycobacterium bovis*-infected from non-infected cattle (Fig. 4.4b). BLASTX queries against a protein reference database provided additional support for 13 out of 36 differentially expressed genes (*SI Appendix*, Table S4.3). The top upregulated non-reference gene supported by the BLASTX query (4.04-fold increase, $P = 1.98x10^5$) 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 [26, 27, 28].

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 [29] 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 [30]; 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 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 ($P = 10^{-4}$) in infected cattle (*SI Appendix*, Table S4.3).

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 ∼100,000 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. S4.12). Similarly, to the transcriptome mapping, sequence reads from Dominette benefitted the least from the extended reference genome

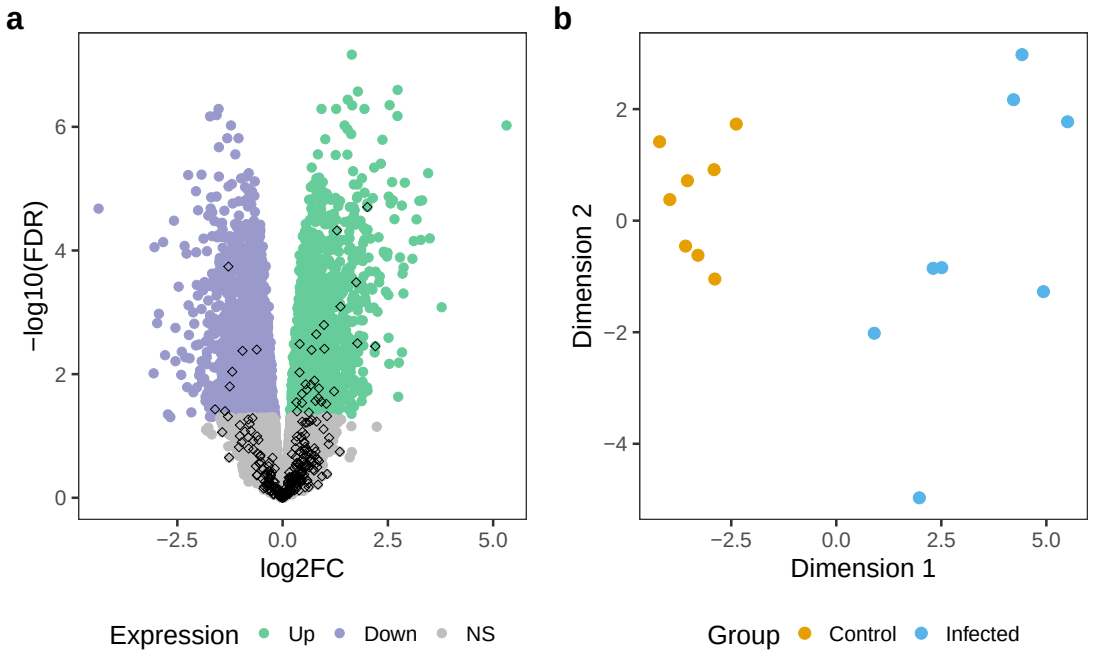


Figure 4.4: **Differentially expressed non-reference genes.**
(a) Volcano plot representing results from the differential expression analysis. Green and purple color indicates genes that are up- and downregulated (FDR $\leq$ 0.05), respectively, in peripheral blood leukocytes of *Mycobacterium bovis*-infected cattle. Diamond shapes indicate the 272 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.

(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 S4.4).

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 ($\geq$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. S4.13).

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 (Transition:Transversion) 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. 4.5a). 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. S4.14). 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. 4.5b,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 S4.5).

4.3 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 [12, 31]. 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 [21] 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

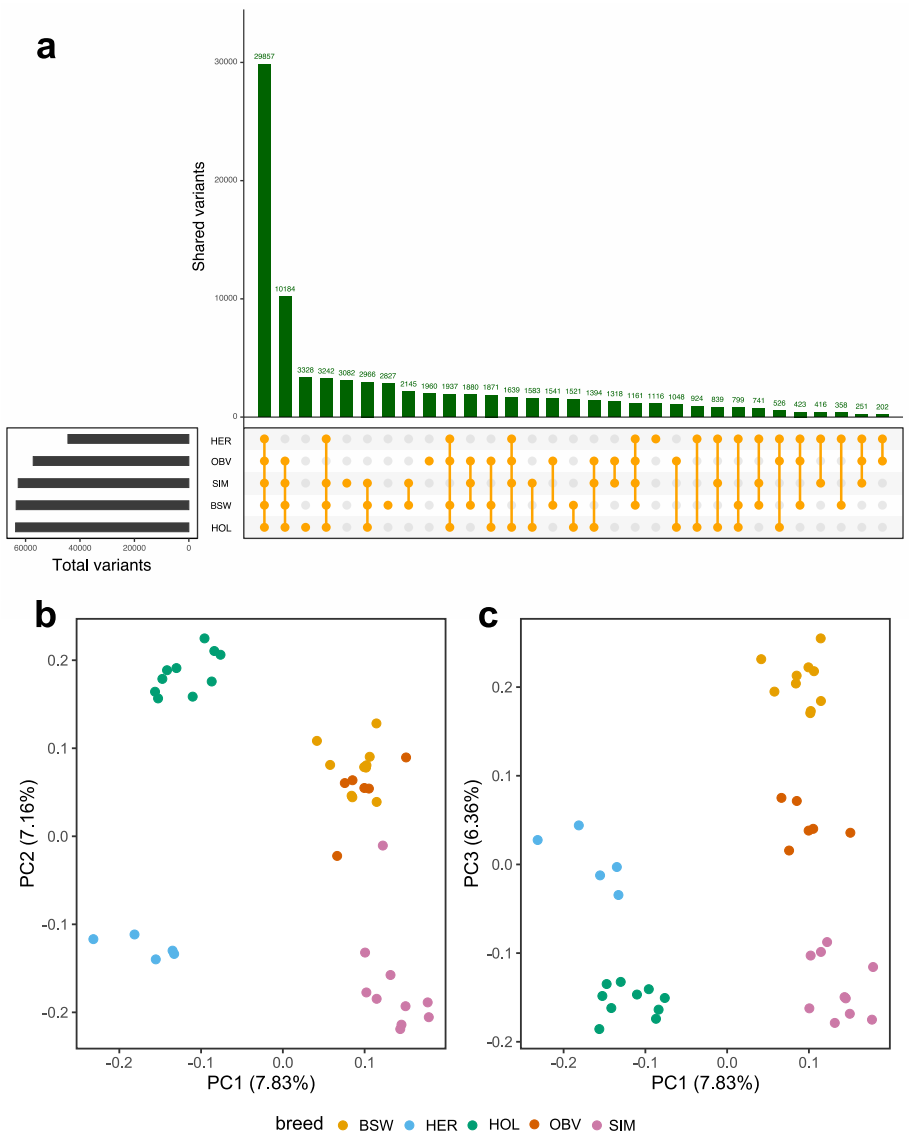


Figure 4.5: **Polymorphic sites detected from non-reference sequences in five breeds.** **(a)** Sharing of 83,250 variants across five taurine cattle breeds (BSW: Brown Swiss, HER: Hereford, HOL: Holstein, OBV: Original Braunvieh, SIM: Simmental). **(b, c)** The top three principal components (PC) of a genomic relationship matrix constructed from non-reference sequence variants separate the animals by breed.

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. Sophisticated algorithms facilitate the reference-free alignment of thousands of assemblies [32]. 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 not part of *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) [33] and goats (38.3 Mb) [34], 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 [35]. A pangenome constructed from domestic cattle and their extant relatives as recently proposed by the Bovine Pangenome Consortium [36] will reveal variants that were lost during domestication and the separation of cattle into specialized breeds [37]. 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 [38, 39, 40], possibly because cattle breeds have 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 [20]. However, as many genome analysis tools still rely on a linear coordinate system, we append the 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 [20]. 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 [21]. 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%) [12, 38]. Many non-reference sequences with repetitive elements were observed at immune gene complex loci, corroborating that these regions are highly repetitive [41]. The immune gene complex loci also contain many non-repetitive non-reference sequences suggesting great allelic diversity which may cause assembly problems [30], 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 other tissues. The discovery of distinct non-reference genes in an independent RNA sequencing dataset from peripheral blood leukocytes of Holstein cattle supports this hypothesis. Some of the non-reference 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 [42, 43]. 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 [44, 45], 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 [42, 46, 47]. Our findings show that more non-reference 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 non-reference 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 [25]. Thus, our multi-assembly graph uncovers func-

tionally 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 ∼100,000 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 appended non-reference sequences resolve misalignments. These findings agree well with results from species other than cattle, including goats, pigs, and humans [33, 34, 39]. In addition, we show that the non-reference 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 non-reference variants in putative coding sequences was only 1.83, they need to be scrutinized for false positives [48]. 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*, Note S4.2). 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*, Note S4.5). Our workflow provides tools to determine the origin of non-reference bases, derive structural variations from multi-assembly graphs, predict non-reference genes and append the non-reference 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 [49].

4.4 Methods

Construction of the multi-assembly graph

We used minigraph [21] (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)[4], Brahman (*Bos taurus indicus*, UOA_Brahman_1, GCF _003369695.1) [4], Highland *(Bos taurus taurus*, ARS_UNL_ Btau-highland_paterna _1.0_alt, GCA_009493655.1) [5], yak (*Bos grunniens*, ARS_UNL_BGru_maternal_1.0_p, GCA_009493645.1) [5]. Additionally, we constructed an assembly from a female Original Braunvieh calf (*Bos taurus taurus*) using PacBio high-fidelity (HiFi) reads (*SI Appendix*, Note S4.1). 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) [22]. 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 [50] as implemented in the R package ape (version 5.4) [51]. 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. S4.1).

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) [21] 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. S4.2):

- 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 ref-

erence 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/) [23] 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) [52] with preset $-xvg$ (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) [53] using the database of repetitive DNA elements from Repbase (release 20181026) [54].

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) [55]. 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 10^{-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) [56] 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 putative protein-coding sequences. To classify the domain and family of the non-reference 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 [57], 10 Brahman [58], 9 Holstein and 10 Jersey [59] cattle, as well as from Dominette - the animal used to construct the ARS-UCD1.2 reference genome [2]. 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) [60], 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.

Non-reference transcripts were assembled *de novo* using StringTie2 (version 2.1.1) [24] 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 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 new 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 [25] 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) [61] 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 [60]. 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 [62]. 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 rep-

resenting five breeds: Original Braunvieh, Brown Swiss, Holstein, Simmental [63], and Hereford (including Dominette, the animal used to construct the ARS-UCD1.2 reference genome) [2, 64] 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) [61] 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) [65] with default parameters. Duplicate reads were marked with Samblaster (version 0.1.24) [66].

We performed multi-sample variant calling (SNP and Indels) on the non-reference sequences using SAMtools (version 1.10) [67] and GATK (version v4.1.9.0) [68] as detailed in Crysnanto et al. [16]. 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: $QD < 2||QUAL < 30||FS > 200||ReadPosRankSum < -20.0||AN < 10$) [16]. A second independent variant discovery and genotyping approach was performed using SAMtools mpileup and bcftools call [67]. The resulting genotypes were subsequently hard-filtered according to parameters recommend by the 1000 Bulls Genomes project ($QUAL < 20||MQ < 30||DP < 10||AN < 10$) [7]. To create a consistent variant representation across both datasets, variants were normalized using vt (version 0.5) [69]. 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 [70].

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). 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 complete sample accessions, multi-assembly graph, non-reference sequences, non-reference genes, transcript abundances and sequence variants detected from non-reference sequences are available via Zenodo (https://doi.org/10.5281/zenodo.4385983) [71].

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) [73] 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

[1] Bovine Genome Sequencing and Analysis Consortium. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science (New York, NY)*, 324(5926):522, 2009.

[2] Benjamin D Rosen, Derek M Bickhart, Robert D Schnabel, Sergey Koren, Christine G Elsik, Elizabeth Tseng, Troy N Rowan, Wai Y Low, Aleksey Zimin, Christine Couldrey, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience*, 9(3):giaa021, 2020.

[3] Sergey Koren, Arang Rhie, Brian P Walenz, Alexander T Dilthey, Derek M Bickhart, Sarah B Kingan, Stefan Hiendleder, John L Williams, Timothy PL Smith, and Adam M Phillippy. De novo assembly of haplotype-resolved genomes with trio binning. *Nature biotechnology*, 36(12):1174–1182, 2018.

[4] Wai Yee Low, Rick Tearle, Ruijie Liu, Sergey Koren, Arang Rhie, Derek M Bickhart, Benjamin D Rosen, Zev N Kronenberg, Sarah B Kingan, Elizabeth 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.

[5] Edward S Rice, Sergey Koren, Arang Rhie, Michael P Heaton, Theodore S Kalbfleisch, Timothy Hardy, Peter H Hackett, Derek M Bickhart, Benjamin D Rosen, Brian Vander Ley, et al. Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. *Gigascience*, 9(4):giaa029, 2020.

[6] Kwondo Kim, Taehyung Kwon, Tadelle Dessie, DongAhn Yoo, Okeyo Ally Mwai, Jisung Jang, Samsun Sung, SaetByeol Lee, Bashir Salim, Jaehoon 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.

[7] Hans D Daetwyler, Aurélien Capitan, Hubert Pausch, Paul Stothard, Rianne Van Binsbergen, Rasmus F Brøndum, Xiaoping Liao, Anis Djari, Sabrina C Rodriguez, Cécile 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.

[8] L Koufariotis, BJ Hayes, M Kelly, BM Burns, R Lyons, P Stothard, AJ 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.

[9] Sara Ballouz, Alexander Dobin, and Jesse A Gillis. Is it time to change the reference genome? *Genome biology*, 20(1):1–9, 2019.

[10] Jacob Pritt, Nae-Chyun Chen, and Ben Langmead. FORGe: prioritizing variants for graph genomes. *Genome biology*, 19(1):1–16, 2018.

[11] Karen HY Wong, Walfred Ma, Chun-Yu Wei, Erh-Chan Yeh, Wan-Jia Lin, Elin HF Wang, Jen-Ping Su, Feng-Jen Hsieh, Hsiao-Jung Kao, Hsiao-Huei Chen, et al. Towards a reference genome that captures global genetic diversity. *Nature communications*, 11(1):1–11, 2020.

[12] Rachel M Sherman, Juliet Forman, Valentin Antonescu, Daniela Puiu, Michelle Daya, Nicholas Rafaels, Meher Preethi Boorgula, Sameer Chavan, Candelaria Vergara, Victor 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.

[13] Lynsey K Whitacre, Polyana C Tizioto, JaeWoo Kim, Tad S Sonstegard, Steven G Schroeder, Leeson J Alexander, Juan F Medrano, Robert D Schnabel, Jeremy F Taylor, and Jared E Decker. What's in your next-generation sequence data? An exploration of unmapped sequence reads from the bovine reference individual. *BMC genomics*, 16(1):1–7, 2015.

[14] Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature biotechnology*, 36(9):875–879, 2018.

[15] Hannes P Eggertsson, Hakon Jonsson, Snaedis Kristmundsdottir, Eirikur Hjartarson, Birte Kehr, Gisli Masson, Florian Zink, Kristjan E Hjorleifsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, et al. Graphtyper enables population-scale genotyping using pangenome graphs. *Nature genetics*, 49(11): 1654, 2017.

[16] Danang Crysnanto, Christine Wurmser, and Hubert Pausch. Accurate sequence variant genotyping in cattle using variation-aware genome graphs. *Genetics Selection Evolution*, 51(1):1–15, 2019.

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

[18] Jia-Ming Song, Zhilin Guan, Jianlin Hu, Chaocheng Guo, Zhiquan Yang, Shuo Wang, Dongxu Liu, Bo Wang, Shaoping Lu, Run Zhou, et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nature Plants*, 6(1):34–45, 2020.

[19] Birte Kehr, Anna Helgadottir, Pall Melsted, Hakon Jonsson, Hannes Helgason, Adalbjörg Jonasdottir, Aslaug Jonasdottir, Asgeir Sigurdsson, Arnaldur Gylfason, Gisli H Halldorsson, et al. Diversity in non-repetitive human sequences not found in the reference genome. *Nature Genetics*, 49(4):588–593, 2017.

[20] Glenn Hickey, David Heller, Jean Monlong, Jonas A Sibbesen, Jouni Sirén, Jordan Eizenga, Eric T Dawson, Erik Garrison, Adam M Novak, and Benedict Paten. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome biology*, 21(1):1–17, 2020.

[21] Heng Li, Xiaowen Feng, and Chong Chu. The design and construction of reference pangenome graphs with minigraph. *Genome biology*, 21(1):1–19, 2020.

[22] Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome biology*, 17(1):1–14, 2016.

[23] Huaiyu Mi, Anushya Muruganujan, Dustin Ebert, Xiaosong Huang, and Paul 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.

[24] Sam Kovaka, Aleksey V Zimin, Geo M Pertea, Roham Razaghi, Steven L Salzberg, and Mihaela Pertea. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome biology*, 20(1): 1–13, 2019.

[25] Kirsten E McLoughlin, Nicolas C Nalpas, Kévin Rue-Albrecht, John A Browne, David A Magee, Kate E Killick, Stephen DE Park, Karsten Hokamp, Kieran G Meade, Cliona O'Farrelly, et al. RNA-seq transcriptional profiling of peripheral blood leukocytes from cattle infected with Mycobacterium bovis. *Frontiers in immunology*, 5:396, 2014.

[26] Jodi L McGill, Randy E Sacco, Cynthia L Baldwin, Janice C Telfer, Mitchell V Palmer, and W Ray 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.

[27] Payal Damani-Yokota, Janice C Telfer, and Cynthia 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.

[28] Hilary E Kennedy, Michael D Welsh, David G Bryson, Joseph P Cassidy, Fiona I Forster, Christopher J Howard, Robert A Collins, and John 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.

[29] Saikou Y Bah, Thorsten Forster, Paul Dickinson, Beate Kampmann, and Peter 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.

[30] K Bakshy, D Heimeier, JC Schwartz, EJ Glass, S Wilkinson, Robin A Skuce, AR Allen, J Young, JC McClure, JB Cole, et al. Development of polymorphic markers in the immune gene complex loci of cattle. *Journal of Dairy Science*, 2021.

[31] Agnieszka A Golicz, Philipp E Bayer, Guy C Barker, Patrick P Edger, HyeRan Kim, Paula A Martinez, Chon Kit Kenneth Chan, Anita Severn-Ellis, W Richard McCombie, Isobel AP Parkin, et al. The pangenome of an agronomically important crop plant Brassica oleracea. *Nature communications*, 7(1): 1–8, 2016.

[32] Joel Armstrong, Glenn Hickey, Mark Diekhans, Ian T Fiddes, Adam M Novak, Alden Deran, Qi Fang, Duo Xie, Shaohong Feng, Josefin Stiller, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, 587(7833):246–251, 2020.

[33] Xiaomeng Tian, Ran Li, Weiwei Fu, Yan Li, Xihong Wang, Ming Li, Duo Du, Qianzi Tang, Yudong Cai, Yiming 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.

[34] Ran Li, Weiwei Fu, Rui Su, Xiaomeng Tian, Duo Du, Yue Zhao, Zhuqing Zheng, Qiuming Chen, Shan Gao, Yudong Cai, et al. Towards the complete goat pan-genome by recovering missing genomic segments from the reference genome. *Frontiers in genetics*, 10:1169, 2019.

[35] XB Qi, Han Jianlin, G Wang, JEO 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.

[36] Timothy Smith, Derek Bickhart, and Benjamin 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.

[37] Aamir W Khan, Vanika Garg, Manish Roorkiwal, Agnieszka A Golicz, David Edwards, and Rajeev 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.

[38] Adam Ameur, Huiwen Che, Marcel Martin, Ignas Bunikis, Johan Dahlberg, Ida Höijer, Susana Häggqvist, Francesco Vezzi, Jessica Nordlund, Pall 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.

[39] Peter A Audano, Arvis Sulovari, Tina A Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, AnneMarie E Welch, Max L Dougherty, Bradley J Nelson, Ankeeta Shah, Susan K Dutcher, et al. Characterizing the major structural variant alleles of the human genome. *Cell*, 176(3):663–675, 2019.

[40] Zhongqu Duan, Yuyang Qiao, Jinyuan Lu, Huimin Lu, Wenmin Zhang, Fazhe Yan, Chen Sun, Zhiqiang Hu, Zhen Zhang, Guichao Li, et al. HUPAN: a pan-genome analysis pipeline for human genomes. *Genome biology*, 20(1):1–11, 2019.

[41] John C Schwartz, Mark S Gibson, Dorothea Heimeier, Sergey Koren, Adam M Phillippy, Derek M Bickhart, Timothy PL Smith, Juan F Medrano, and John 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.

[42] Mingzhou Li, Lei Chen, Shilin Tian, Yu Lin, Qianzi Tang, Xuming Zhou, Diyan Li, Carol KL Yeung, Tiandong Che, Long 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.

[43] Graham M Hughes, Emma SM Boston, John A Finarelli, William J Murphy, Desmond G Higgins, and Emma C Teeling. The birth and death of olfactory receptor gene families in mammalian niche adaptation. *Molecular biology and evolution*, 35(6):1390–1406, 2018.

[44] Mitchell R Vollger, Philip C Dishuck, Melanie Sorensen, AnneMarie E Welch, Vy Dang, Max L Dougherty, Tina A Graves-Lindsay, Richard K Wilson, Mark JP Chaisson, and Evan E Eichler. Long-read sequence and assembly of segmental duplications. *Nature methods*, 16(1):88–94, 2019.

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

[46] Sean P Gordon, Bruno Contreras-Moreira, Daniel P Woods, David L Des Marais, Diane Burgess, Shengqiang Shu, Christoph Stritt, Anne C Roulin, Wendy Schackwitz, Ludmila Tyler, et al. Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nature communications*, 8(1):1–13, 2017.

[47] Agnieszka A Golicz, Philipp E Bayer, Prem L Bhalla, Jacqueline Batley, and David Edwards. Pangenomics comes of age: from bacteria to plant and animal applications. *Trends in Genetics*, 36(2):132–145, 2020.

[48] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43 (5):491, 2011.

[49] Jouni Sirén, Jean Monlong, Xian Chang, Adam M Novak, Jordan M Eizenga, Charles Markello, Jonas Sibbesen, Glenn Hickey, Pi-Chuan Chang, Andrew Carroll, et al. Genotyping common, large structural variations in 5,202 genomes using pangenomes, the Giraffe mapper, and the vg toolkit. *bioRxiv*, 2020.

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

[51] Emmanuel Paradis and Klaus Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3):526–528, 2019.

[52] Mikko Rautiainen and Tobias Marschall. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome biology*, 21(1):1–28, 2020.

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

[54] Weidong Bao, Kenji K Kojima, and Oleksiy Kohany. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile Dna*, 6(1):1–6, 2015.

[55] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nature methods*, 12(1):59–60, 2015.

[56] Mario Stanke and Stephan Waack. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, 19(suppl_2):ii215–ii225, 2003.

[57] Ruidong Xiang, Ben J Hayes, Christy J Vander Jagt, Iona M MacLeod, Majid Khansefid, Phil J Bowman, Zehu Yuan, Claire P Prowse-Wilkins, Coralie M Reich, Brett 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.

[58] LT Nguyen, A Reverter-Gomez, A Canovas, B Venus, A Islas-Trejo, SA Lehnert, JF Medrano, SS Moore, and MR Fortes. P1012 Liver transcriptome from pre versus post-pubertal Brahman heifers. *Journal of Animal Science*, 94(suppl_4):20–21, 2016.

[59] SM Salleh, Gianluca Mazzoni, P Løvendahl, and Haja 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.

[60] Daehwan Kim, Joseph M Paggi, Chanhee Park, Christopher Bennett, and Steven L Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology*, 37 (8):907–915, 2019.

[61] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890, 2018.

[62] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[63] Irene M Häfliger, Marlene Sickinger, Mark Holsteg, Leif M Raeder, Manfred Henrich, Siegfried Marquardt, Cord Drögemüller, and Gesine Lühken. An IL17RA frameshift variant in a Holstein cattle family with psoriasis-like skin alterations and immunodeficiency. *BMC genetics*, 21:1–10, 2020.

[64] Amy E Young, Tamer A Mansour, Bret R McNabb, Joseph R Owen, Josephine F Trott, C Titus Brown, and Alison L Van Eenennaam. Genomic and phenotypic analyses of six offspring of a genome-edited hornless bull. *Nature biotechnology*, 38(2):225–232, 2020.

[65] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 2013.

[66] Gregory G Faust and Ira M Hall. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*, 30(17):2503–2505, 2014.

[67] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25 (16):2078–2079, 2009.

[68] R Poplin, V Ruano-Rubio, MA DePristo, TJ Fennell, MO Carneiro, and GA Van der Auwera. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, 201178, 2017.

[69] Adrian Tan, Gonçalo R Abecasis, and Hyun Min Kang. Unified representation of genetic variants. *Bioinformatics*, 31(13):2202–2204, 2015.

[70] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome biology*, 17 (1):1–14, 2016.

[71] Danang Crysnanto, Alexander S. Leonard, Zih Hua Fang, and Hubert 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. Accessed 8 January 2021.

[72] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

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

Chapter 5

General Discussion

This thesis is the first to investigate the utility of genome graph-based sequence variant analysis approaches in the cattle genome. Thereby, this thesis offers a novel paradigm in the analysis of livestock genomes through accounting for genetic diversity in all analyses involved. The graph-based approaches introduced here facilitate thorough variation-aware genetic analyses because individual DNA sequences are compared to a set of haplotypes observed in the population rather than to the linear reference genome. Out of species with gigabase-sized genomes, only the human genome has been investigated with graph-based approaches. Within this thesis, I constructed the first genome graphs in a livestock species and performed different analyses to investigate the utility of genome graph-based approaches (Table 5.1).

Using three different variation-aware genome graphs, this thesis demonstrates that genome graphs outperform linear genomes across a suite of genomic analyses. Chapter 3 and 4 showed that graph-based reference structures enable improvements in mapping rate and resolve misalignments introduced by using linear coordinates. These mapping improvements facilitate accurate and unbiased genotyping. Chapter 4 showed that genotyping based on the graph-based alignment yielded a more balanced support of both reference and alternate alleles. These findings suggest that genome graphs are particularly useful for analyses which are sensitive to allelic dosage, such as allele-specific expression analyses. More importantly, the multi-assembly graphs constructed in Chapter 5 reveal abundant biologically-relevant sequences which are missing in the current ARS-UCD1.2 *Bos taurus* reference genome. Genome analyses are currently blind to the variations in these missing segments. Thus, the pangenome graph-based approach introduced in Chapter 5 makes this so far unused source of variations amenable to genomic analysis. Chapter 5 also provides an example how these hitherto neglected sequences enable a better biological understanding of the molecular underpinnings of phenotypic variations.

5.1 The application of graph genomes in cattle population

The feasibility of graph-based genomic methods on the cattle genome

Chapter 2 investigated the utility of region-specific graphs for the genotyping of polymorphic sites in the cattle genome. To this end, variation-aware graphs were constructed and augmented with variants discovered from linear read alignments of the same sequenced cohort. The workflow was established with a modified version of the *Graphtyper*

Table 5.1: **Three genome-graph approaches investigated in this thesis**

	Chapter 2	**Chapter 3**	**Chapter 4**
Graph types	Local (region-specific) variation graphs	Whole-genome (full) variation graphs	Multi-assembly genome graphs
Graph constructor	*Graphtyper*	*vg toolkit*	*minigraph*
Source of variations added to the graphs	Cohort-specific variants of 49 Original Braunvieh cattle	External (known) variants of 288 cattle from four breeds (Original Braunvieh (OBV), Brown Swiss, Fleckvieh, Holstein)	6 genome assemblies (OBV, Hereford, Angus, Highland, Brahman, Yak)
Application of the graphs	• Refined genotyping from linear alignments	• Variant prioritization • Genotyping from full-graph alignment • Assessment of reference bias	• Non-reference sequences extraction • Prediction of the novel genes • Transcription potential of non-reference sequences • Genetic variants in non-reference sequences
Benefits	• Computationally efficient	• Incorporate known (external variations) • Full-graph based alignment • More extensive downstream tools that can process	• Include structural variations diverged between assemblies • Computationally efficient
Limitations	• Need initial global read alignment by a linear mapper • Region-specific graphs • Limited to small variations discovered in the cohort	• Computationally expensive • Limited by small variations	• Not including small variations • Impacted by the graph backbone and order of assembly included • Limited downstream tools

software that was compatible with cattle chromosome complement. Although the pipeline is not full graph-based, due to its dependency on variants discovered from linear alignments and global read placement by a linear mapper, this simple graph-based implementation has outperformed the current-state-of-the-art linear mapping-based sequence variant genotyping approaches. Variant genotyping was highly accurate as indicated by multiple metrics including genotype concordance, non-reference sensitivity, non-reference discrepancy, and mendelian consistency in parent-offspring pairs, suggesting that graph-based methods are readily applicable for genomic analyses of the cattle genome.

Local graph genotyping is competitive with state-of-art linear-genome based methods

The computational requirement (both memory and time) is lower for *Graphtyper*-based than *GATK*–based variant discovery, which is a widely-applied workflow that also performs local read re-alignment. Therefore, the application of a region-specific graph-based method is also computationally competitive with methods that rely on a linear genome. In fact, *Graphtyper* has been applied to genotype thousands of human DNA samples demonstrating that it is applicable to genotype variants at the population scale [1, 2]. However, *Graphtyper* struggles with gaps and potential miss-assemblies that were numerous in the bovine UMD3.1 assembly. An additional analysis conducted in Chapter 2 provides evidence that these problems are mostly resolved when better reference genomes (e.g. ARS-UCD1.2) are used [3]. Thus, this thesis suggests that graph-based methods will benefit from the current influx of reference-quality assemblies across a wide-range of species.

Chapter 2 further demonstrated that genotypes produced by graph-based analysis are compatible with current state-of-the-art downstream tools. First, the genotype likelihoods produced by *Graphtyper* may serve as input and benefit from *Beagle* phasing and imputation, even yield higher genotype concordance compared to imputation using genotypes from a linear-mapping based methods. Secondly, we discovered more than 17 million variants from 49 key ancestor animals of the Original Braunvieh cattle breed using *Graphtyper* and used these genotypes to assess genomic diversity [4].

5.2 Prioritization of variants to be included in the graphs

Instead of relying on genetic variants from the same cohort, informative graphs can also be constructed using external variants. The study presented in Chapter 3 utilizes

a catalogue of variants discovered from close to 300 cattle from four major European cattle breeds to build variation-aware graphs. This approach showed that graph-based analysis can leverage, in principle, on a readily available variant database.

It is well known that variant prioritization is crucial to construct informative graph genomes [5, 6]. Chapter 3 showed that adding random unphased variants increases graph complexity without benefiting on read mapping accuracy. However, variant prioritization based on allele frequency increases the read mapping accuracy. The addition of variants with frequencies between 0.01 to 0.1 did not further improve the mapping accuracy. Thus, Chapter 3 provides a guideline to prioritize an optimal number of variants considered to create an informative yet computationally tractable graph genomes. The addition of variants beyond this threshold will not lead to further gains in mapping accuracy. The analysis presented in Chapter 3 revealed that this threshold is population- and species-specific. For example, the negative impact of rare variants on mapping accuracy was more pronounced in human than cattle populations, possibly due to a higher prevalence of low-frequency variants in the human genome.

Chapter 3 showed that improvements over linear references were similar to pangenome graphs and population (breeds) specific graphs. A similar finding has recently reported from a pan-human consensus reference [7]. This further suggests that building a unified cattle pangenome graph is feasible and likely preferred over generating multiple population-specific graphs. Due to low effective population size, a common set of variants to be added to the pangenome can be detected from few key ancestor animals, which have been compiled for instance by the 1000 Bull Genomes Project [8]. Chapter 3 shows that this observation holds for variation-aware graphs from four European cattle breeds. These breeds share more than 80% of the variations. Yet, it remains to be investigated if such a graph is also applicable to genetically-diverged breeds. Possibly, a set of prioritized variants from genetically-diverged breeds can be added to the graph while the slight increase in graph complexity is paid off with mapping improvements. The pervasive introgression and admixture across *Bos* species seem to indicate that this is a viable strategy [9]. Ideally a cattle pangenome graph includes variants from all global breeds (including understudied breeds), which will provide insight into an unbiased picture of the cattle diversity.

5.3 Investigation of inaccessible genetic variations with multi-assembly graphs

A multi-assembly graph provides a platform to investigate genetic variations

Beyond integrating small variations as performed in the Chapters 2 and 3, graph genomes provide a powerful framework to investigate large variations that segregate between individuals [2, 10, 11]. So far, only a few studies have attempted to characterize structural variations in the cattle genome [12, 13, 14, 15, 16]. However, large variations on overall affect longer genomic regions than small variations and have a more drastic effect on the gene functions [17, 18]. Thus, the contribution of structural variations to the genetic architecture of complex traits is likely to be under-appreciated in the cattle genome. For example, a recent study utilizing data from the 1000 Bull Genomes Project [15] found an overrepresentation of SV affecting expanding gene families that might provide novel and enhanced features. So far, most studies were limited to deletions or duplications of reference genome segments. Little is known about large sequences that segregate in the population but are absent in the reference genome.

Using a novel multi-assembly graph approach, the analyses presented in Chapter 4 integrated six assemblies from taurine cattle and their close relatives, Brahman and yak. The analysis recovered about 70 megabases of autosomal sequence not found in the ARS-UCD1.2 *Bos taurus* reference genome, containing thousands of structural variations. An independent alignment of long-read sequencing data validated nearly three-quarters of the structural variations in taurine and indicine breeds, giving confidence that that most of them are real variations rather than artifacts from mis-assembly. Moreover, it also indicates that these variations are prevalent across multiple cattle breeds. The *minigraph* algorithm applied to construct the multi-assembly graph does not consider variations smaller than 50 bp. Thus, the 70 Mb value reported in the Chapter 4 likely underestimates the full diversity between the six individual genomes. However, even with a such simplified graph, biologically relevant information can be retrieved, suggesting an enormous potential of applying pangenome graph approaches in the cattle population.

Segments not included in the reference genome are biologically-relevant

Sequences not included in the reference genome might contain variations contributing to the differentiation, adaptation, and evolution of breeds. The analyses presented in Chapter 4 revealed that polymorphic sites in non-reference sequences separate animals by breeds. Interestingly, some of these hitherto understudied sequences contain variations annotated with a high impact on the protein function. Thus, the use of a

pangenome graph expands our understanding of the bovine genome architecture.

A large amount of the non-reference sequences are specific to yak (30 Mb). These sequences might contain ancestral or wild-relative alleles that were lost during domestication of modern cattle or genomic sequences that shaped the evolutionary history of cattle. Further, the multi-assembly graph uncovered about 15 Mb non-reference sequences from individual taurine cattle genomes, indicating that the Hereford-based reference genome does not even accommodate the genetic diversity of closely-related breeds. This value aligns well with an estimate for a diverged single human genome that differs at about 16 Mb from the reference [19]. Intriguingly, the bovine pangenome revealed 4.4 megabases that were found in all assemblies but not in the *Bos taurus* reference genome. These sequences are likely mis-assembled or deleted in the reference animal (known as muted gaps [20]). Because the multi-assembly graph in Chapter 4 contains only a single indicine and yak animal, a more thorough analysis on breed-specific variations was not attempted. However, such an analysis seems to be warranted once more animals of the same breed have been added to the graph.

5.4 Functional characterization of the non-reference sequences

Chapter 4 further demonstrated that pangenome graphs facilitate the utilization of so far neglected sources of variations for functional genomic analysis.

Non-reference sequences are enriched with repeat elements

Repetitive elements account for the more than three-quarters of the non-reference sequences. More than half of these repeat sequences belong to LINE/L1. LINE/L1 is still active in the bovine genomes and transposition of these elements might lead to structural variations that alter gene structure or affect gene expression [15, 21, 22]. The high prevalence of these elements among the non-reference sequences suggests that this family of repetitive elements contributes to variable sequences across different bovine genomes that might shape the bovine evolution, although the details of the events need to be explored further.

Hundreds of transcriptionally active genes identified from non-reference sequences

Chapter 4 also reports on an array of analyses of the non-repetitive elements of the non-reference sequences that were conducted to uncover biologically-relevant sequences that are not included in the current *Bos taurus* reference genome. Specifically, 142 genes

were identified and expressed in breeds of cattle but not in the reference animals. Functional analysis indicated that the genes related to immune response are over-represented among the non-reference genes. Immune genes are highly polymorphic and contribute to genetic divergence and speciation [23]. Specifically, the Major Histocompatibility Locus at the cattle chromosome 23, which is one of the regions harboring an excess of variations in the multi-assembly graph, is a well-known hotspot of structural variations and one of the most diverse regions in the bovine genome [16].

Novel biological insights uncovered from the non-reference sequences

More importantly, Chapter 4 shows that these hitherto unused functionally-relevant sequences provide novel insights into biological processes. Specifically, the use of a pangenome helped to expand our understanding of the biology of *M. bovis* infections in cattle. Differentially expressed non-reference genes might contribute to the variability in response to infections. This information might be valuable for selecting disease-resistant animals. The top downregulated non-reference gene encoding LILRA5 (Leukocyte Immunoglobulin-like Receptor 5) resided in an unplaced contig in the linear reference genome. Because this gene is assembled completely in some of the assemblies used in constructing the multi-assembly graph, its placement to an autosomal region was possible, making it amenable for differential expression analysis. Presence and absence of LILRA5 has been reported among different yak assemblies [24]. Another top differentially expressed non-reference gene encodes workshop-cluster (WC) 1.1, which has been shown repeatedly to be affected by copy number variations [12, 13, 25, 26]. The WC gene family is unique to cattle, sheep, and pig [27]. It encodes pattern recognition in gamma delta T cells that its up-regulation might reduce the disease susceptibility. Previous studies have also reported on the transcriptome dynamic of the non-reference sequences, including genes that exhibit tissue-specific expression [28, 29, 30], which again corroborates that non-reference sequences contain functionally-relevant elements.

5.5 Construction of comprehensive and informative pangenome graphs for cattle

Building a comprehensive pangenome graph across global cattle breeds

The bovine multi-assembly graphs constructed in Chapter 4 revealed that about 6% of the pangenome is variable across assemblies. This value is similar to values reported in human, pig, and goat pangenomes [30, 31, 32] but considerably lower than those estimated for plant pangenomes [33, 34, 35], likely because plant genomes are affected by polyploidization, higher repeat content, and larger effective population size [36]. How-

ever, the size of the bovine pangenome still grows when more genomes are added, indicating that the analysis presented in Chapter 4 is not exhaustive and that the variable part of the bovine pangenomes might be higher than estimated. Adding distant assemblies also recovers more variable non-reference sequences. For example, including an assembly from yak recovered the largest amount of diverged sequences not yet characterized. Similarly, expanding the pangenome graph with a recently available gaur assembly[1] increased the size of the pangenome by 20 Mb including 13 megabases private to gaur (Fig. 5.1). Yet, it needs to be seen whether the pangenome will still grow when many more distant assemblies are added to the graph. To this end, this thesis provides the computational framework to construct and characterize the pangenome with a flexible number of input genomes (https://github.com/AnimalGenomicsETH/bovine-graphs).

The construction of a comprehensive pangenome representing global cattle diversity is the major aim of the Bovine Pangenome Consortium [37]. Chapter 4 provides an initial framework to build such a novel reference structure. A multi-assembly graph built from representative DNA sequences of different cattle breeds might also be a starting point for the construction of a comprehensive bovine pangenome graph that accommodates the full spectrum of genetic variation. The sample selection should be carefully considered to maximize diversity (e.g., some proposed methods [38, 39]). The optimal sample selection that includes comprehensive and diverse breeds, including under-represented and wild and undomesticated relatives of cattle, helps to characterize the pangenome of Bovinae that will reveal the true extent of genetic diversity. Since generating reference-quality genome assemblies at the population scale is still cost-prohibitive, an initial assembly-based strategy might be followed by augmenting the multi-assembly graph with known small variations. While small variations can be obtained readily from public databases, Chapter 3 showed that this step is ideally done by iterative augmentation of variations discovered directly from the graphs. Moreover, it seems important to integrate haplotype information in order to indicate biologically plausible allele combinations. The recent development of the so-called dynamic genome graph is appealing as it can be iteratively updated once additional genomes become available and be subdivided into smaller graphs facilitating detailed inspection on the population of interest [40].

Towards establishing highly informative graph genomes that integrate functional genomics resources

In addition to be comprehensive, graph genomes should be at least as informative as the reference sequence. In their current implementation, graph genomes appear as static

[1]The gaur assembly is available at the NCBI Genome with the accession of GCA_014182915.1 ARS_UOA_Gaur_1

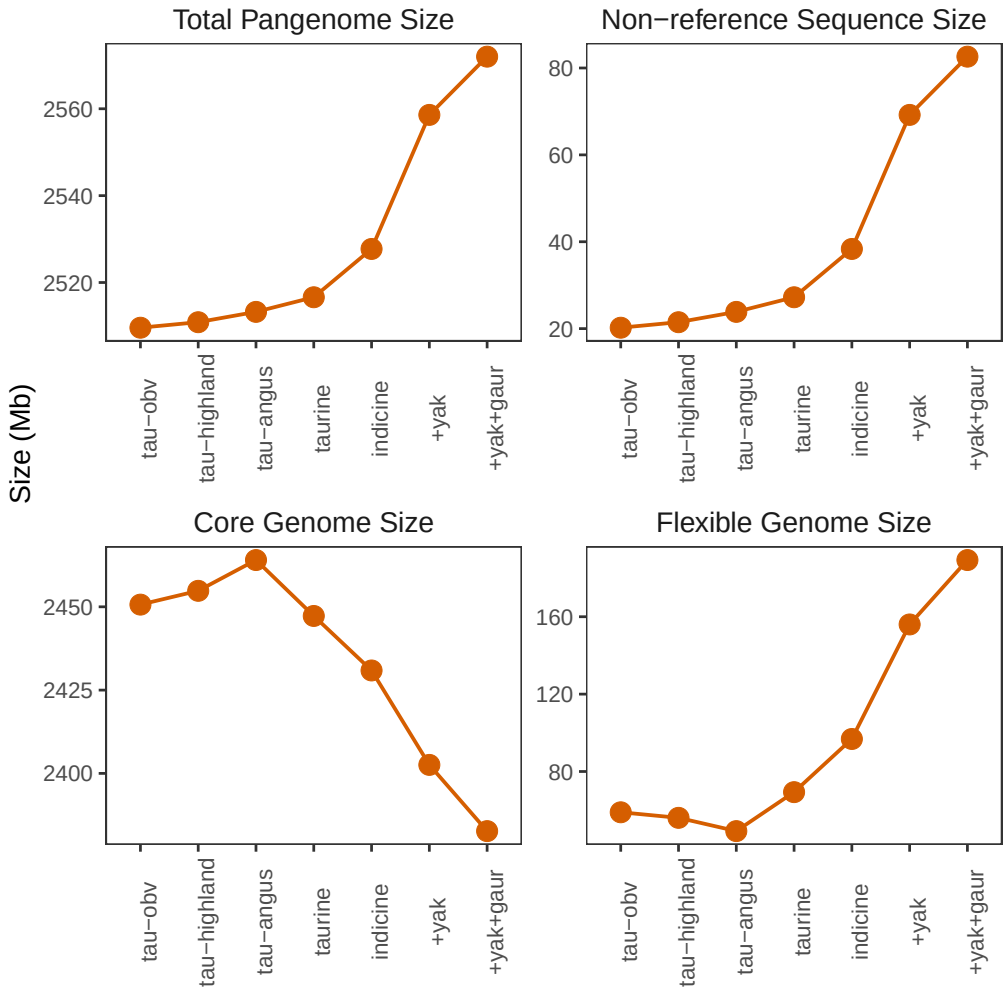


Figure 5.1: **Profile of the pangenome graph**
Pangenome graphs were constructed as in Chapter 4 (4 taurine breeds, 1 indicine breed, 1 yak) and complemented with a recently available gaur assembly. tau-*X* denotes a graph with taurine assemblies but excluding breed *X*. Taurine indicates a graph with four taurine breeds. TauInd is a graph consisting of taurine + brahman genomes. +yak and +yak+gaur indicate the TauInd graph with an addition of yak and yak and gaur assembly, respectively. The core and flexible genomes indicate sequences in pangenome shared in all and not in all breeds, respectively.

entities containing only DNA sequence information. However, the non-linear structure opens the possibility to include additional information in the graph other than DNA sequences, such as allele frequency, phenotype status of individuals (assigned to haplotypes traversing the nodes), or different layers of functional epigenomic data. As a proof of concept, an analysis presented in Chapter 4 showed that labelling the nodes to track sample information enables characterization the origin of the non-reference sequences. For this purpose, a strategy is needed that can compactly store metadata information from large number of samples in the graphs e.g. Sirén et al. [41].

Recent studies have examined the possibility of building pangenome graphs that contain information beyond DNA sequences. Sibbesen et al. [42] showed that adding splice information into a pangenome graph may outperform state-of-the-art RNA sequencing alignment and variant genotyping from linear reference genomes for the analysis of allele-specific expression. Hokin et al. [43] added genotype information and assigned the disease status of samples, enabling an association study directly from genotype graphs (termed as *Pangenome Wide Association Study*). They found regions harboring complex variations that are associated with complex traits that were missed by traditional GWAS that rely on variants called from linear alignments. On the same line, Kaye and Wasserman [44] proposed a Genome Atlas as an informative pangenome representation in which the graph's nodes are labelled with an unique ID that assigned functional metadata. The connections between nodes are not limited by sequence proximity, e.g. nodes could also be linked because of sharing annotation, which can be flexibly tuned.

In such an implementation, the pangenome graph can be used as a reference structure for multiple layers of epi-genomics data. This approach is readily feasible in many livestock species for which large amounts of functional omics data have been generated [45]. Overall, these graph resources will be highly valuable for the livestock genomics community to catalogue the global livestock diversity in order to perform comprehensive comparative genomics or even to identify beneficial alleles that are relevant for adaptation to future environmental changes.

5.6 Challenges to construct comprehensive pangenome graphs

Impact of the genome assembly quality on the reliability of the graph-based analysis

The quality of the assemblies being integrated into the graphs is important. Chapter 2 showed that in regions with unresolved segmental duplications, the graph computation time increased substantially, indicating that the incomplete or flawed assembly of such regions increases graph complexity. Particularly, with the *minigraph* approach, one assembly is used as the backbone of the graph and the pangenome is iteratively built by augmenting other genomes to this backbone. Therefore, the quality of the backbone assembly is critical for accurate and complete pangenome representation, especially to retrieve the true sequences diverged across animals rather than technical artifacts due to the incomplete assembly. Chapter 4 and Fig. 5.2 demonstrated that the use of the Highland or OBV assembly as a backbone leads to a larger pangenome and less non-reference sequences detected from the other assemblies. This finding possibly suggests that these

two assemblies are more complete than other assemblies which seems to corroborate findings from the initial analysis of the Highland assembly [46].

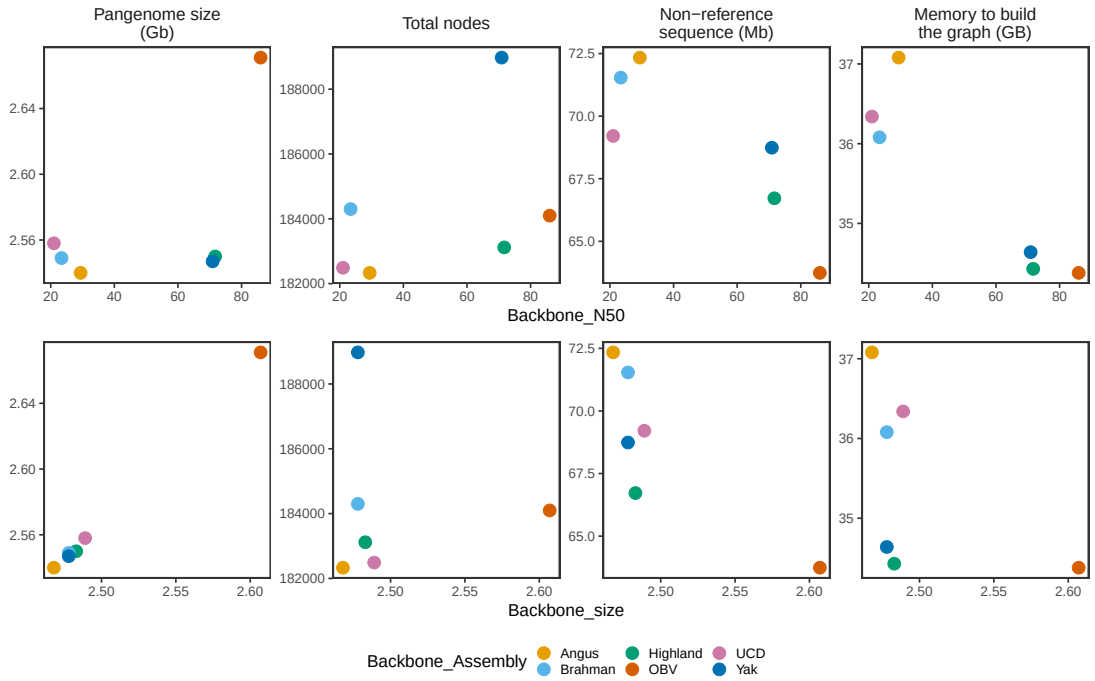


Figure 5.2: **Correlation between the backbone assembly quality and the profile of the pangenome graph**
A colored dot represents the backbone assembly from which that the graph was built from. N50 represents the assembly contiguity with a higher number reflects a more contiguous assembly.

Additionally, the pangenome will benefit from the use of haplotype-resolved assemblies. The mapping algorithm in *vg* (Chapter 3) utilizes phasing information to prioritize read alignments conforming to biologically plausible haplotypes, thus reducing mapping ambiguity. Moreover, haplotype switches in collapsed assemblies might limit the interpretation of long-range information encoded in the paths. The benefit of using haplotype-resolved over primary assemblies has recently been shown in human pangenome, that phasing information helps to infer the genotypes of low-coverage regions facilitating imputation-like strategies performed directly from the graphs [47, 48].

Technological advancements in long-read sequencing particularly with the development of the highly-accurate circular consensus sequencing [49] facilitate the cost-effective production of high-quality genome assemblies. The multi-assembly graph constructed in Chapter 4 integrated a bovine genome assembly that was generated using HiFi reads. There were 104-116 Mb sequences from the HiFi-based Original Braunvieh assembly not included in the graphs when other assemblies were the backbone of the

multi-assembly graph. These sequences are primarily composed of DNA satellites, suggesting that HiFi-reads enable a better assembly of so far difficult-to-assemble regions in the cattle genome, such as telomeric and centromeric sequences. Due to lower quality of X, Y chromosomes, and unplaced contigs, the analyses in this thesis were restricted to the autosomes. The high quality of the novel HiFi-based assemblies now provides an opportunity to also investigate highly polymorphic or repeat regions and sex chromosomes [50, 51], thus revealing a more accurate and complete pangenome.

Scalable approaches for building comprehensive pangenome graphs across hundreds of assemblies

Beyond generating assemblies, scalable approaches that can efficiently construct and characterize a pangenome from many assemblies are needed. The pangenome graph in Chapter 4 was built computationally efficient using the *minigraph*. However, it is unable to represent the full spectrum of genomic diversity as it included only structural variations longer than 50 bp (Table 5.2). Thus, to exploit the full potential of the pangenome, full graph models that can accommodate all haplotypes of the individuals in the population and their sites of variations, are required. The development a more comprehensive genome graph such as *pggb* (https://github.com/pangenome/pggb) or *cactus* [52] is promising, because these tools can perform reference-free multi-genome alignment to generate a full graph containing both short and long variations. Utilizing a full pangenome graph as implemented in *pggb* or *cactus* may uncover more variable and non-reference sequences than *minigraph*, as an initial analysis revealed (Table 5.2). However, this approach is computationally demanding for whole-genome applications, likely because the resulting graph is more complex due to many small nodes (Table 5.2). Additionally, without anchoring the pangenome on well-established reference coordinates, complex and highly repetitive genomic regions tend to form highly tangled regions in the graphs which are difficult to interpret [36]. Therefore, a thorough analysis to assess differences between various multi-assembly graph implementations is required. A strategy proposed by the Human Pangenome Reference Consortium to integrate 350 diverse human assemblies is supposed to define the first *de facto* standard in the field.

Stable ecosystem of tools and adoption of graph genomes in the genomics community

A stable framework to efficiently store, modify, and handle complex graphs for routine genomic analyses remains to be developed. Many analyses presented in this thesis were not fully graph-based as they depend on the graph's transformation into linear coordinates to make the graph amenable to current tools. For example, sequence variant genotyping in Chapter 3 was based on projections onto reference sequence paths.

Table 5.2: **Comparison of methods to build the multi-assembly graphs.**
Ref nodes refer to the node contained sequences from the ARS-UCD1.2 reference genome and non-ref nodes contained sequences from the other breeds but not in the reference assembly. Core nodes and flexible represent nodes with sequences shared in all breeds and not in all breeds, respectively. R-R, R-NR, NR-NR denote edges connecting ref-ref nodes, ref-non-ref nodes, and non-ref-non-ref nodes respectively.

Parameter	Unit	Minigraph pipeline	pggb pipeline	Cactus pipeline
Average memory	Gb	1.7	12.5	11.6
CPU time	hours	0.05	7.23	10.98
All nodes	n	1,136	804,723	843,177
Total length	bp	42,671,567	43,495,189	43,583,632
Average Node length	bp	37562	54	51
Reference nodes	n	770	534,993	545,952
Total length ref nodes	bp	42,350,435	42,316,615	42,350,435
Non-reference nodes	n	366	269,730	297,225
Total length non-ref nodes	bp	321,132	1,178,574	1,233,197
Total edges	n	1,630	1,384,318	1,142,667
R-R edges	n	904	706,505	570,277
R-NR edges	n	705	631,949	524,483
NR-NR edges	n	21	45,864	47,907
Node to Edge Ratio	ratio	1.43	1.72	1.35
Core nodes	n	441	270,044	274,134
Core length	bp	42,071,986	41,546,904	41,577,514
Flexible nodes	n	695	534,679	569,043
Flexible length	bp	59,9581	1,948,285	2,006,118
Core proportion	%	98.59%	95.52%	95.39%
Flexible proportion	%	1.41%	4.48%	4.60%

* The multi-assembly graph was built from chromosome 25 of 4 taurine assemblies (Hereford, Angus, Highland, Original Braunvieh) and 1 indicine (Brahman) assembly. The minigraph pipeline was implemented as in the Chapter 4. The pggb pipeline was run with the recommended parameters (`-s 100000 -p 90 -n 10`, https://github.com/pangenome/pggb) and the cactus pipeline was based on the suggested within-species pangenome pipeline (https://github.com/ComparativeGenomicsToolkit/cactus). Both pggb and cactus pipeline implement a full graph model that includes complete variations, meanwhile minigraph only considers variations longer than 50 bp.

Thus, the reported improvement of genotyping accuracy over linear alignments might be more pronounced once all analyses are performed directly on the graph. Moreover, multiple fragmented graph implementations for specific use cases with poor interoperability among the tools hamper the development of widely-accepted graph-based genomics approaches. For example, due to different specifications, the graph structure from *minigraph* (Chapter 4) is not compatible with extensive graph operations that have been implemented in *vg* (Chapter 3). As the graph genome framework reaches maturity, the genomics community ideally will agree on a widely-accepted standard that ensures long-term stability, similar to tools development for the linear reference genome (e.g., *BAM*, *VCF*) [53]. A wider adoption of graph-based analysis will naturally foster the development of efficient tools to process these new richer reference structures (e.g., [54, 55]).

Reluctance of the genomics community to transition to graph-based approaches results in slower adoption of the methods. It is clear that a transition to a graph-based reference will require a new paradigm and huge efforts to adjust downstream tools that rely on a linear representation of the genome. Additionally, instead of a ready-to-use linear genome, graph genomes need a more involved construction process (see Chapter 3 Methods). However, this thesis clearly showed that the increase in the analysis complexity is outweighed by novel intriguing insights. Moreover, graph-based structures are required to compactly integrate an ever-increasing amount of genomic resources. To increase the appeal of graph-based genomes, it is highly desirable to have a robust graph-genome-based visualization for interactive explorations of the graph structure (e.g. coloring paths according to breeds that might help pinpoint segments differentiating between lineages). However, implementations that can accommodate across zoom levels and finer details are still not fully operable [56, 57, 58]. In the short term, graph-based approaches may be used for intermediate steps which are hidden from the end user, i.e., the analysis is performed on graphs but the output is projected back to the linear space. Thus, graphs might supplement rather than completely replace linear genomes [11, 59, 60, 61]. The *Graphtyper* pipeline as implemented in Chapter 2 follows this paradigm.

Outlook

This thesis presents the first implementations of graph-based reference structures in cattle. Pangenome graphs provide a framework for accurate, unbiased and complete representation of sequence variation within a species, including those that are missed in routine genomic analysis because of the incompleteness of a single linear reference genome. The graph-based approaches presented in this thesis may serve as a starting point for many analyses that have either not yet been possible or were less accurate due to using the linear reference sequence. Importantly, this thesis provides a computational framework to integrate and exploit an ever-increasing amount of genomic resources (including genome assemblies and their sites of variation). On the one hand, this is relevant for collaborative initiatives to catalogue the complete species diversity such as the Bovine Pangenome Consortium. On the other hand, the computational framework developed and implemented in this thesis is broadly applicable to many species. Importantly, comprehensive comparative genomic analyses on the pangenome graph might help identify genomic features that are conserved or diverged between breeds and species that might underly the adaptive traits or domestication which can then be exploited to accelerate genetic progress [45, 62].

Potential applications of genome graphs to enhance livestock genomics are discussed below

Unbiased genomic analyses using genome graphs

Genome graph approaches provide an opportunity to revisit genomic analyses that suffer from reference bias, such as allele-specific expression (ASE), which attempts to detect gene expression imbalance between paternal and maternal-derived alleles [63]. ASE is known to be pervasive in the cattle genome [64] and affects complex traits in livestock such as meat quality [65, 66]. The current ASE detection method primarily relies on RNA-sequencing alignments to a linear genome which is prone to reference allele bias. To overcome this issue, reference sequences are commonly modified to match the alleles from the transcriptome [67]. However, this strategy is imperfect as it needs two rounds of read mapping, is limited to SNPs, and can still underestimate the overall expression levels [68]. Genome graphs can represent both paternal and maternal alleles in a coherent structure that can mitigate this issue. The split-read mapping capability that has been recently implemented in the *vg toolkit* [42] facilitates direct mapping of transcriptome data against genome graphs. Therefore, it is appealing to perform a more accurate ASE analysis in livestock using the graph genome approach.

Comprehensive variations from pangenome might contribute to the missing heritability and improve genomic predictions

Even the most comprehensive catalogues of genetic variations available to date cannot capture the full heritability of traits, widely known as missing heritability [69]. For example, a large meta-analysis on stature in cattle identified 163 lead variants, but these variants only explain about 13.8% of the heritability of stature [70]. There were some proposals explaining the sources of missing heritability, such as the contribution of rarer variants [71] that can be recovered when considering large mapping cohorts that have been genotyped directly for whole-genome variations [72]. However, complex structural variations and sequences not present in the reference genome which are not routinely assessed might as well contribute to the missing heritability [73, 74]. The effect of large variations can be completely missed, which undermine its contribution to the genetic of traits.

Several studies in humans [2, 10, 75] have attempted to integrate sequence-resolved large structural variations that were detected using long read sequencing into pangenome graphs. These variations may then serve as a reference for reference-guided variant discovery from short-read sequencing data. The graph-based structures developed in this thesis provide an appealing resource to call genotypes at large structural variants from the vast amount of whole-genome short-read re-sequencing data that have been collected in many breeds of cattle. These genotypes may then be used for robust genetic studies that might uncover some part of the missing heritability.

Genotypes at genome-wide variants are frequently used to predict the animal's genetic merit, widely known as genomic prediction [76]. Genomic prediction typically relies on SNPs and small insertion and deletion polymorphisms detected from a linear reference genome. Recent efforts aimed at including genotypes from structural variations in the genomic prediction models. However, these structural variations only resulted in small improvements in prediction accuracy over pure SNP-based prediction [77, 78]. This might be partly due to an incomplete representation of structural variations from short-read sequencing data aligned to linear coordinates. Pangenome graphs offer the ability to catalogue more accurate and unbiased variations from the population, particularly at genomic regions that are missing in the reference genome. These hitherto neglected variants may improve the prediction accuracy, thus leading to additional genetic gain. Additionally, informative graph genomes that integrate diverse functional omics data might be used to prioritize variants in genomic prediction. MacLeod et al. [79] showed that stratification of variants with functional omics data improves prediction accuracy over treating all variants equally.

Sequence variants in the pangenome might be causative for agriculturally important traits

Most of the genomic analyses in livestock rely on genetic markers discovered from a linear reference genome. Thus, association testing between phenotypes and markers is blind to variants in segments that are missing in the linear reference sequence. Several studies in plants and humans have shown that important QTL may be missed in genomic regions that are absent from the reference [80, 81, 82]. Additionally, the fine mapping of causative variants is challenging in genomic regions harboring structural variants that are not part of the reference sequences. On the other hand, the contribution of large variations to the genetic architecture of complex traits may be substantial. Chiang et al. [17] and Chaisson et al. [83] suggest that large structural variations are more likely to be associated with GWAS signals due to having larger impacts on gene expression than SNPs. Song et al. [82] performed GWAS between phenotypes and presence-absence variations of pangenome segments (termed as PAV GWAS) across *Brassica* accessions to identify large insertions, not part of the reference sequences, as causal variants for agriculturally important traits. Moreover, a pangenome-based analysis of more than 15,000 Icelandic genomes uncovered a common 766 bp insertion [80] that is associated with a complex trait. These series of studies highlight interesting areas for potential applications of pangenome-based analyses to dissect the genetic architecture of complex traits, which have not been applied to livestock populations. Hayes and Daetwyler [8] noted that the rate of causal variant identification for complex traits has been very slow in cattle, which might partially be due to not considering many variants that are not accessible from the existing linear reference genome.

Resources to catalogue and preserve the complete genetic diversity

Domestication and selection of livestock species resulted in a considerable reduction of the genetic diversity compared to the wild relatives (termed as *the cost of domestication*) [84]. Selection for desirable genes might be accompanied by unintentional removal of beneficial variants related to disease-, parasite- or heat resilience relevant to potentially changing environmental conditions. Thus, the cosmopolitan breeds might be more susceptible to environmental stress. For example, breeding for milk yield in dairy cattle is accompanied with undesired impacts on fertility [85] and there is a negative genetic correlation between milk yield and mastitis resistance [86].

Targeting non-domesticated relatives in the pangenome might help to identify genetic diversity which has been lost due to domestication and breeding that might be favorable for the future environmental changes [87]. Such a *super pangenome* of extant and extinct relatives of domestic species might uncover alleles that were lost during domestication that can be re-introgressed into the modern breeds. Of note, genome

assemblies of undomesticated Bovinae members of bison (*Bison bison*) [88] and gaur (*Bos gaurus*) have been created recently, providing an opportunity to enrich the bovine pangenome with more diversity (Fig. 5.1). Additionally, the trio binning assembly technique performs better for trios with diverged parents [46, 89]. This provides exciting opportunities to generate assemblies for understudied or undomesticated cattle relatives that will generate diverse collections cattle assemblies that become ideal resources to construct a comprehensive bovine pangenome.

References

[1] Hannes P Eggertsson, Hakon Jonsson, Snaedis Kristmundsdottir, Eirikur Hjartarson, Birte Kehr, Gisli Masson, Florian Zink, Kristjan E Hjorleifsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, et al. Graphtyper enables population-scale genotyping using pangenome graphs. *Nature genetics*, 49(11), 2017.

[2] Hannes P Eggertsson, Snaedis Kristmundsdottir, Doruk Beyter, Hakon Jonsson, Astros Skuladottir, Marteinn T Hardarson, Daniel F Gudbjartsson, Kari Stefansson, Bjarni V Halldorsson, and Pall Melsted. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature communications*, 10(1):1–8, 2019.

[3] Benjamin D Rosen, Derek M Bickhart, Robert D Schnabel, Sergey Koren, Christine G Elsik, Elizabeth Tseng, Troy N Rowan, Wai Y Low, Aleksey Zimin, Christine Couldrey, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience*, 9(3):giaa021, 2020.

[4] Meenu Bhati, Naveen Kumar Kadri, Danang Crysnanto, and Hubert Pausch. Assessing genomic diversity and signatures of selection in Original Braunvieh cattle using whole-genome sequencing data. *BMC genomics*, 21(1):1–14, 2020.

[5] Jacob Pritt, Nae-Chyun Chen, and Ben Langmead. FORGe: prioritizing variants for graph genomes. *Genome biology*, 19(1):1–16, 2018.

[6] Chirag Jain, Neda Tavakoli, and Srinivas Aluru. A variant selection framework for genome graphs. *bioRxiv*, 2021.

[7] Benjamin Kaminow, Sara Ballouz, Jesse Gillis, and Alexander Dobin. Virtue as the mean: Pan-human consensus genome significantly improves the accuracy of RNA-seq analyses. *bioRxiv*, 2020.

[8] Ben J Hayes and Hans D Daetwyler. 1000 bull genomes project to map simple and complex genetic traits in cattle: applications and outcomes. *Annual review of animal biosciences*, 7:89–102, 2019.

[9] Dong-Dong Wu, Xiang-Dong Ding, Sheng Wang, Jan M Wójcik, YI Zhang, Małgorzata Tokarska, Yan Li, Ming-Shan Wang, Omar Faruque, Rasmus Nielsen, et al. Pervasive introgression facilitated domestication and adaptation in the Bos species complex. *Nature ecology & evolution*, 2(7):1139–1145, 2018.

[10] Sai Chen, Peter Krusche, Egor Dolzhenko, Rachel M Sherman, Roman Petrovski, Felix Schlesinger, Melanie Kirsche, David R Bentley, Michael C Schatz, Fritz J Sedlazeck, et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome biology*, 20(1):1–13, 2019.

[11] Jouni Sirén, Jean Monlong, Xian Chang, Adam M Novak, Jordan M Eizenga, Charles Markello, Jonas Sibbesen, Glenn Hickey, Pi-Chuan Chang, Andrew Carroll, et al. Genotyping common, large structural variations in 5,202 genomes using pangenomes, the Giraffe mapper, and the vg toolkit. *Biorxiv*, 2020.

[12] George E Liu, Yali Hou, Bin Zhu, Maria Francesca Cardone, Lu Jiang, Angelo Cellamare, Apratim Mitra, Leeson J Alexander, Luiz L Coutinho, Maria Elena Dell'Aquila, et al. Analysis of copy number variations among diverse cattle breeds. *Genome research*, 20(5):693–703, 2010.

[13] Derek M Bickhart, Yali Hou, Steven G Schroeder, Can Alkan, Maria Francesca Cardone, Lakshmi K Matukumalli, Jiuzhou Song, Robert D Schnabel, Mario Ventura, Jeremy F Taylor, et al. Copy number variation of individual cattle genomes using next-generation sequencing. *Genome research*, 22(4):778–790, 2012.

[14] Mekki Boussaha, Diane Esquerré, Johanna Barbieri, Anis Djari, Alain Pinton, Rabia Letaief, Gérald Salin, Frédéric Escudié, Alain Roulet, Sébastien Fritz, et al. Genome-wide study of structural variants in bovine holstein, montbéliarde and normande dairy breeds. *PloS one*, 10(8):e0135931, 2015.

[15] Long Chen, Amanda J Chamberlain, Coralie M Reich, Hans D Daetwyler, and Ben J Hayes. Detection and validation of structural variations in bovine whole-genome sequence data. *Genetics Selection Evolution*, 49(1):1–13, 2017.

[16] Yan Hu, Han Xia, Mingxun Li, Chang Xu, Xiaowei Ye, Ruixue Su, Mai Zhang, Oyekanmi Nash, Tad S Sonstegard, Liguo Yang, George E Liu, and Yang Zhou. Comparative analyses of copy number variations between Bos taurus and Bos indicus. *BMC Genomics*, 21(1):682, 2020. ISSN 1471-2164. doi: 10.1186/s12864-020-07097-6.

[17] Colby Chiang, Alexandra J Scott, Joe R Davis, Emily K Tsang, Xin Li, Yungil Kim, Tarik Hadzic, Farhan N Damani, Liron Ganel, Stephen B Montgomery, et al. The impact of structural variation on human gene expression. *Nature genetics*, 49(5):692–699, 2017.

[18] Alexandra J Scott, Colby Chiang, and Ira M Hall. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *bioRxiv*, 2021.

[19] John Huddleston, Mark JP Chaisson, Karyn Meltz Steinberg, Wes Warren, Kendra Hoekzema, David Gordon, Tina A Graves-Lindsay, Katherine M Munson, Zev N Kronenberg, Laura Vives, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome research*, 27(5):677–685, 2017.

[20] Peter A Audano, Arvis Sulovari, Tina A Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, AnneMarie E Welch, Max L Dougherty, Bradley J Nelson, Ankeeta Shah, Susan K Dutcher, et al. Characterizing the major structural variant alleles of the human genome. *Cell*, 176(3):663–675, 2019.

[21] David L Adelson, Joy M Raison, and Robert C Edgar. Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *Proceedings of the National Academy of Sciences*, 106(31):12855–12860, 2009.

[22] Christine R Beck, José Luis Garcia-Perez, Richard M Badge, and John V Moran. LINE-1 elements in structural variation and disease. *Annual review of genomics and human genetics*, 12:187–215, 2011.

[23] Lei Chen, Qiang Qiu, Yu Jiang, Kun Wang, Zeshan Lin, Zhipeng Li, Faysal Bibi, Yongzhi Yang, Jinhuan Wang, Wenhui Nie, et al. Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science*, 364(6446), 2019.

[24] Qiu-mei Ji, Jin-wei Xin, Zhi-xin Chai, Cheng-fu Zhang, Yangla Dawa, Sang Luo, Qiang Zhang, Zhandui Pingcuo, Min-Sheng Peng, Yong Zhu, et al. A chromosome-scale reference genome and genome-wide genetic variations elucidate adaptation in yak. *Molecular ecology resources*, 21(1):201–211, 2021.

[25] Chuang Chen, Carolyn TA Herzig, Leeson J Alexander, John W Keele, Tara G McDaneld, Janice C Telfer, and Cynthia L Baldwin. Gene number determination and genetic polymorphism of the gamma delta T cell co-receptor WC1 genes. *BMC genetics*, 13(1):1–17, 2012.

[26] Wai Yee Low, Rick Tearle, Ruijie Liu, Sergey Koren, Arang Rhie, Derek M. Bickhart, Benjamin D. Rosen, Zev N. Kronenberg, Sarah B. Kingan, Elizabeth Tseng, Françoise Thibaud-Nissen, Fergal J. Martin, Konstantinos Billis, Jay Ghurye, Alex R. Hastie, Joyce Lee, Andy W.C. Pang, Michael P. Heaton, Adam M. Phillippy, Stefan Hiendleder, Timothy P.L. Smith, and John L. Williams. Haplotype-Resolved Cattle Genomes Provide Insights Into Structural Variation and Adaptation. *Nature Communications*, 11 (1), aug 2020. ISSN 2041-1723. doi: 10.1101/720797.

[27] Derek M Bickhart and George E Liu. The challenges and importance of structural variation detection in livestock. *Frontiers in genetics*, 5:37, 2014.

[28] Karen HY Wong, Michal Levy-Sakin, and Pui-Yan Kwok. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nature communications*, 9(1):1–9, 2018.

[29] Karen HY Wong, Walfred Ma, Chun-Yu Wei, Erh-Chan Yeh, Wan-Jia Lin, Elin HF Wang, Jen-Ping Su, Feng-Jen Hsieh, Hsiao-Jung Kao, Hsiao-Huei Chen, et al. Towards a reference genome that captures global genetic diversity. *Nature communications*, 11(1):1–11, 2020.

[30] Ran Li, Weiwei Fu, Rui Su, Xiaomeng Tian, Duo Du, Yue Zhao, Zhuqing Zheng, Qiuming Chen, Shan Gao, Yudong Cai, et al. Towards the complete goat pan-genome by recovering missing genomic segments from the reference genome. *Frontiers in genetics*, 10, 2019.

[31] Mingzhou Li, Lei Chen, Shilin Tian, Yu Lin, Qianzi Tang, Xuming Zhou, Diyan Li, Carol KL Yeung, Tiandong Che, Long 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.

[32] Zhongqu Duan, Yuyang Qiao, Jinyuan Lu, Huimin Lu, Wenmin Zhang, Fazhe Yan, Chen Sun, Zhiqiang Hu, Zhen Zhang, Guichao Li, et al. HUPAN: a pan-genome analysis pipeline for human genomes. *Genome biology*, 20(1):1–11, 2019.

[33] Agnieszka A Golicz, Philipp E Bayer, Guy C Barker, Patrick P Edger, HyeRan Kim, Paula A Martinez, Chon Kit Kenneth Chan, Anita Severn-Ellis, W Richard McCombie, Isobel AP Parkin, et al. The pangenome of an agronomically important crop plant Brassica oleracea. *Nature communications*, 7(1): 1–8, 2016.

[34] Sean P Gordon, Bruno Contreras-Moreira, Daniel P Woods, David L Des Marais, Diane Burgess, Shengqiang Shu, Christoph Stritt, Anne C Roulin, Wendy Schackwitz, Ludmila Tyler, et al. Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nature communications*, 8(1):1–13, 2017.

[35] Lei Gao, Itay Gonda, Honghe Sun, Qiyue Ma, Kan Bao, Denise M Tieman, Elizabeth A Burzynski-Chang, Tara L Fish, Kaitlin A Stromberg, Gavin L Sacks, et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature genetics*, 51(6):1044–1051, 2019.

[36] Li Lei, Eugene Goltsman, David Goodstein, Guohong Albert Wu, Daniel S Rokhsar, and John P Vogel. Plant Pan-Genomics Comes of Age. *Annual Review of Plant Biology*, 72, 2021.

[37] Tim Smith. Individual Breed Genome Assembly to Create the Cattle Pangenome. In *Online Abstracts: International Plant and Animal Genomes XXVIII Conference*, page W120, San Diego, 2020.

[38] Roger Ros-Freixedes, Serap Gonen, Gregor Gorjanc, and John M Hickey. A method for allocating low-coverage sequencing resources by targeting haplotypes rather than individuals. *Genetics Selection Evolution*, 49(1):78, 2017. ISSN 1297-9686.

[39] Timothy Rhyker Ranallo-Benavidez, Zachary H Lemmon, Sebastian Soyk, Sergey Aganezov, William J Salerno, Rajiv C McCoy, Zachary B Lippman, Michael C Schatz, and Fritz J Sedlazeck. Optimized sample selection for cost-efficient long-read population sequencing. *Genome Research*, 2021.

[40] Jordan M Eizenga, Adam M Novak, Emily Kobayashi, Flavia Villani, Cecilia Cisar, Simon Heumos, Glenn Hickey, Vincenza Colonna, Benedict Paten, and Erik Garrison. Efficient dynamic variation graphs. *Bioinformatics*, 2020.

[41] Jouni Sirén, Erik Garrison, Adam M Novak, Benedict Paten, and Richard Durbin. Haplotype-aware graph indexes. *Bioinformatics*, 36(2):400–407, 2020.

[42] Jonas A Sibbesen, Jordan M Eizenga, and Adam M Novak. Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. *bioRxiv*, 2021.

[43] Samuel Hokin, Alan Cleary, and Joann Mudge. Disease association with frequented regions of genotype graphs. *medRxiv*, 2020.

[44] Alice M Kaye and Wyeth W Wasserman. The genome atlas: Navigating a new era of reference genomes. *Trends in Genetics*, 2021.

[45] Emily L Clark, Alan L Archibald, Hans D Daetwyler, Martien AM Groenen, Peter W Harrison, Ross D Houston, Christa Kühn, Sigbjørn Lien, Daniel J Macqueen, James M Reecy, et al. From FAANG to fork: application of highly annotated genomes to improve farmed animal production. *Genome Biology*, 21(1):1–9, 2020.

[46] Edward S Rice, Sergey Koren, Arang Rhie, Michael P Heaton, Theodore S Kalbfleisch, Timothy Hardy, Peter H Hackett, Derek M Bickhart, Benjamin D Rosen, Brian Vander Ley, et al. Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. *Gigascience*, 9(4):giaa029, 2020.

[47] Jana Ebler, Wayne E Clarke, Tobias Rausch, Peter A Audano, Torsten Houwaart, Jan Korbel, Evan E Eichler, Michael C Zody, Alexander T Dilthey, and Tobias Marschall. Pangenome-based genome inference. *bioRxiv*, 2020.

[48] Peter Ebert, Peter A Audano, Qihui Zhu, Bernardo Rodriguez-Martin, David Porubsky, Marc Jan Bonder, Arvis Sulovari, Jana Ebler, Weichen Zhou, Rebecca Serra Mari, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, 372(6537), 2021.

[49] Aaron M Wenger, Paul Peluso, William J Rowell, Pi-Chuan Chang, Richard J Hall, Gregory T Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D Olson, Armin Töpfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen-Shan Chin, Adam M Phillippy, Michael C Schatz, Gene Myers, Mark A DePristo, Jue Ruan, Tobias Marschall, Fritz J Sedlazeck, Justin M Zook, Heng Li, Sergey Koren, Andrew Carroll, David R Rank, and Michael W Hunkapiller. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10):1155–1162, 2019.

[50] Glennis A Logsdon, Mitchell R Vollger, PingHsun Hsieh, Yafei Mao, Mikhail A Liskovykh, Sergey Koren, Sergey Nurk, Ludovica Mercuri, Philip C Dishuck, Arang Rhie, et al. The structure, function and evolution of a complete human chromosome 8. *Nature*, pages 1–7, 2021.

[51] Karen H Miga, Sergey Koren, Arang Rhie, Mitchell R Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, Edmund Howe, David Porubsky, Glennis A Logsdon, et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, 585(7823):79–84, 2020.

[52] Joel Armstrong, Glenn Hickey, Mark Diekhans, Ian T Fiddes, Adam M Novak, Alden Deran, Qi Fang, Duo Xie, Shaohong Feng, Josefin Stiller, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, 587(7833):246–251, 2020.

[53] James K Bonfield, John Marshall, Petr Danecek, Heng Li, Valeriu Ohan, Andrew Whitwham, Thomas Keane, and Robert M Davies. HTSlib: C library for reading/writing high-throughput sequencing data. *GigaScience*, 10(2):giab007, 2021.

[54] Yutong Qiu and Carl Kingsford. Constructing smaller genome graphs via string compression. *bioRxiv*, 2021.

[55] Tizian Schulz, Roland Wittler, Sven Rahmann, Faraz Hach, and Jens Stoye. Detecting High Scoring Local Alignments in Pangenome Graphs. *bioRxiv*, 2020.

[56] Toshiyuki T Yokoyama, Yoshitaka Sakamoto, Masahide Seki, Yutaka Suzuki, and Masahiro Kasahara. MoMI-G: modular multi-scale integrated genome graph browser. *BMC bioinformatics*, 20(1):1–14, 2019.

[57] Wolfgang Beyer, Adam M Novak, Glenn Hickey, Jeffrey Chan, Vanessa Tan, Benedict Paten, and Daniel R Zerbino. Sequence tube maps: making graph genomes intuitive to commuters. *Bioinformatics*, 35(24), 2019.

[58] Jordan M Eizenga, Adam M Novak, Jonas A Sibbesen, Simon Heumos, Ali Ghaffaari, Glenn Hickey, Xian Chang, Josiah D Seaman, Robin Rounthwaite, Jana Ebler, et al. Pangenome graphs. *Annual Review of Genomics and Human Genetics*, 21:139–162, 2020.

[59] Daehwan Kim, Joseph M Paggi, Chanhee Park, Christopher Bennett, and Steven L Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology*, 37 (8):907–915, 2019.

[60] Ivar Grytten, Knut D Rand, Alexander J Nederbragt, and Geir K Sandve. Assessing graph-based read mappers against a baseline approach highlights strengths and weaknesses of current methods. *BMC genomics*, 21:1–9, 2020.

[61] Heng Li, Xiaowen Feng, and Chong Chu. The design and construction of reference pangenome graphs with minigraph. *Genome biology*, 21(1):1–19, 2020.

[62] Sylvain Foissac, Sarah Djebali, Kylie Munyard, Nathalie Vialaneix, Andrea Rau, Kevin Muret, Diane Esquerré, Matthias Zytnicki, Thomas Derrien, Philippe Bardou, et al. Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC biology*, 17(1):1–25, 2019.

[63] Stephane E Castel, François Aguet, Pejman Mohammadi, Kristin G Ardlie, and Tuuli Lappalainen. A vast resource of allelic expression data spanning human tissues. *Genome biology*, 21(1):1–12, 2020.

[64] Amanda J Chamberlain, Christy J Vander Jagt, Benjamin J Hayes, Majid Khansefid, Leah C Marett, Catriona A Millen, Thuy TT Nguyen, and Michael E Goddard. Extensive variation between tissues in allele specific expression in an outbred mammal. *BMC genomics*, 16(1):1–20, 2015.

[65] Gabriel M Guillocheau, Abdelmajid El Hou, Cédric Meersseman, Diane Esquerré, Emmanuelle Rebours, Rabia Letaief, Morgane Simao, Nicolas Hypolite, Emmanuelle Bourneuf, Nicolas Bruneau, et al. Survey of allele specific expression in bovine muscle. *Scientific reports*, 9(1):1–11, 2019.

[66] Jennifer Jessica Bruscadin, Marcela Maria de Souza, Karina Santos de Oliveira, Marina Ibelli Pereira Rocha, Juliana Afonso, Tainã Figueiredo Cardoso, Adhemar Zerlotini, Luiz Lehmann Coutinho, Simone Cristina Méo Niciura, and Luciana Correia de Almeida Regitano. Muscle allele-specific expression QTLs may affect meat quality traits in Bos indicus. *Scientific Reports*, 11(1):1–14, 2021.

[67] Mazdak Salavati, Stephen J Bush, Sergio Palma-Vera, Mary EB McCulloch, David A Hume, and Emily L Clark. Elimination of reference mapping bias reveals robust immune related allele-specific expression in crossbred sheep. *Frontiers in genetics*, 10:863, 2019.

[68] Bryce Van De Geijn, Graham McVicker, Yoav Gilad, and Jonathan K Pritchard. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature methods*, 12(11):1061–1063, 2015.

[69] Brendan Maher. Personal genomes: The case of the missing heritability. *Nature News*, 456(7218):18–21, 2008.

[70] Aniek C Bouwman, Hans D Daetwyler, Amanda J Chamberlain, Carla Hurtado Ponce, Mehdi Sargolzaei, Flavio S Schenkel, Goutam Sahana, Armelle Govignon-Gion, Simon Boitard, Marlies Dolezal, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nature genetics*, 50(3):362–367, 2018.

[71] Oscar Gonzalez-Recio, Hans D Daetwyler, Iona M MacLeod, Jennie E Pryce, Phil J Bowman, Ben J Hayes, and Michael E Goddard. Rare variants in transcript and potential regulatory regions explain a small percentage of the missing heritability of complex traits in cattle. *PloS one*, 10(12), 2015.

[72] Pierrick Wainschtein, Deepti P Jain, Loic Yengo, Zhili Zheng, L Adrienne Cupples, Aladdin H Shadyab, Barbara McKnight, Benjamin M Shoemaker, Braxton D Mitchell, Bruce M Psaty, et al. Recovery of trait heritability from whole genome sequence data. *BioRxiv*, page 588020, 2019.

[73] Emmanuelle Génin. Missing heritability of complex diseases: case solved? *Human genetics*, 139(1): 103–113, 2020.

[74] Frances Theunissen, Loren L Flynn, Ryan S Anderton, Frank Mastaglia, Julia Pytte, Leanne Jiang, Stuart Hodgetts, Daniel K Burns, Ann Saunders, Sue Fletcher, et al. Structural variants may be a source of missing heritability in sALS. *Frontiers in neuroscience*, 14, 2020.

[75] Glenn Hickey, David Heller, Jean Monlong, Jonas A Sibbesen, Jouni Sirén, Jordan Eizenga, Eric T Dawson, Erik Garrison, Adam M Novak, and Benedict Paten. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome biology*, 21(1):1–17, 2020.

[76] Theo HE Meuwissen, Ben J Hayes, and Michael E Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, 2001.

[77] A Hay El Hamidi, Yuri T Utsunomiya, Lingyang Xu, Yang Zhou, Haroldo HR Neves, Roberto Carvalheiro, Derek M Bickhart, Li Ma, Jose Fernando Garcia, and George E Liu. Genomic predictions combining SNP markers and copy number variations in Nellore cattle. *BMC genomics*, 19(1):1–8, 2018.

[78] Long Chen, Jennie E Pryce, Ben J Hayes, and Hans D Daetwyler. Investigating the Effect of Imputed Structural Variants from Whole-Genome Sequence on Genome-Wide Association and Genomic Prediction in Dairy Cattle. *Animals*, 11(2):541, 2021.

[79] IM MacLeod, PJ Bowman, CJ Vander Jagt, M Haile-Mariam, KE Kemper, AJ Chamberlain, C Schrooten, BJ Hayes, and ME Goddard. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC genomics*, 17(1):1–21, 2016.

[80] Birte Kehr, Anna Helgadottir, Pall Melsted, Hakon Jonsson, Hannes Helgason, Adalbjörg Jonasdottir, Aslaug Jonasdottir, Asgeir Sigurdsson, Arnaldur Gylfason, Gisli H Halldorsson, et al. Diversity in non-repetitive human sequences not found in the reference genome. *Nature Genetics*, 49(4):588–593, 2017.

[81] Joseph L Gage, Brieanne Vaillancourt, John P Hamilton, Norma C Manrique-Carpintero, Timothy J Gustafson, Kerrie Barry, Anna Lipzen, William F Tracy, Mark A Mikel, Shawn M Kaeppler, et al. Multiple Maize Reference Genomes Impact the Identification of Variants by Genome-Wide Association Study in a Diverse Inbred Panel. *The plant genome*, 12(2), 2019.

[82] Jia-Ming Song, Zhilin Guan, Jianlin Hu, Chaocheng Guo, Zhiquan Yang, Shuo Wang, Dongxu Liu, Bo Wang, Shaoping Lu, Run Zhou, et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nature Plants*, 6(1):34–45, 2020.

[83] Mark JP Chaisson, Ashley D Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J Gardner, Oscar L Rodriguez, Li Guo, Ryan L Collins, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature communications*, 10(1):1–16, 2019.

[84] Gillian P McHugo, Michael J Dover, and David E MacHugh. Unlocking the origins and biology of domestic animals using ancient DNA and paleogenomics. *BMC biology*, 17(1):1–20, 2019.

[85] JE Pryce, MD Royal, PC Garnsworthy, and IL Mao. Fertility in the high-producing dairy cow. *Livestock production science*, 86(1-3):125–135, 2004.

[86] Zexi Cai, Magdalena Dusza, Bernt Guldbrandtsen, Mogens Sandø Lund, and Goutam Sahana. Distinguishing pleiotropy from linked QTL between milk production traits and mastitis resistance in Nordic Holstein cattle. *Genetics Selection Evolution*, 52:1–15, 2020.

[87] Aamir W Khan, Vanika Garg, Manish Roorkiwal, Agnieszka A Golicz, David Edwards, and Rajeev 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.

[88] Jonas Oppenheimer, Benjamin D Rosen, Michael P Heaton, Brian L Vander Ley, Wade R Shafer, Fred T Schuetze, Brad Stroud, Larry A Kuehn, Jennifer C McClure, Jennifer P Barfield, et al. A reference genome assembly of American bison, Bison bison bison. *Journal of Heredity*, 112(2):174–183, 2021.

[89] Michael P Heaton, Timothy PL Smith, Derek M Bickhart, Brian L Vander Ley, Larry A Kuehn, Jonas Oppenheimer, Wade R Shafer, Fred T Schuetze, Brad Stroud, Jennifer C McClure, et al. A reference genome assembly of Simmental cattle, Bos taurus taurus. *Journal of Heredity*, 2021.

Supplementary Material
Chapter 2

Additional file 2.1
Instruction to compile a Graphtyper version modified for the cattle chromosome complement

Modified Graphtyper for variant discovery and genotyping in cattle The most convenient way to run a *Graphtyper* version compiled for the bovine chromosome complement is to use *Docker* (which deals with all required dependencies). The command below starts to download modified *Graphtyper* software hosted at the Dockerhub:

```
docker run --rm cdanang/graphtyper_cattle graphtyper
```

We built the docker images using *Ubuntu* 18.04 as a base image. If you are working on a Linux 64-bit machine you could also get a static executable with command below. We placed the *Graphtyper* binary in /usr/local/bin) and executing command below will copy the *Graphtyper* binary from docker images to the current working directory:

```
docker run --rm -v ${PWD}:/io cdanang/graphtyper_cattle \
cp /usr/local/bin/graphtyper /io

### And then run the software as a standard binary
./graphtyper
```

If you prefer to modify and build a modified version of Graphtyper for the bovine chromosome complement directly from the source, please follow the instructions below:

1. Clone the *Graphtyper Github*

```
git clone --recursive https://github.com/DecodeGenetics/graphtyper.git
```

2. Create a new *branch* at this specific commit tag. We built graphtyper at this specific commit hash (04ab5ee460fa36129fb0d8ea5d4b72adc3836f52), to compile at the same software version that we use in the paper, please use this commit tag. We named the branch as *cattle modification*

```
git checkout -b cattle_modification \
04ab5ee460fa36129fb0d8ea5d4b72adc3836f52
```

3. Change directory into *graphtyper* and modify the chromosomal specifications in the files include *graphtyper/graph/absolute_position.hpp* and *src/typer/vcf.cpp* using UMD 3.1 cattle chromosomal names and lengths. The first modification enables all cattle chromosomes (esp. for chromosome number > 23) as the current software release set the maximum allowed length for each chromosomes according to the human GRChb37 and GRCh38. The second modifications are required that the respective chromosomal information is written to the *vcf header*.

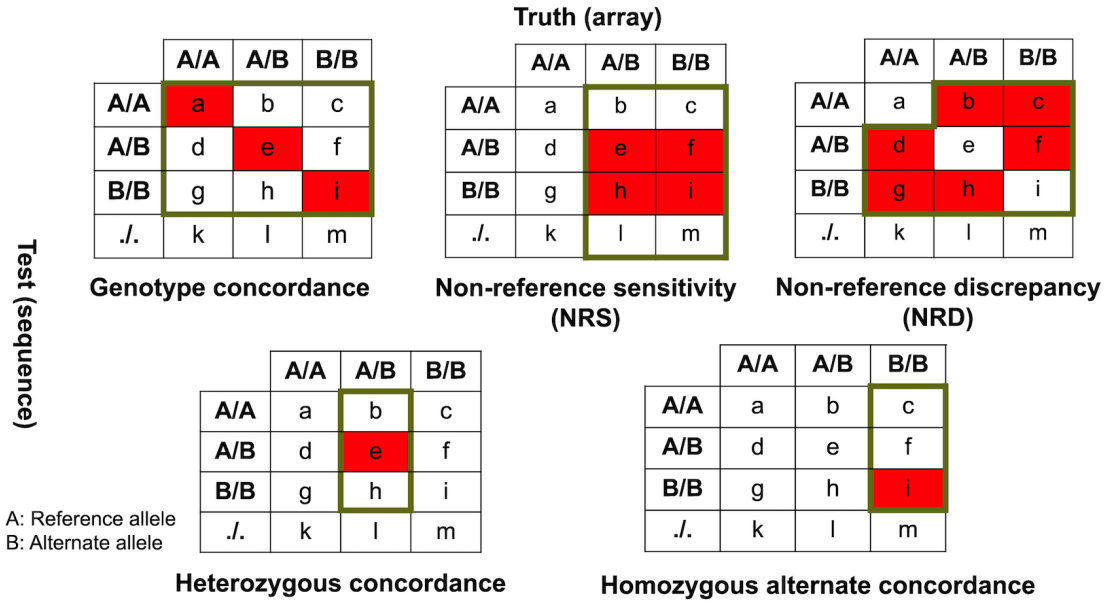
4. Make sure that these dependencies are installed:
   - C++ compiler with C++11 supported (we tested gcc 4.8.5 or gcc 6.3.0
   - Boost$\geq$1.57.0
   - zlib$\geq$1.2.8
   - libbz2
   - liblzma
   - Autotools, Automake, libtool, Make, and CMake>=2.8.8

5. Follow installation procedures as below. This will put the software in *releasebuild/bin/graphtyper*

```
mkdir -p release-build && cd release-build
cmake ..
make -j4 graphtyper
bin/graphtyper # Run Graphtyper with modified cattle chromosome
specifications
```

Additional file 2.2
Properties of the different metrics used for the evaluation of sequence variant genotyping accuracy.

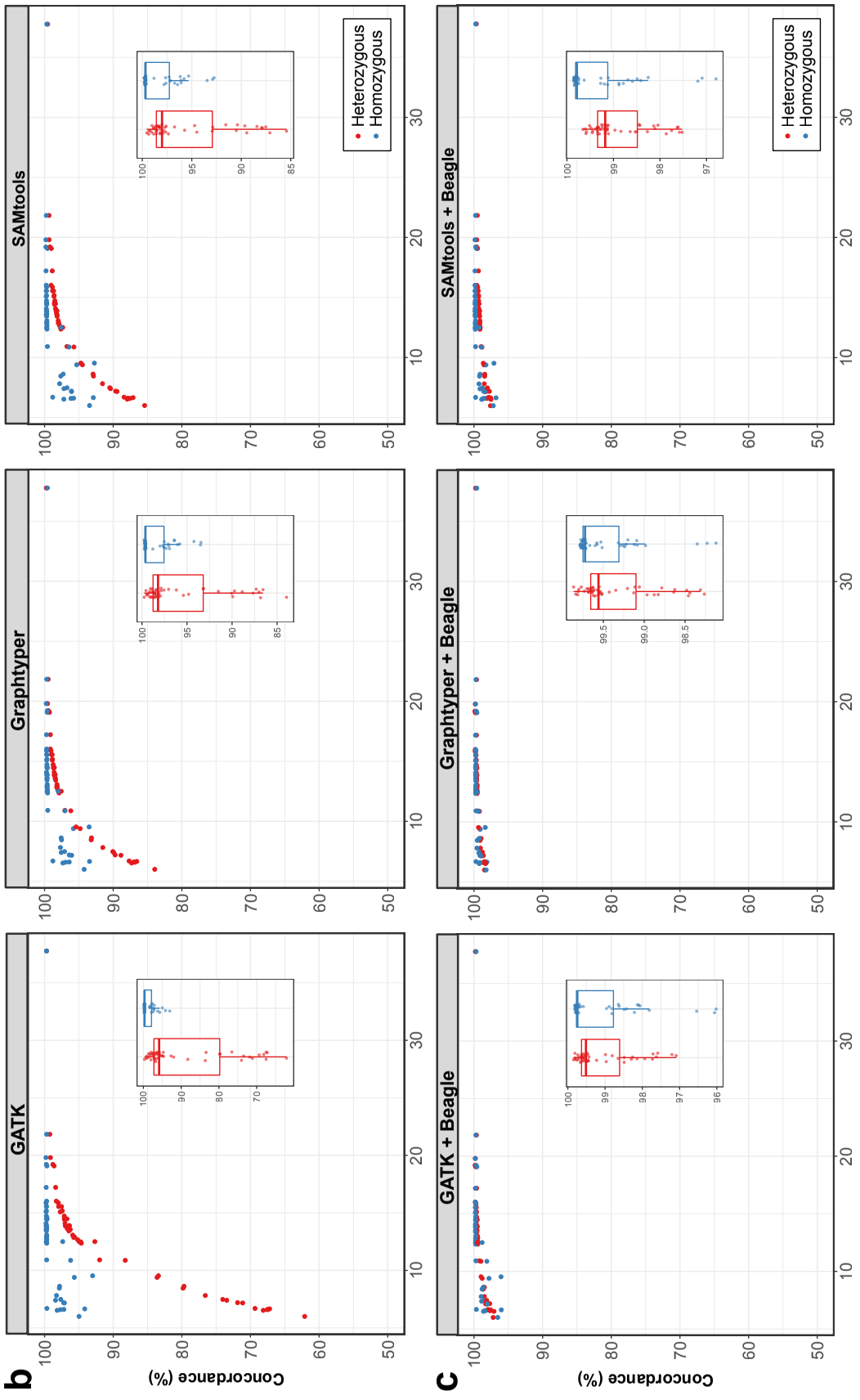
The metrics were calculated using the sum of the red cells as numerator and the cells within the green frame as denominator.



Additional file 2.3
Concordance statistics

The concordance of heterozygous and alternate homozygous genotypes in 49 Original Braunvieh cattle (**a**) and the concordance at the different sequencing depth for the (**b**) raw and (**c**) imputed datasets.

	Heterozygous concordance				Homozygous concordance			
	raw	imp	raw	imp	raw	imp	raw	imp
GATK	89.17	99.11	89.24	99.21	*98.74*	99.18	*98.75*	99.27
Graphtyper	*95.79*	*99.36*	*95.82*	*99.44*	98.55	*99.51*	98.59	*99.57*
SAMtools	95.73	98.91	95.77	98.99	98.46	99.37	98.49	99.41


b
c
GATK
Graphtyper
SAMtools
GATK + Beagle
Graphtyper + Beagle
SAMtools + Beagle
Concordance (%)
Heterozygous
Homozygous

Additional file 2.4

Sequence variant genotyping quality for 18 and 31 animals that were sequenced at a lower and higher than 12-fold sequencing coverage, respectively.

Asterisks denote significant differences with the best value (italic) for a respective parameter.

Coverage less than 12

	Genotype concordance				Non-reference sensitivity				Non-reference discrepancy			
	full		filtered		full		filtered		full		filtered	
	raw	imp	raw	imp	raw	imp	raw	imp	raw	imp	raw	imp
GATK	90.99***	98.7***	91.02***	98.82***	85.63***	98.91	85.51***	98.73	14.64***	2.09***	14.59***	1.91***
Graphtyper	94.89	99.07	94.91	99.17	96.44	99	96.13	98.71	8.04	1.49	8	1.31
SAMtools	94.87	98.61***	94.89	98.67***	96.24***	98.94	95.75***	98.45***	8.11	2.24***	8.09	2.11***

Coverage more than 12

	Genotype concordance				Non-reference sensitivity				Non-reference discrepancy			
	full		filtered		full		filtered		full		filtered	
	raw	imp	raw	imp	raw	imp	raw	imp	raw	imp	raw	imp
GATK	98.73***	99.66	98.76***	99.71	98.3***	99.61	98.14***	99.39	1.8***	0.48*	1.76***	0.42
Graphtyper	99.26	99.67	99.3	99.72	99.25	99.54***	98.88	99.16***	1.04	0.45	0.99	0.4
SAMtools	99.21***	99.59***	99.24***	99.62***	99.21**	99.58***	98.51***	98.79***	1.12***	0.58***	1.08***	0.54***

Additional file 2.5
Twelve 1-Mb regions for which *Graphtyper* initially failed to genotype sequence variants

The algorithm either ran out of memory or exceeded the allocated runtime (12 h). Graphtyper eventually produced genotypes for the sequence variants when these regions were re-run in 10-kb segments.

No	Chromosome	Region (Mb)
1	1	0-1
2	1	145-146
3	3	69-70
4	7	58-57
5	8	110-111
6	12	76-77
7	23	26-27
8	23	29-30
9	26	50-51
10	27	37-38
11	28	39-40
12	29	30-31

Additional file 2.6
Variant filtration using *GATK*

The best practice guidelines for variant discovery using *GATK* recommend sequence variants to be filtered using Variant Quality Score Recalibration (VQSR) because it implements advanced machine learning-based methods to differentiate between true and false-positive variants. However, VQSR relies on sets of high confidence truth/training variants, which are currently not (publicly) available in cattle. Thus, we ran *GATK* with best practice recommendations for variant filtering when applying VQSR is not possible, i.e., we used a generic baseline hard-filtering threshold for each variant annotation (see https://gatkforums.broadinstitute.org/GATK/discussion/2806/howto-apply-hard-filters-to-acall-set). This threshold-based filtering is commonly applied the cattle genomics community [1, 2]

To facilitate running the VQSR module in sheep and goat, i.e., species where sets of truth/training variants are not (publicly) available, Alberto et al. [3] used an intersection of high confidence variants that had been discovered from multiple variant callers as truth/training sets, i.e., they derived truth/training sets directly from the analyzed data. We implemented their approach to apply *GATK* VQSR to our variant dataset. Training and truth sets were constructed using the overlap of the filtered variants from the *GATK*, *Graphtyper* and *SAMtools* pipelines (truth=false, training=true, known=false, prior= 10) and markers from the BovineHD BeadChip (truth=true, training=true, known=false, prior= 15), respectively. Moreover, we used variants listed in dbSNP (version 150) as known variants (truth=false, training=false, known=true, prior=3.0). Following *GATK* VQSR, we retained variants in the 99.9% tranche sensitivity threshold (best practice).

Variant filtration using *GATK* VQSR removed more variants from the raw data than *GATK* hard filtering (Table SN21). However, VQSR retained more HD SNPs than *GATK* hard filtering, possibly reflecting bias that results from the use of HD SNPs as training/truth sets. The values of the concordance statistics (genotype concordance, non-reference sensitivity, nonreference discrepancy) were almost identical between *GATK* VQSR and *GATK* hard filtration (Table SN22) indicating that the choice of either filtration option does not notably affect the concordance between sequence-derived and BovineHD SNP array-derived genotypes. These findings are in line with Vander Jagt et al. [4] who showed that the concordance between microarray-called and sequence-derived genotypes is almost identical using either *GATK* VQSR or the *GATK* 1000 bull genomes project hard filters, even though they used stringently filtered truth/training sets based on a more comprehensive catalogue of variants than in our study. Interestingly, in agreement with Vander Jagt et al. [4], the proportion of opposing homozygous genotypes in sire/son-pairs (which does not suffer from ascertainment bias because it is calculated using sequence-derived SNPs) is less using *GATK* hard filter than *GATK* VQSR.

The performance of *GATK* VQSR may be assessed using the novel variant sensitivity tranche plot (Figure SN21). In the lowest 90% tranches (highest specificity) the filtering

model still retained many false positive variants (orange box and low Ti/Tv ratio). However, when the 99.9% tranche sensitivity is used as filtration criterion as recommended by the *GATK* best practice guidelines, a high proportion of true positive variants is removed from the data.

Overall, our findings suggest that

(i) *GATK* VQSR removes more variants from the data than *GATK* hard filtering,

(ii) *GATK* VQSR does not notably improve the concordance between sequence derived and microarray-called genotypes compared to *GATK* hard filtering,

(iii) the proportion of opposing homozygous genotypes in sire/son-pairs is higher using *GATK* VQSR than *GATK* hard filtering, and

(iv) improving VQSR may be possible by providing more sophisticated truth/training variant datasets produced by orthogonal sequencing technology other than the ones used for training, e.g. Li et al. [5]

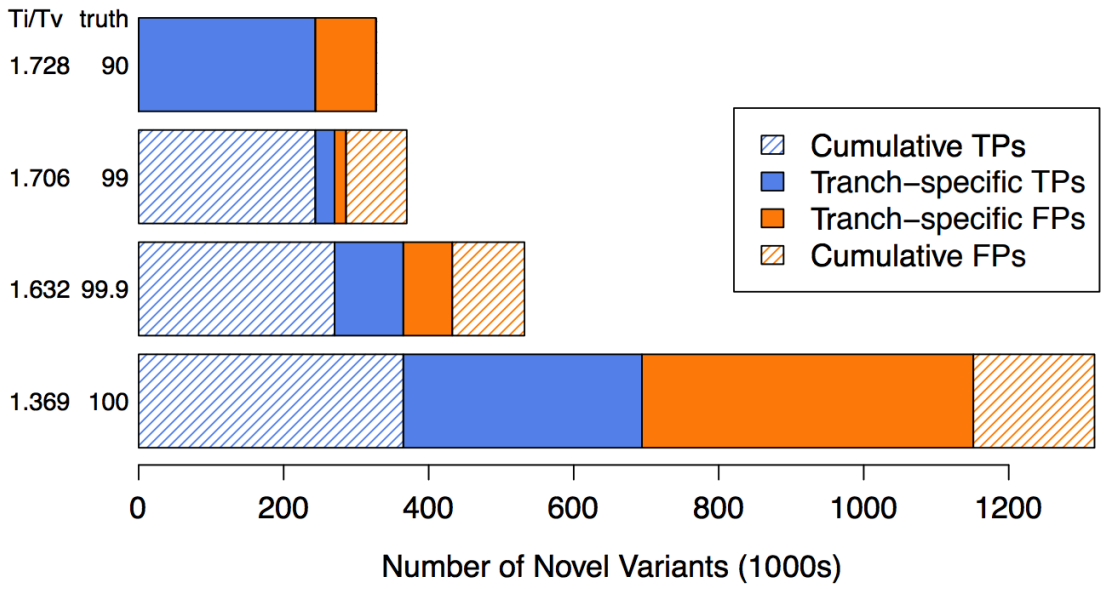
Table SN21 Comparison of variants statistics between unfiltered and filtered datasets using either hard-filtering or VQSR.

	GATK full	*GATK* hard-filter	*GATK* VQSR
Total SNPs	18,594,182	17,248,593	16,537,577
Biallelic	18,347,962	17,111,806	16,430,734
Multi-allelic	246,220	136,787	106,843
Ti/Tv ratio	2.09	2.17	2.16
BovineHD	99.46	99.21	99.38
BovineSNP50	99.14	98.91	98.98

Table SN22 The concordance statistics between hard-filtered and VQSR

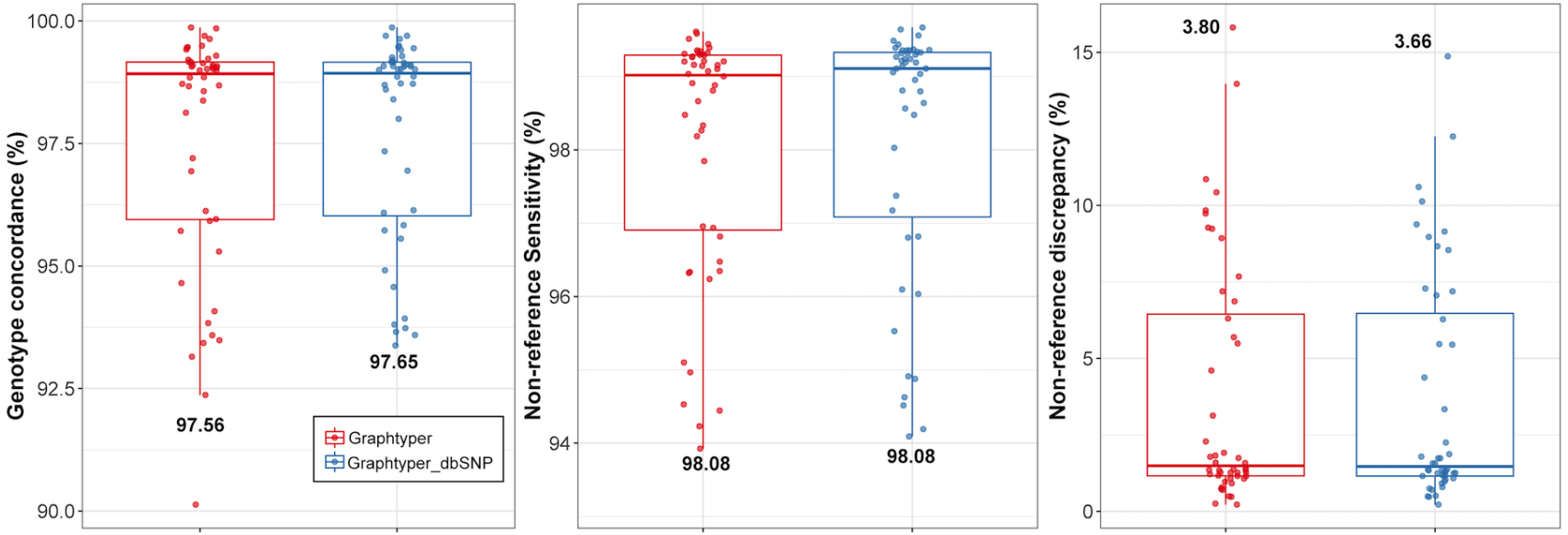
	Genotype concordance	Non-reference sensitivity	Non-reference discrepancy	Opposing Homozygous
GATK hard-filter	96.02	93.67	6.3	0.72
GATK VQSR	96.01	93.77	6.32	0.75

Figure SN21 Tranche sensitivity plot of novel variants as reported by the VQSR model fitting



Additional file 2.7

Accuracy and sensitivity of sequence variant genotyping on bovine chromosome 25 from a variation-aware genome graph that incorporated 2,143,417 dbSNP variants as prior known variants.



Supplementary References

[1] L Koufariotis, BJ Hayes, M Kelly, BM Burns, R Lyons, P Stothard, AJ 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.

[2] Ningbo Chen, Yudong Cai, Qiuming Chen, Ran Li, Kun Wang, Yongzhen Huang, Songmei Hu, Shisheng Huang, Hucai Zhang, Zhuqing Zheng, et al. Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia. *Nature Communications*, 9(1):1–13, 2018.

[3] Florian J Alberto, Frédéric Boyer, Pablo Orozco-terWengel, Ian Streeter, Bertrand Servin, Pierre De Villemereuil, Badr Benjelloun, Pablo Librado, Filippo Biscarini, Licia Colli, et al. Convergent genomic signatures of domestication in sheep and goats. *Nature Communications*, 9(1):1–9, 2018.

[4] CJ Vander Jagt, AJ Chamberlain, RD Schnabel, BJ Hayes, and HD Daetwyler. Which is the best variant caller for large whole-genome sequencing datasets. In *Proceedings of the 11th world congress on genetics applied to livestock production*, pages 11–16, 2018.

[5] Heng Li, Jonathan M Bloom, Yossi Farjoun, Mark Fleharty, Laura Gauthier, Benjamin Neale, and Daniel MacArthur. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nature methods*, 15(8):595–597, 2018.

Supplementary Material
Chapter 3

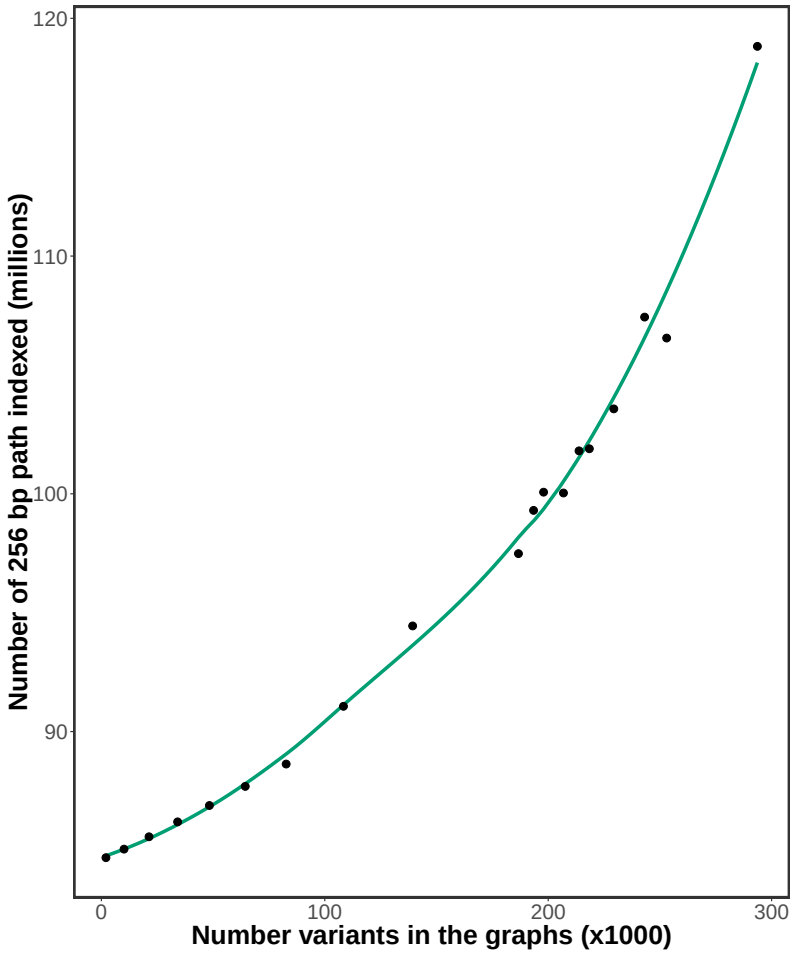


Figure S3.1: **Number of 256 bp haplotype paths in the graphs with an increasing number of variants added to the graphs.**
The line plot is fitted using loess function in *R*.

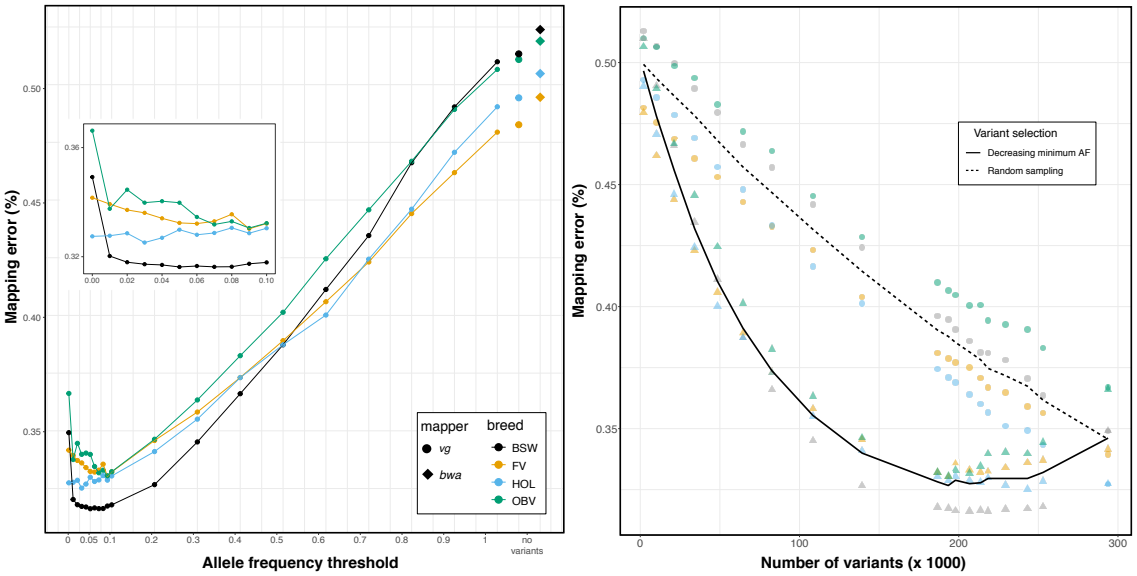


Figure S3.2: **Single-end mapping accuracy using genome graphs that contained variants filtered for allele frequency.**
(a) Proportion of incorrectly mapped reads for four breed-specific augmented genome graphs. Diamonds and large dots represent results from linear mapping using *BWA mem* and *vg*, respectively. The inset is a larger representation of the mapping accuracy for alternate allele frequency thresholds less than 0.1. (b) Read mapping accuracy for breed-specific augmented graphs that contained variants that were either filtered for alternate allele frequency (triangles) or sampled randomly (circles) from all variants detected within a breed. The dashed and solid line represents the average proportion of mapping errors across four breeds using variant prioritization and random sampling.

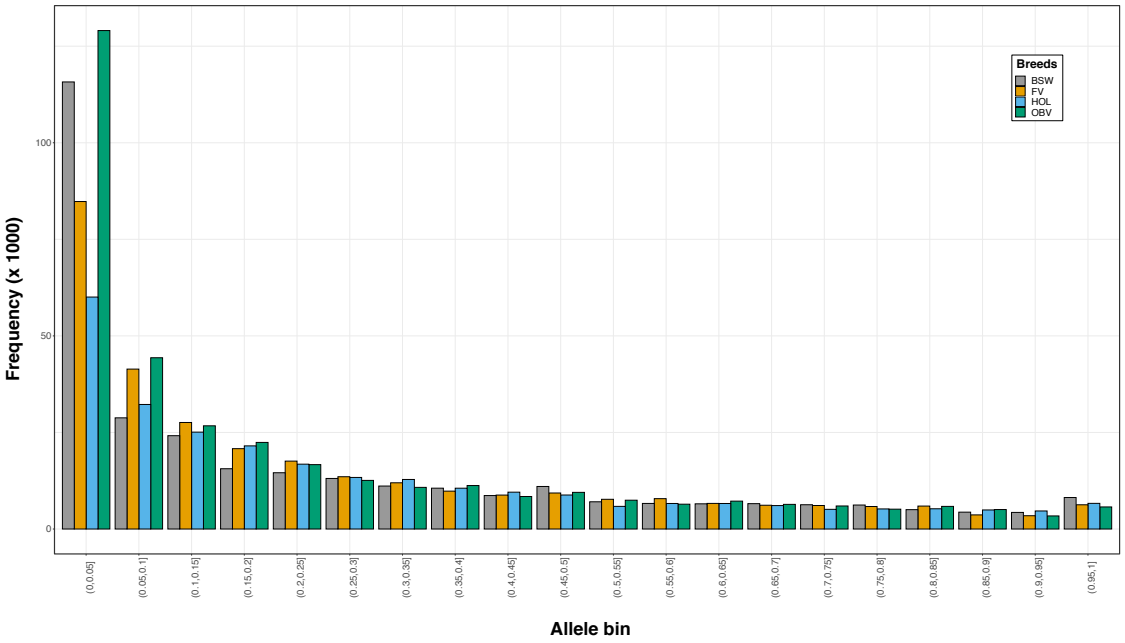


Figure S3.3: **Number of variants detected on chromosome 25 in 82 BSW, 49 FV, 49 HOL and 108 OBV cattle.**
Variants are binned according to allele frequency.

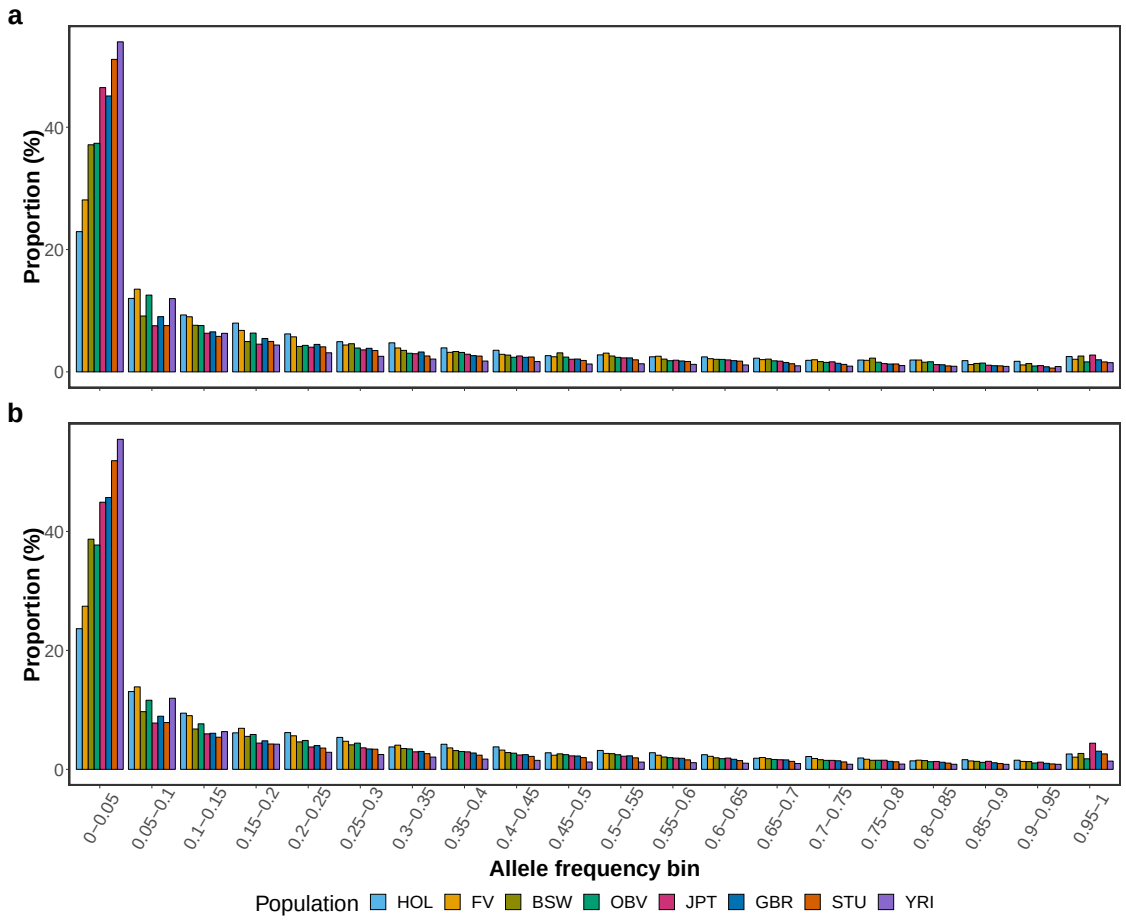


Figure S3.4: **Distribution of alternate allele frequencies in four cattle breeds and four human populations based on (a) bta25 and human chromosome 19 used for graph construction, and (b) whole genome variants.**
The bars indicate the proportion of sequence variants for 20 allele frequency classes. Different colour indicates cattle breeds (HOL, FV, BSW, OBV) and human populations (JPT, GBR, STU, YRI).

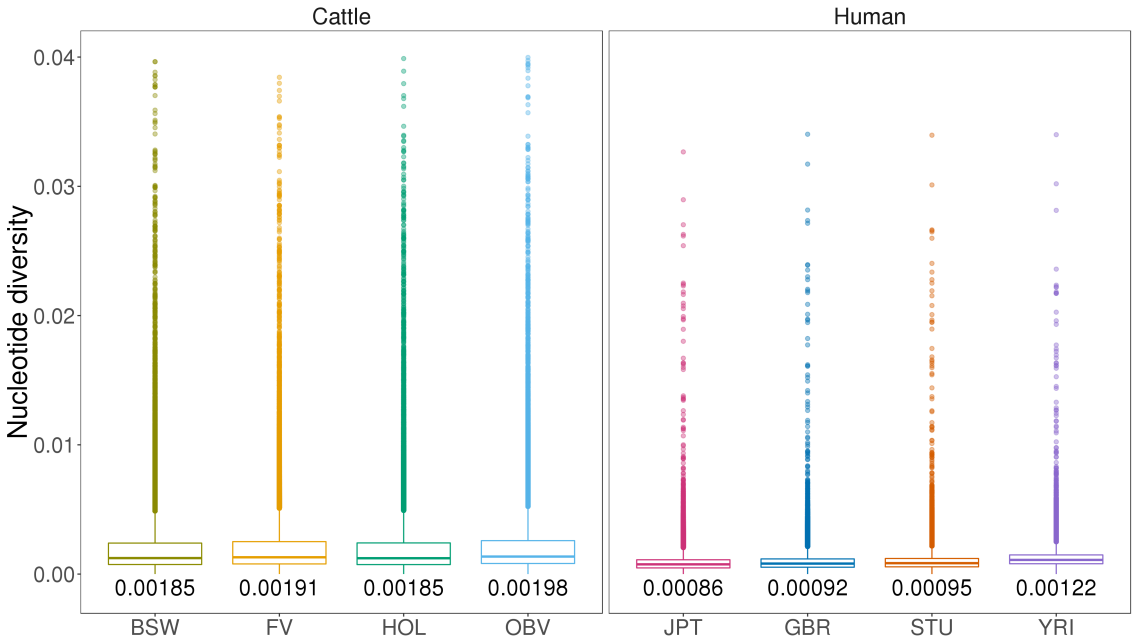


Figure S3.5: **Nucleotide diversity (π) based on whole genome autosomal variants in cattle and human.**
Nucleotide diversity (π) from each population calculated using vcftools with 10 kb non-overlapped windows based on whole genome autosomal variants. Number under the boxplot indicates average across windows.

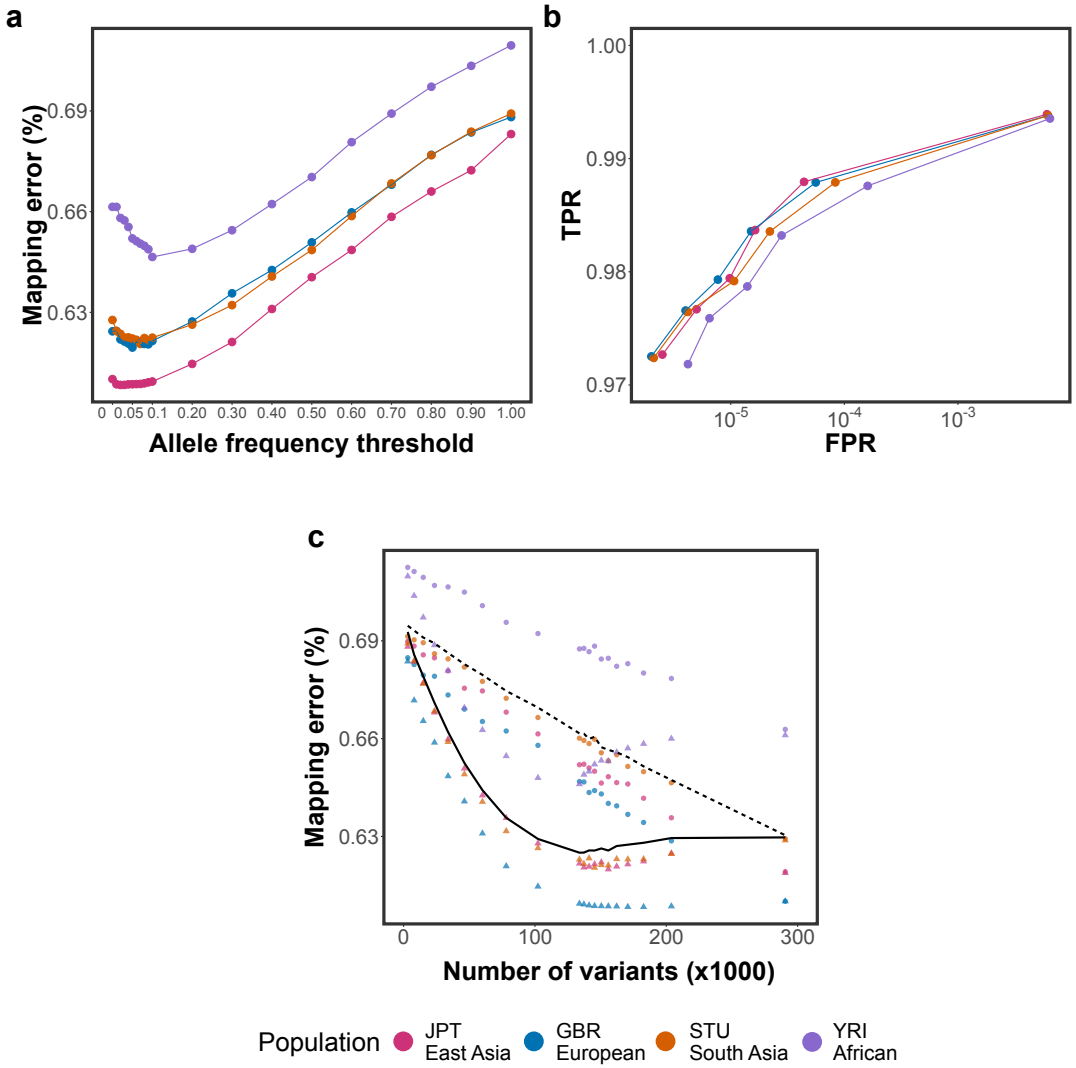


Figure S3.6: **Single-end mapping accuracy using four human population-specific augmented graphs.**
(a) Proportion of incorrectly mapped reads for four populationspecific augmented genome graphs (b) True positive (sensitivity) and false positive mapping rate (specificity) parameterized based on the mapping quality for the best performing graph from each population. (c) Read mapping accuracy for population specific augmented graphs that contained variants that were either filtered for alternate allele frequency (triangles) or sampled randomly (circles) from all variants detected within a population. The dashed and solid line represents the average proportion of mapping errors across four populations using variant prioritization and random sampling.

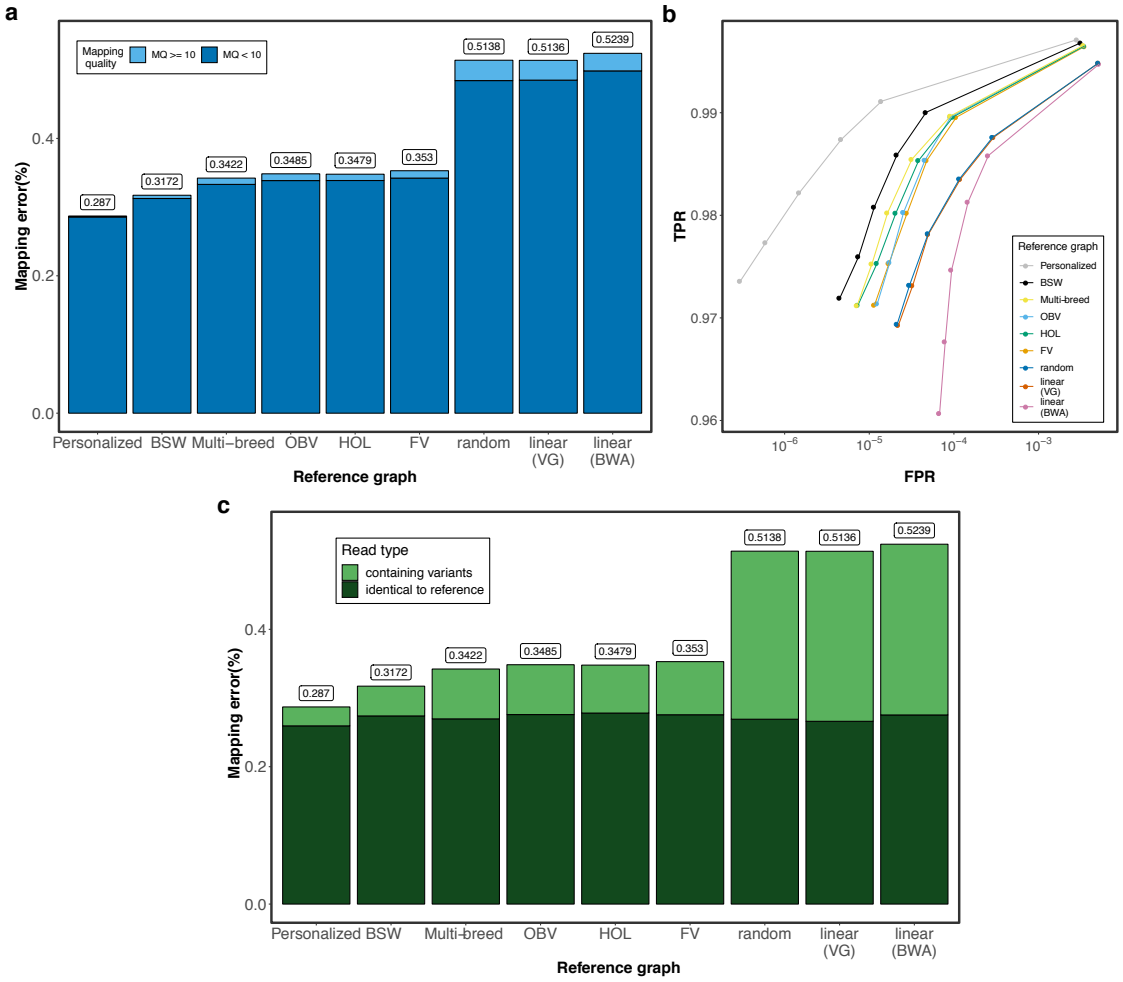


Figure S3.7: **The accuracy of mapping simulated BSW single-end reads to variation-aware and linear reference structures.**
(a) Proportion of BSW single-end reads that mapped erroneously against breed-specific augmented graphs, random graphs or linear reference sequences. Dark and light blue colours represent the proportion of incorrectly mapped reads with mapping quality (MQ)<10 and MQ>10, respectively. (b) True positive (sensitivity) and false positive mapping rate (specificity) parameterized based on the mapping quality. (c) Dark and light green colours represent the proportion of incorrectly mapped reads that matched corresponding reference nucleotides and contained non-reference alleles, respectively

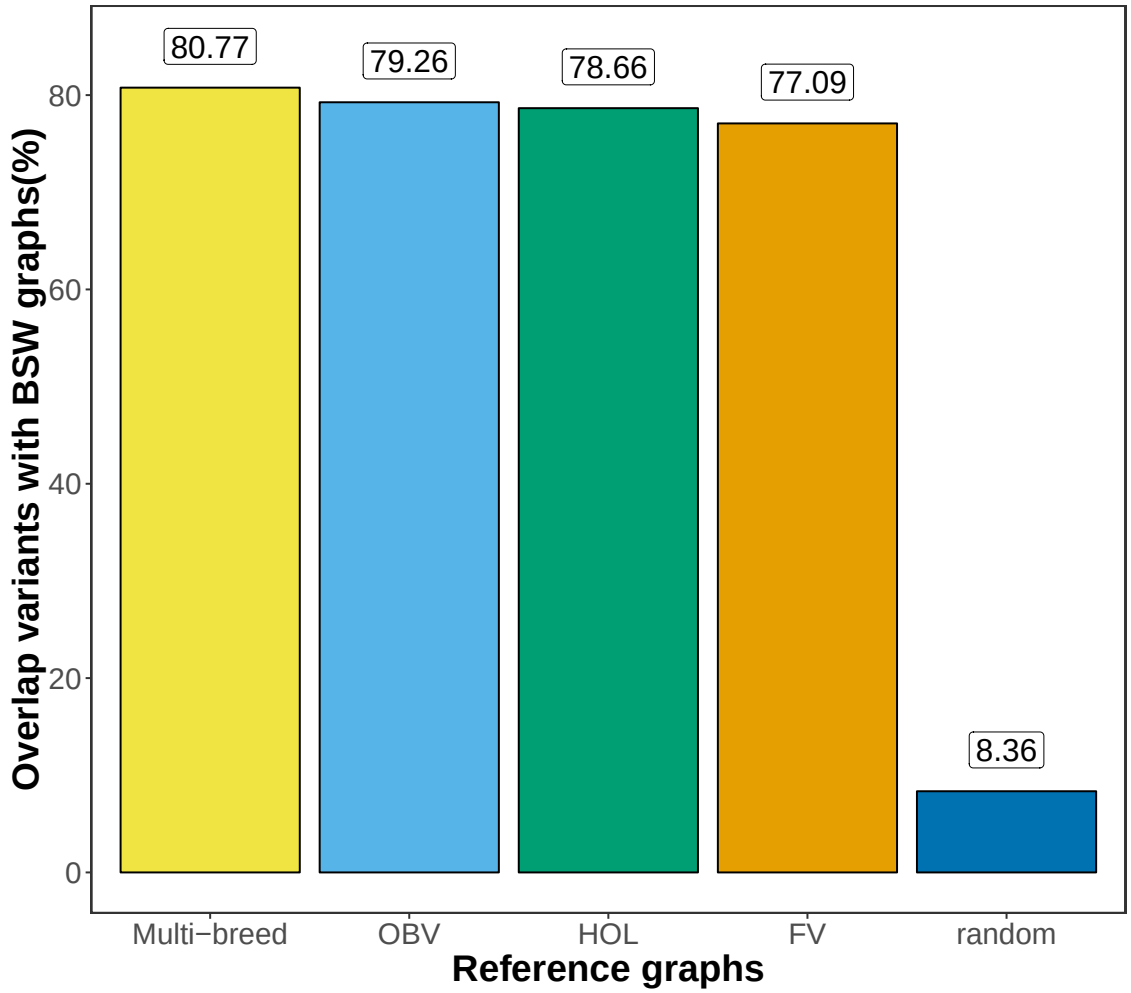


Figure S3.8: **Overlap of the variants**
(N=243,145) between the BSW-and all other variation-aware reference graphs. The values are averaged across 10 replicates.

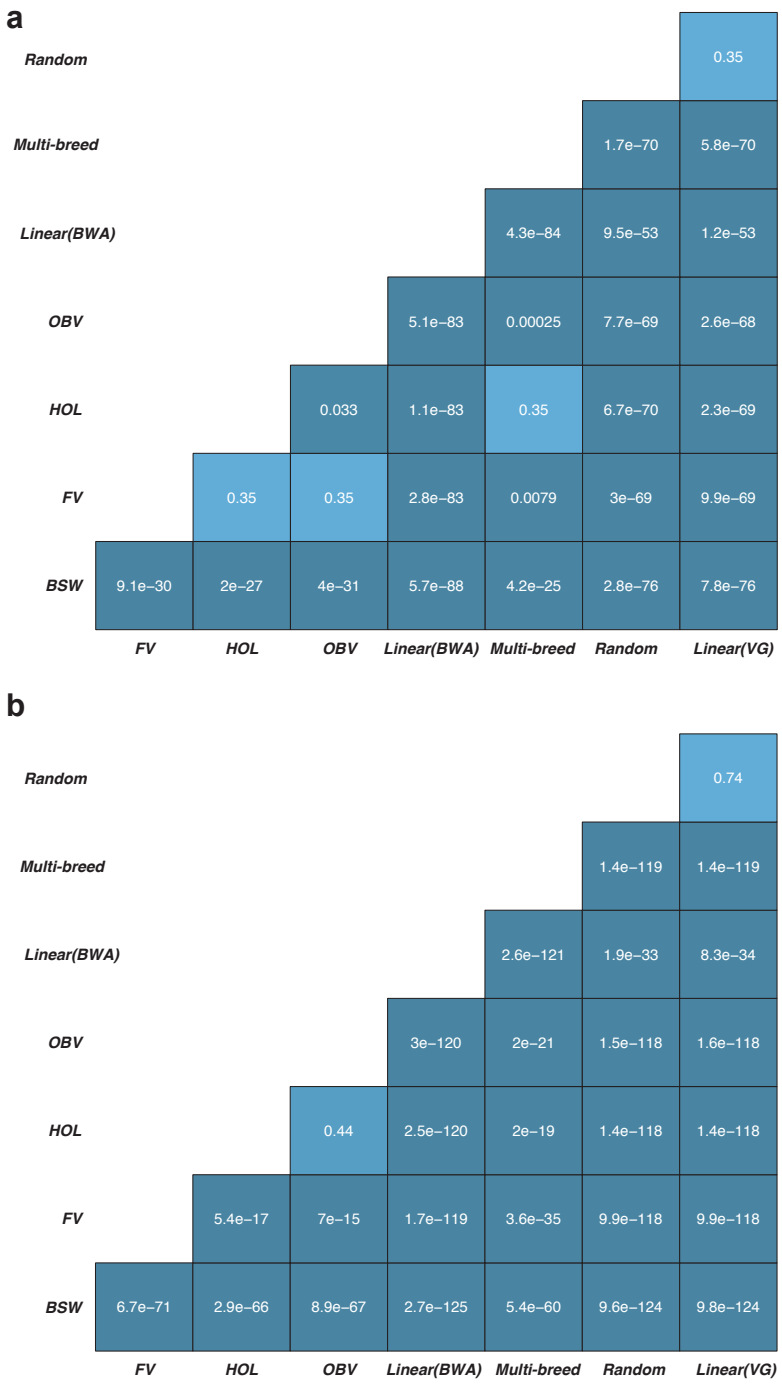


Figure S3.9: **Pairwise heatmap of *P-values* from *t tests***
comparing 8 graph-based mapping scenarios for (a) paired- and (b) single-end reads. The P-values are adjusted for multiple testing using Bonferroni-correction.

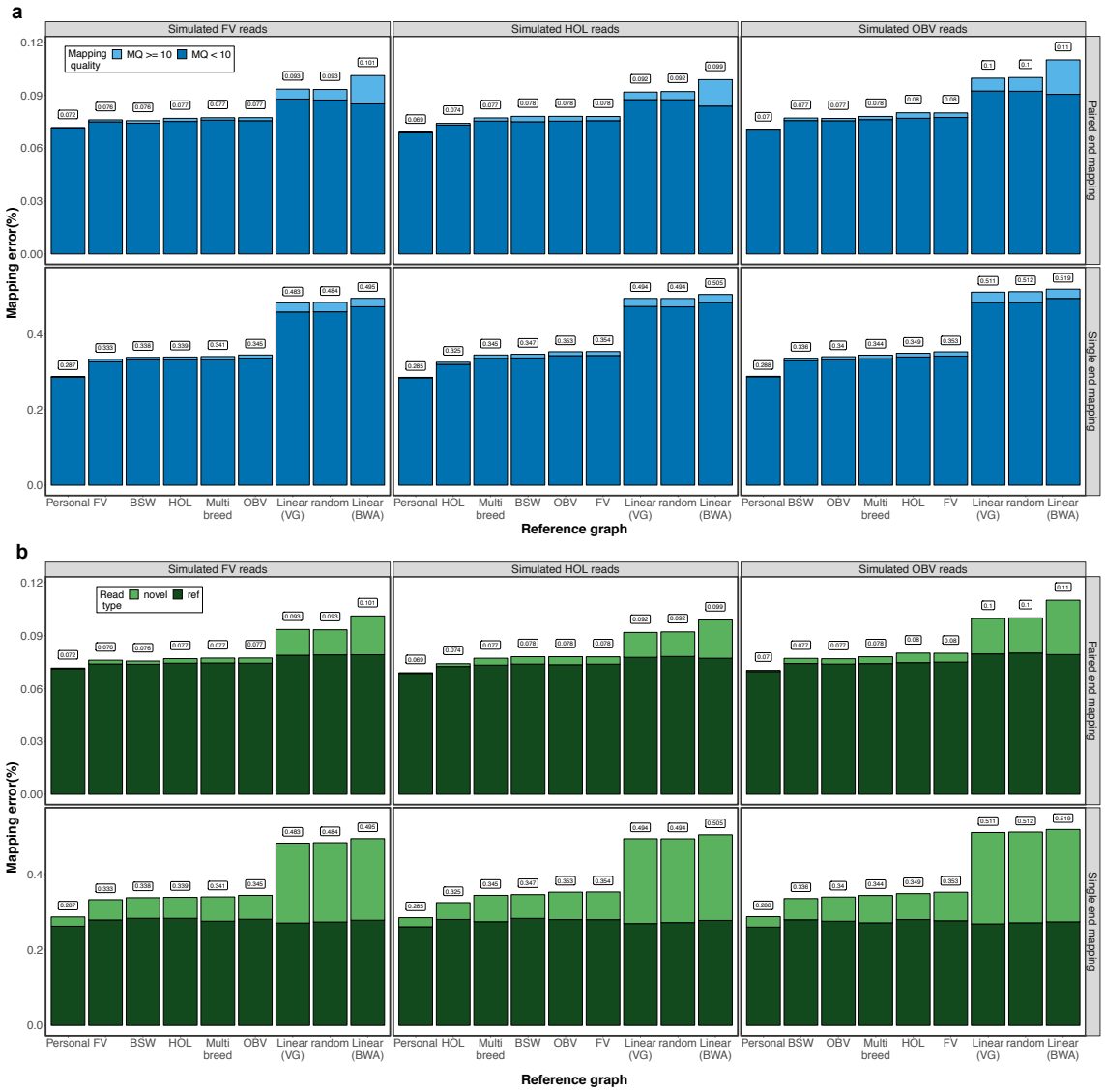


Figure S3.10: **The accuracy of mapping simulated FV, HOL and OBV reads to variation-aware and linear reference structures.**
(a) Proportion of reads that mapped erroneously against personalized graphs, breed-specific augmented graphs, random graphs or linear reference sequences. Dark and light blue colours represent the proportion of incorrectly mapped reads with mapping quality (MQ)<10 and MQ>10, respectively. The upper and lower panels reflect paired-end and single-end reads, respectively. (b) Dark and light green colours represent the proportion of incorrectly mapped reads that matched corresponding reference nucleotides and contained non-reference alleles, respectively. The upper and lower panels reflect paired-end and single-end reads, respectively

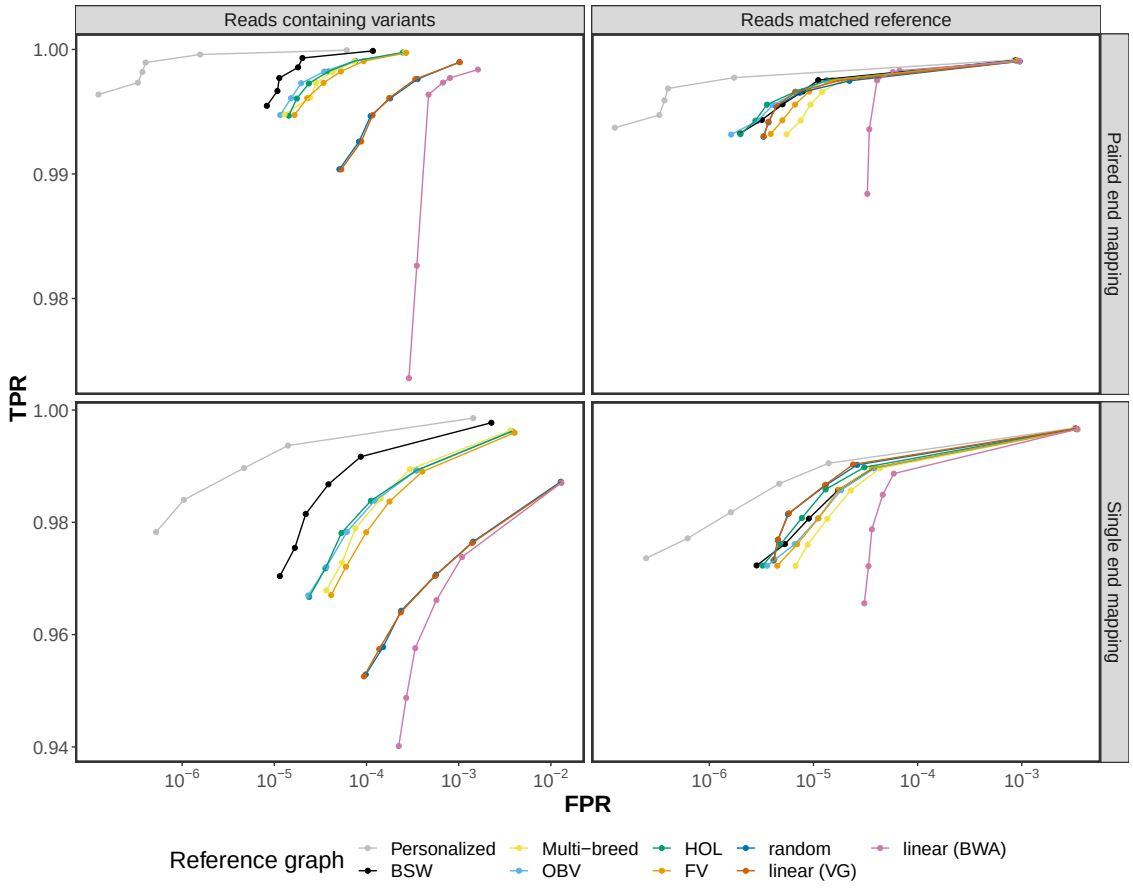


Figure S3.11: **ROC curves split by read's novelty**
Cumulative *True positive* and *False positive rate* at different mapping quality thresholds visualized as Receiver Operating Characteristic (ROC) curves for reads than contain variants and match corresponding reference alleles. The upper and lower panels represent results from paired- and single-end reads.

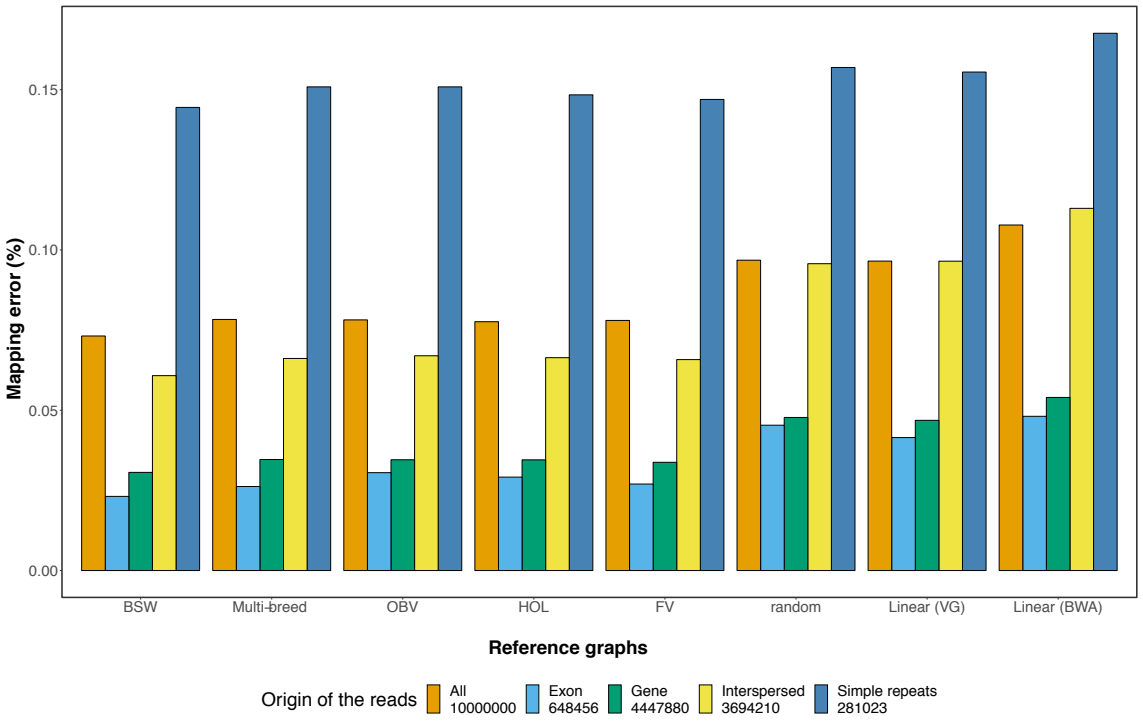


Figure S3.12: **Mapping accuracy for reads originating from different genomic features.** The origin of 10 million simulated reads was determined based on the Bos taurus ARS-UCD1.2 ensembl 99 annotations (exonic and genic) and the ARS-UCD1.2 repeat regions labelled by Repeat Masker (Interspersed duplications including SINEs, LINEs, LTR, and DNA transposable elements, and simple repeats which contain low-complexity and simple repetitive regions). Different colour indicates the proportion of erroneously mapped reads for each annotation category. The orange bars represent the average proportion of mis-mapped reads for six graph-based (BSW, Multi-breed, OBV, HOL, FV, random) and two linear (VG, BWA) reference structures. Reads were simulated from haplotypes of a BSW individual.

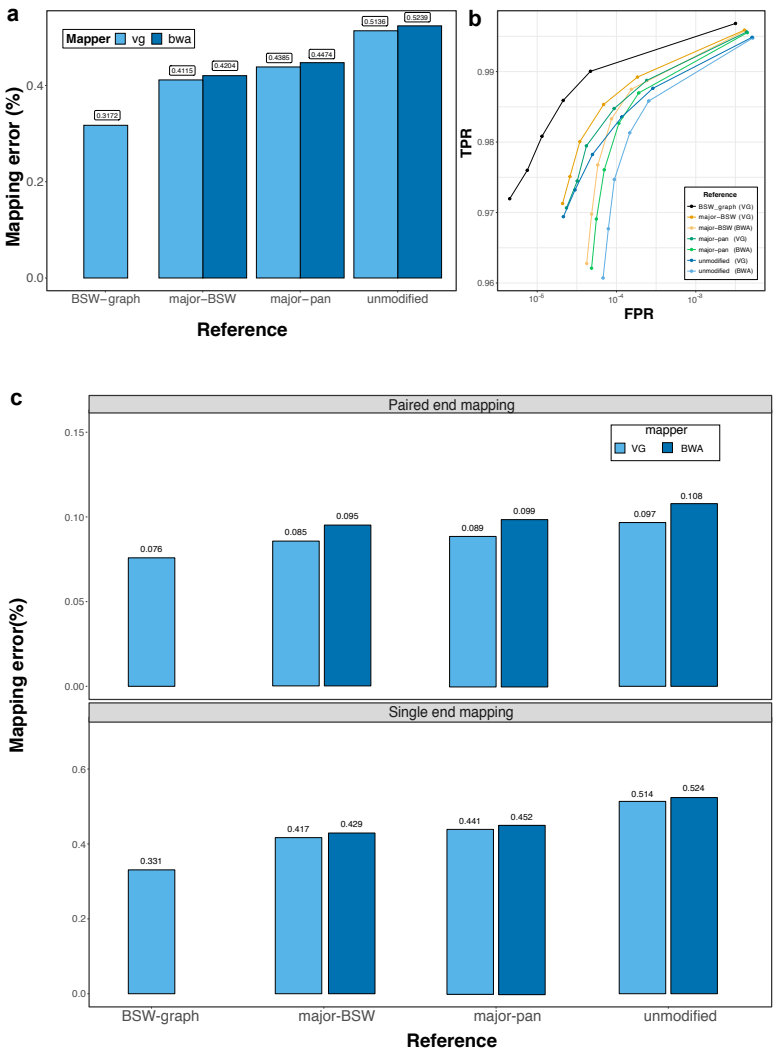


Figure S3.13: **Single-end read mapping accuracy using breed-specific augmented genome graphs and consensus linear reference sequences.**
(a) Dark and light blue represent the proportion of reads that mapped incorrectly using *BWA mem* and *vg*, respectively, to the BSW-specific augmented reference graph (BSW-graph), the BSW-specific (major-BSW) and multi-breed linear consensus sequence (major-pan) and the bovine linear reference sequence (unmodified). (b) True positive (sensitivity) and false positive mapping rate (specificity) parameterized based on the mapping quality. (c) Paired- and single-end read mapping accuracy using breed-specific augmented genome graphs and consensus linear reference sequences that were only adjusted at SNPs.

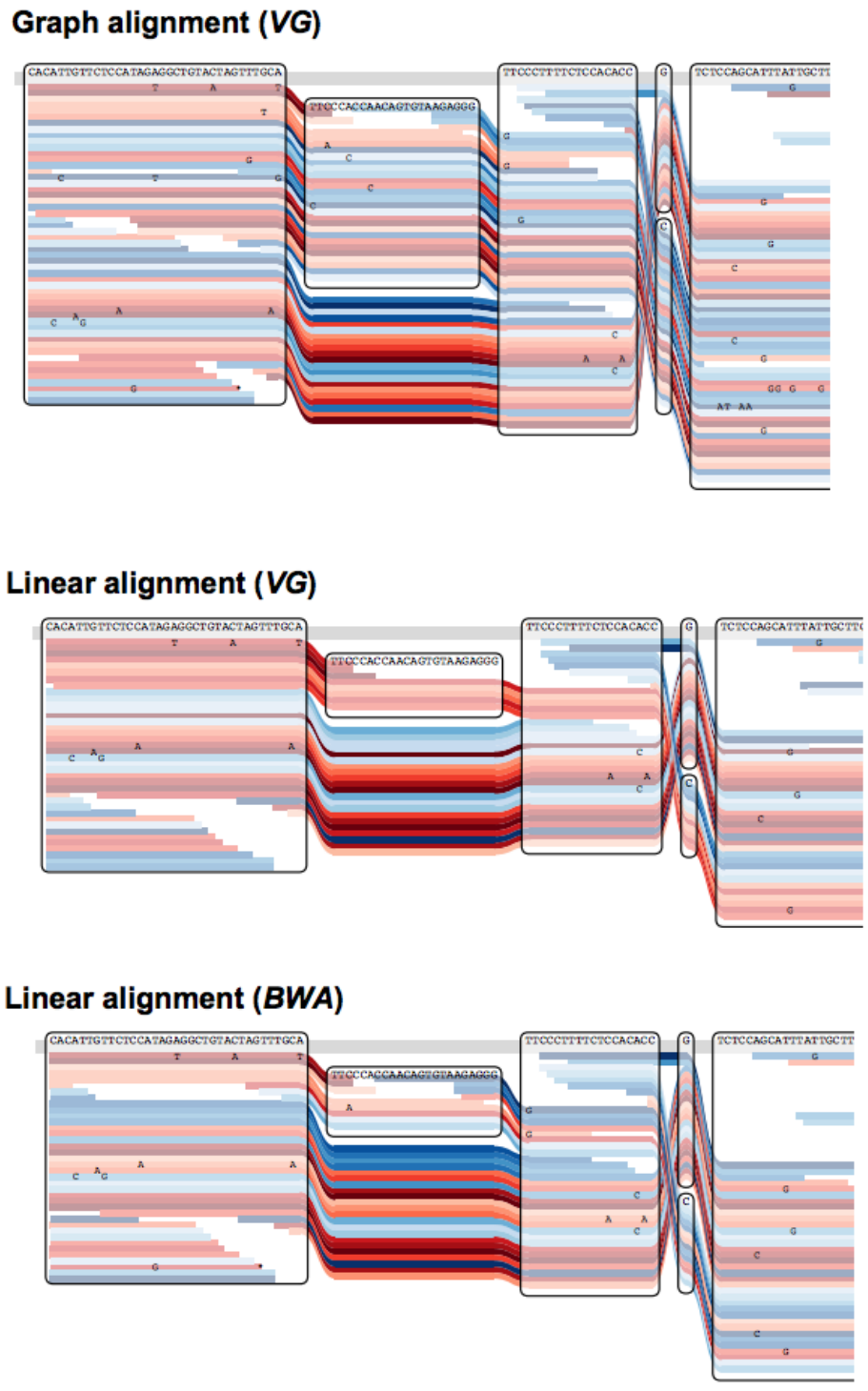


Figure S3.14: **Graph alignment visualization.**
Visualization of a 23-bp insertion at Chr10: 5,941,270 in graph and linear alignments using the *sequence tube map* tool [1]. The variant was called heterozygous from the linear alignment, but the allelic ratio was highly biased towards the reference allele. Visual inspection suggests that more reads supporting the alternate allele are present in the graph alignments. Red and blue colour indicates forward and reverse reads, respectively. The reads from the linear alignment were realigned to the variation-aware graph for the purpose of the visualisation.

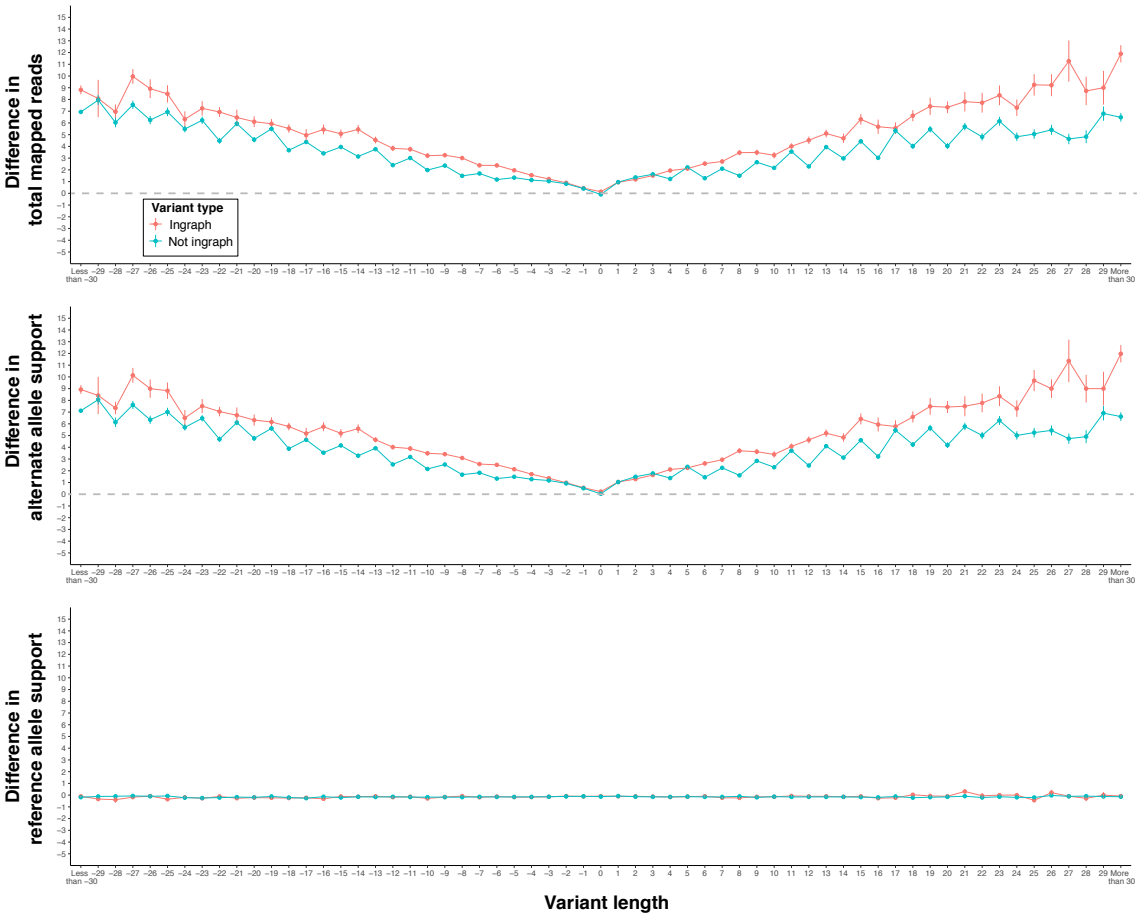


Figure S3.15: **Difference in the total of mapped reads, and reads support for reference and alternate alleles**
between the graph-based and BWA alignments for deletions, SNPs and insertions. Positive values indicate a larger number of reads for graph-based alignments. The dashed grey line indicates equal support for graphbased and linear alignments. The circles represent the mean (± standard error of mean) values at a given variant length. Red and green colour indicates that the alternate allele is included and not included in the graph, respectively.

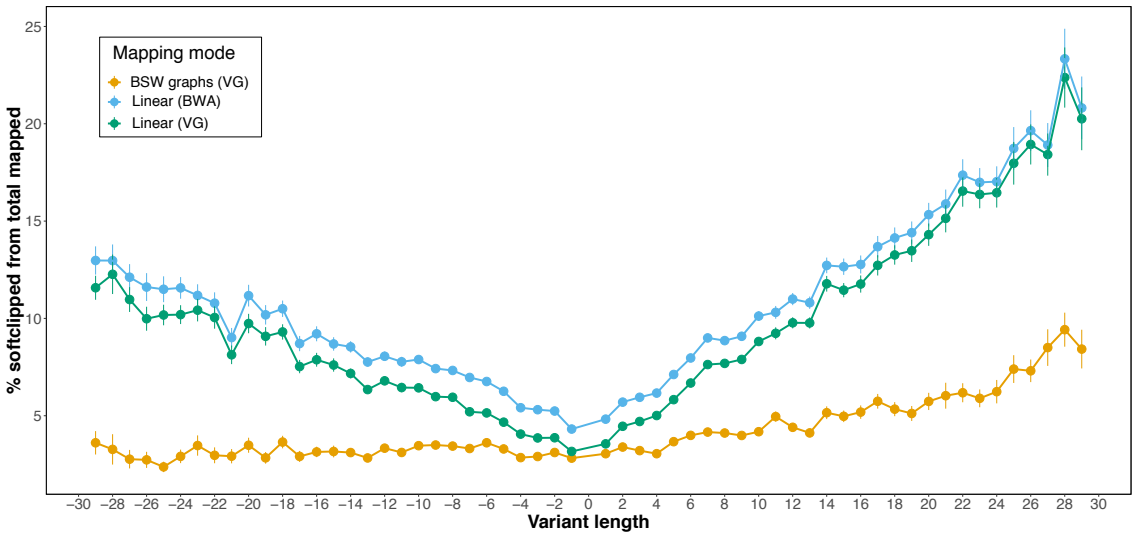


Figure S3.16: **Proportion of soft-clipped reads at heterozygous sites in graph (*vg*) and linear (*vg* and *BWA*) alignments.**
We considered only variants for which the alternate allele was already included in the graph. The circles represent the mean (± standard error of mean) values at a given variant length.

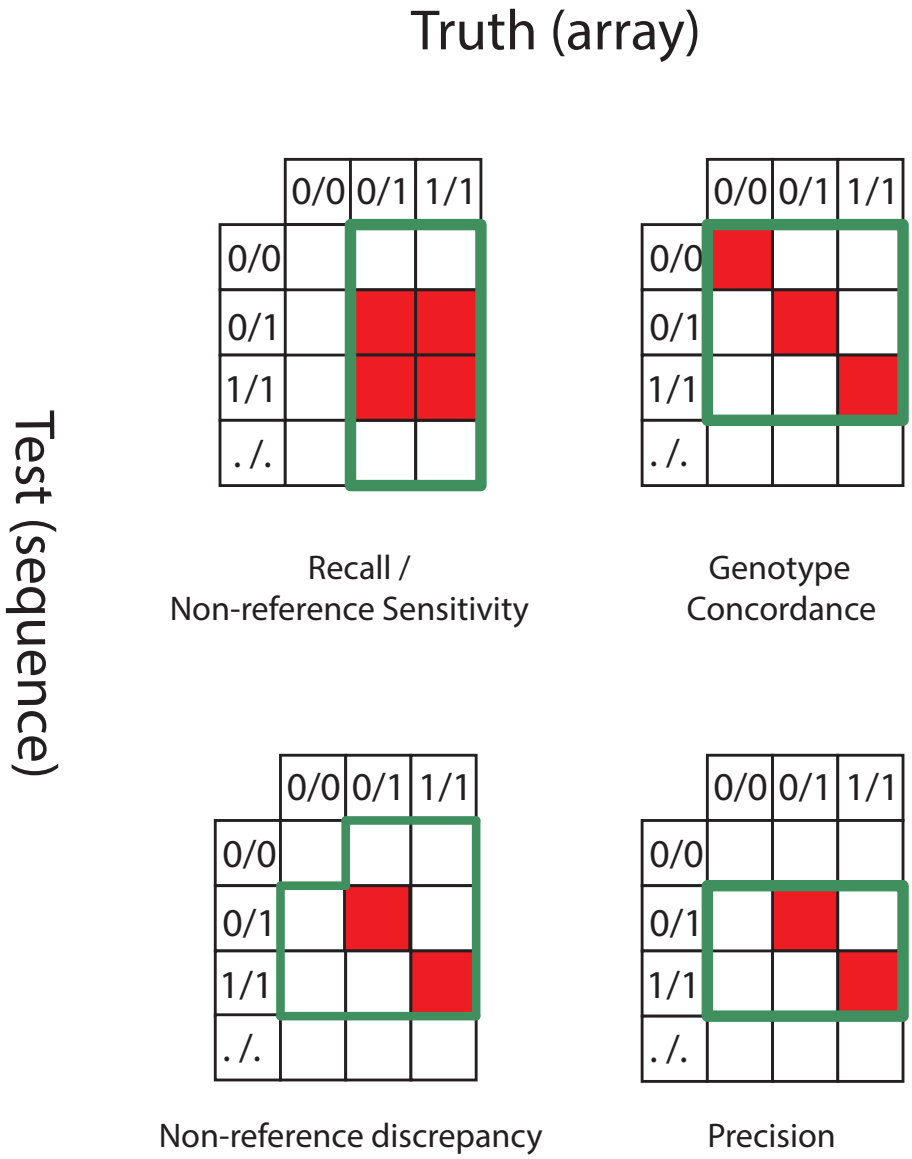


Figure S3.17: **Genotype concordance matrices for four quality parameters.**
For each metric, we divided the sum of the red cells by the sum of the cells within the green frame.

Note S3.1

Comparison of variant prioritization approaches

We applied FORGe [2] to prioritize variants to be added to the Brown Swiss reference graph for chromosome 25. Specifically, we considered the four variant ranking approaches implemented in FORGe and compared the mapping accuracy from the resulting graphs with a graph that was constructed with variants selected based on an allele frequency threshold.

The following prioritization approaches were investigated:

1. Pop Cov: variants ranked based on allele frequency
2. Pop Cov + blowup: variants ranked based on allele frequency and proximity (variants that are nearby receive lower scores)
3. Hybrid: variants ranked based on allele frequency and how the variants affect the resulting k-mer profile of the genome graph (variants that would increase the repetiveness of the resulting graph receive lower scores)
4. Hybrid + blowup: hybrid methods + considering variant proximity
5. AF threshold: variants ranked based on allele frequency (AF, as applied in our paper).

We refer to the FORGe paper [2] for a detailed description on the implementation of the variant prioritization methods 1-4. For each prioritization approach, we constructed a number of graphs that included the top x% of the ranked variants, where x ranged from 1 to 100 with steps of 10 (e.g., a graph constructed with x=10 included 34,715 out of 347,147 bta25 Brown Swiss variants). We then mapped paired-end reads simulated form a Brown Swiss animal (as detailed in the Material and Methods part of the main manuscript) to the graphs in order to calculate mapping accuracy.

Graphs constructed with variants that were prioritized solely using allele frequency (as applied in our current paper and the Pop Cov method of FORGe) enable the most accurate mapping of reads (Table SN31 and Figure SN31). Considering additional factors other than allele frequency did not lead to further accuracy improvements. The mapping accuracy of the Pop Cov and AF threshold strategies was virtually identical when the same number of variants was used. The most accurate Pop Cov approach corresponds to an alternate allele frequency threshold of 0.06.

Table SN31: Comparison of the most accurate graph from each ranking method

Ranking methods	Minimum mapping error	Number of variants in the graphs with maximum accuracy
PopCov	0.0722	208288
PopCov + blowup	0.0730	208288
Variant frequency	0.0723	208288
Hybrid	0.0749	347147
Hybrid + blowup	0.0749	347147

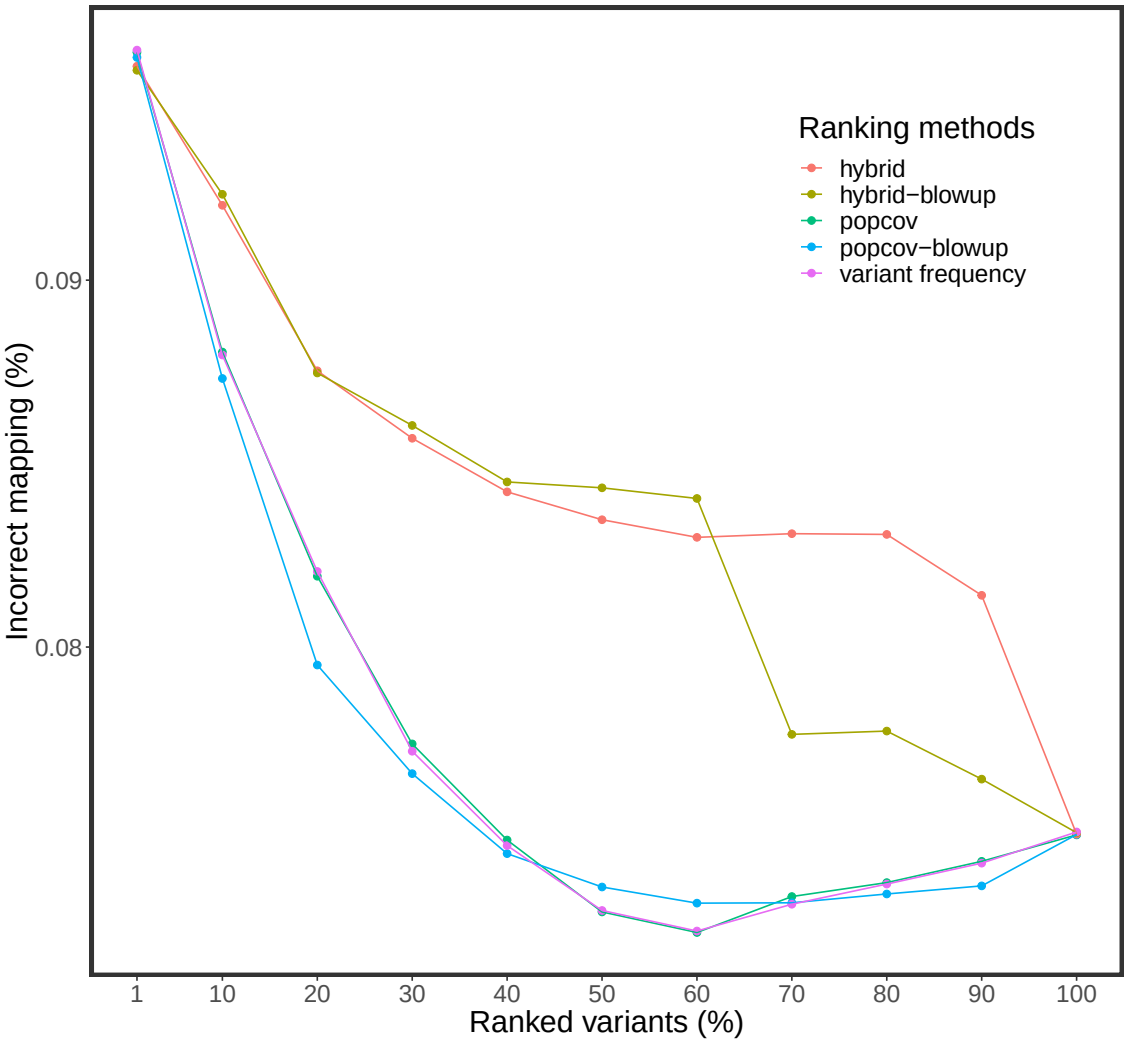


Figure SN31: Comparison of different variant prioritization strategies.
Proportion of incorrectly mapped reads for graphs constructed with five variant prioritization approaches.

Note S3.2

Adjusted (tuned) linear mapping approach

We followed the proposed approach outlined by Grytten et al. [3] to adjust the default parameters of *BWA mem* in order to also consider sub-optimal alignments. First, we reduce the D value (default 0.5) to consider more alternative alignment positions. However, the mapping performance changed only marginally.

Second, we ran *Minimap2* in short read mode (-ax sr) to find all suboptimal alignments. Subsequently, we retained for each read the read placement from either *BWA mem* or *Minimap2* that had the higher alignment score. For reads that had identical alignment score and position for both linear mappers, we retained the lower mapping quality score. For all other cases, we retained the *BWA mem* alignment.

We made two observations (Figure SN32):

1. The overall mapping accuracy increased mainly due to a smaller number of incorrectly placed reads that had high mapping quality (MQ $>$ 10). This indicates that the tuned linear mapping approach assigns the quality of the alignments better.
2. We found an improvement in mapping accuracy only on reads that are identical to the reference, but not on reads that contain variants.

While Grytten et al. [3] observed that an adjusted parameter setting of *BWA mem* and subsequent application of *Minimap2* led to considerable accuracy improvements, the gain in accuracy was low in our study. The proportion of simulated reads with variants was twice as high (19.16% vs. 10.6%) in our study than in Grytten et al. [3], because the average number of polymorphic sites per genome was almost two-fold higher in cattle than humans.

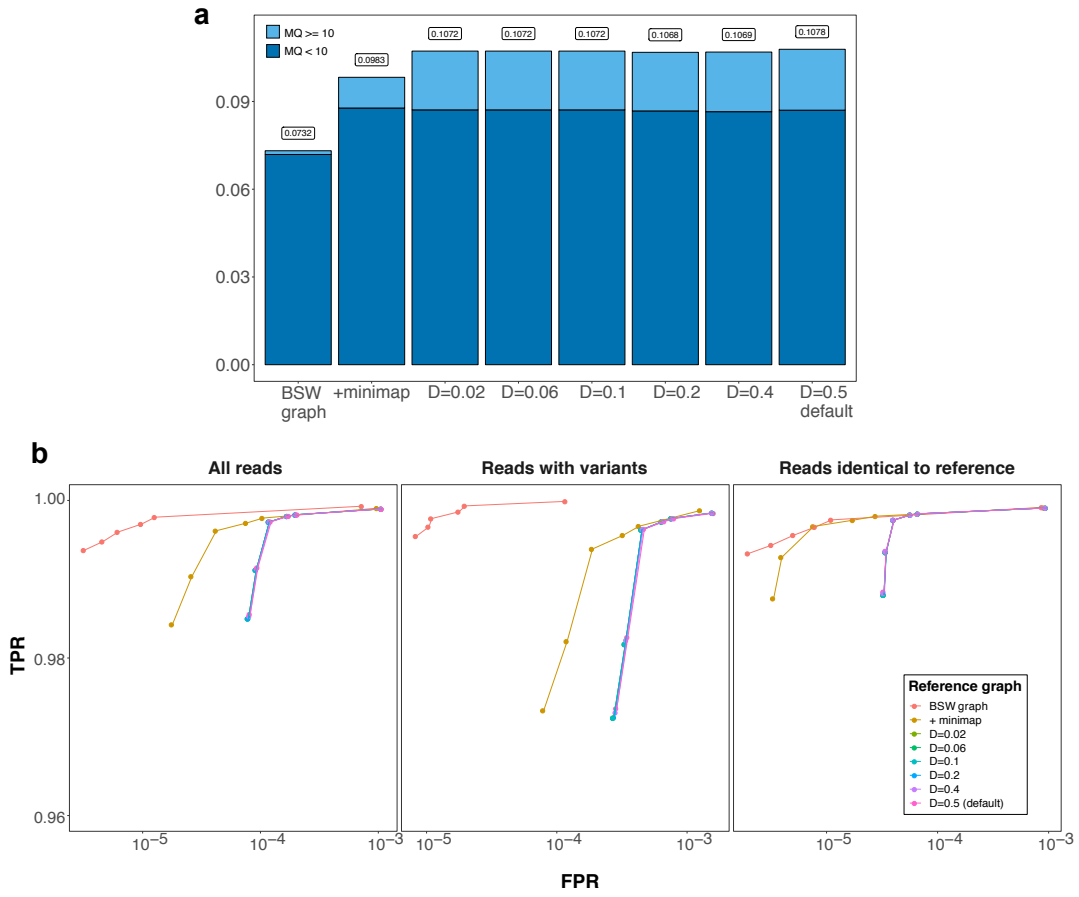


Figure SN32: Mapping accuracy of paired-end reads simulated form a Brown Swiss animal using different mapping approaches.
(a) Proportion of simulated reads with mapping errors for different mapping scenarios. (b) True positive and false positive rate parameterized on mapping quality for the different scenarios.

Note S3.3

Integrating structural variants into the graphs

We investigated the effect of including longer (structural) variants. For this purpose, we first called and genotyped structural variants using *Delly* [4] from 82 Brown Swiss samples that had been sequenced using short-reads (see Material and Methods part of the main manuscript). We discovered 157 precise SVs on bovine chromosome 25 that had an average length of 178 bp. We then combined these variants with 243,145 SNPs and Indels that were discovered using *GATK*. We used the bta25 ARS-UCD1.2 reference as a backbone and constructed four graphs: (i) SNPs (+Indels) from *GATK*, (i) SVs from *Delly*, (iii) SNPs (+Indels) from *GATK* + SVs from *Delly*, (iv) empty (only the backbone, no variants). We simulated 10 million paired end reads from haplotypes of one Brown Swiss animal (SAMEA6272105, that had 121,996 SNPs + Indels and 57 SVs that were included in the graph). The simulated reads were mapped to the different graphs using *vg*.

Table SN32: Mapping accuracy for graphs that contained different variant types
MQ=0 and MQ $<$ 10 indicates the proportion of reads mapped with mapping quality 0 and less than 10, respectively.

Graphs	Variants in the graphs	MQ=0 (%)	MQ<10 (%)	Mapping error (%)
Linear	0	0.15474	0.22310	0.08599
SNP	243,145	0.15366	0.21804	0.07995
SV	157	0.15508	0.22390	0.08629
SNP + SV	243,145 + 157	0.15458	0.21900	0.08003

Adding SVs that were detect from short sequencing reads to the graph marginally affected the mapping performance. Actually, the mapping accuracy decreased slightly when SVs were added. Read mapping accuracy improvements were attributable to the SNPs and Indels detected using *GATK*.

Table S3.1: **Properties of autosomal variants detected in human (JPT, GBR, STU, YRI) and bovine (HOL, FV, BSW, OBV) populations**

Species	Population	Number of samples	Variant count	Average per sample	Singleton variants	Variants with allele frequency < 0.05
Human	JPT	104	12,433,397	4,020,815	2,836,542 (22.81%)	5,580,288 (44.88%)
	GBR	91	13,148,448	4,011,102	2,878,144 (21.88 %)	6,005,303 (45.67%)
	STU	102	15,264,479	4,096,457	4,024,478 (26.34%)	7,915,678 (51.85%)
	YRI	108	22,420,039	4,863,955	4,702,120 (20.97%)	12,431,887 (55.45%)
Cattle	HOL	49	16,762,842	6,841,965	1,713,642 (10.22%)	3,964,699 (23.65%)
	FV	49	18,638,951	6,955,100	2,272,546 (12.19%)	5,112,547 (27.42%)
	BSW	82	20,446,693	6,983,517	3,957,703 (19.35%)	7,913,226 (38.70%)
	OBV	104	21,875,164	7,111,562	3,124,950 (14.28%)	8,250,961 (37.71%)

Table S3.2: **Properties of variants detected on human chromosome 19 and bovine chromosome 25 in human (JPT, GBR, STU, YRI) and bovine (HOL, FV, BSW, OBV) populations**

Species	Population	Number of samples	Variant count	Average per sample	Singleton variants	Variants with allele frequency < 0.05
Human	JPT	104	291,303	88,945	66,944 (22.98%)	135,289 (46.44%)
	GBR	91	306,304	90,988	64,119 (20.93 %)	138,076 (45.07%)
	STU	102	355,107	94,253	93,116 (26.22%)	181,300 (51.05%)
	YRI	108	521,021	118,429	106,734 (20.49%)	280,960 (53.92%)
Cattle	HOL	49	295,801	121,114	30,543 (10.32%)	67827 (22.92%)
	FV	49	336,390	125,597	43,783 (13.01%)	94,577 (28.11%)
	BSW	82	347,402	124,209	53,773 (15.47%)	128,990 (37.12%)
	OBV	104	387,855	126,158	47,498 (12.24%)	144,958 (37.37%)

Table S3.3: **Concordance between array-called and sequence variant genotypes that were discovered from either graph or linear alignments using *Samtools*, *GATK*, or *Graphtyper*.**
Numbers represent average values (± standard deviation) of 10 BSW animals for the raw (Full) and hard-filtered (Filtered) genotypes.

	Full			Filtered		
Samtools						
	Graph	Linear	Linear	Graph	Linear	Linear
	VG	VG	BWA	VG	VG	BWA
Genotype concordance	98.50(1.07)	98.47(1.07)	98.53(1.03)	98.53(1.07)	98.50(1.07)	98.55(1.04)
NR-sensitivity (Recall)	98.53(0.37)	98.52(0.39)	98.53(0.39)	97.48(0.36)	97.45(0.35)	97.53(0.36)
NR-discrepancy	2.21(1.60)	2.24(1.60)	2.17(1.55)	2.17(1.60)	2.20(1.61)	2.13(1.56)
Precision	98.90(0.83)	98.89(0.83)	98.93(0.81)	98.91(0.83)	98.90(0.83)	98.94(0.82)
GATK						
Genotype concordance	97.26(2.24)	97.24(2.25)	97.38(2.15)	97.26(2.25)	97.25(2.25)	97.39(2.15)
NR-sensitivity (Recall)	98.17(0.94)	98.16(0.94)	98.23(0.87)	98.14(0.94)	98.12(0.94)	98.18(0.87)
NR-discrepancy	4.09(3.38)	4.10(3.39)	3.89(3.23)	4.08(3.38)	4.09(3.39)	3.88(3.23)
Precision	98.90(0.83)	98.90(0.83)	98.94(0.80)	98.91(0.83)	98.91(0.83)	98.95(0.80)
Graphtyper						
Genotype concordance	98.57(1.01)	98.57(1.01)	98.61(0.97)	98.61(1.03)	98.61(1.03)	98.64(0.99)
NR-sensitivity (Recall)	98.34(0.54)	98.36(0.55)	98.37(0.53)	96.14(0.54)	96.13(0.54)	96.17(0.52)
NR-discrepancy	2.08(1.49)	2.08(1.50)	2.02(1.44)	2.01(1.50)	2.01(1.50)	1.97(1.45)
Precision	98.85(0.80)	98.84(0.81)	98.87(0.79)	98.89(0.82)	98.89(0.82)	98.91(0.80)

Table S3.4: **Accession numbers of the animals** used for variant detection, read simulation, sequence read mapping and genotyping

Accession	Breed	Graph construction	Simulation	Validation genotyping	Validation Mapping bias	Coverage
SAMEA4827645	OBV	x				14.41
SAMEA4827646	OBV	x				12.9
SAMEA4827647	OBV	x				14.79
SAMEA4827648	OBV	x				10.76
SAMEA4827649	OBV	x				11.55
SAMEA4827650	OBV	x				10.29
SAMEA4827651	OBV	x				14.76
SAMEA4827652	OBV	x				10.65
SAMEA4827653	OBV	x				9.69
SAMEA4827654	OBV	x				10.72
SAMEA4827655	OBV	x				11.32
SAMEA4827656	OBV	x				11.83
SAMEA4827657	OBV	x				8.47
SAMEA4827658	OBV	x				9.69
SAMEA4827659	OBV	x				9.52
SAMEA4827660	OBV	x				10.04
SAMEA4827661	OBV	x				9.68
SAMEA4827662	OBV	x				17.37
SAMEA4827663	OBV	x				11.2
SAMEA4827664	OBV	x				11.29
SAMEA4827665	OBV	x				13.07
SAMEA4827666	OBV	x				11.23
SAMEA4827667	OBV	x				10.99
SAMEA4827668	OBV	x				10.93
SAMEA4827669	OBV	x				12.89
SAMEA4827670	OBV	x				12.18
SAMEA4827671	OBV	x				11.35
SAMEA4827672	OBV	x				10.49
SAMEA4827673	OBV	x				10.31
SAMEA4827674	OBV	x				12.58
SAMEA5059741	OBV	x				4.58
SAMEA5059742	OBV	x				3.76
SAMEA5059743	OBV	x	x			22.33
SAMEA5059744	OBV	x				3.93
SAMEA5059745	OBV	x				4.31
SAMEA5059746	OBV	x				4.29
SAMEA5059747	OBV	x				4.58
SAMEA5059748	OBV	x				5.08
SAMEA5059749	OBV	x				5.19
SAMEA5059750	OBV	x				3.91
SAMEA5059751	OBV	x				5.59
SAMEA5059752	OBV	x				3.89
SAMEA5059753	OBV	x				4.18
SAMEA5059754	OBV	x				3.49
SAMEA5059755	OBV	x				7.49
SAMEA5059756	OBV	x				6.65
SAMEA5059757	OBV	x				5.74
SAMEA5059758	OBV	x				5.1
SAMEA6272117	OBV	x				6.43
SAMEA5059759	OBV	x				3.97
SAMEA5159792	BSW	x				10.68
SAMEA5159791	BSW	x				10.22

Continuation of Table S3.4

Accession	Breed	Graph construction	Simulation	Validation genotyping	Validation Mapping bias	Coverage
SAMEA5159788	BSW	x				10.71
SAMEA5159783	BSW	x				11.91
SAMEA5159785	BSW	x				11.94
SAMEA5159799	BSW	x				10.25
SAMEA5159787	BSW	x				13.63
SAMEA5159761	BSW	x				16.46
SAMEA5159782	BSW	x				11.47
SAMEA5159775	BSW	x				10.14
SAMEA5159786	BSW	x				12.04
SAMEA5159784	BSW	x				11.88
SAMEA5159798	BSW	x				12.79
SAMEA5159781	BSW	x				12.65
SAMEA5159780	BSW	x				12.41
SAMEA5159777	BSW	x				9.8
SAMEA5159797	BSW	x				11.98
SAMEA5159774	BSW	x				9.46
SAMEA5159769	BSW	x				12.3
SAMEA5159778	BSW	x				13.03
SAMEA5159771	BSW	x				10.92
SAMEA5159779	BSW	x				10.63
SAMEA5159772	BSW	x				11.88
SAMEA5159773	BSW	x				10.77
SAMEA5159793	BSW	x				12.6
SAMEA5159770	BSW	x				10.01
SAMEA5159795	OBV	x				12.58
SAMEA5159768	OBV	x				8.69
SAMEA5159796	OBV	x				11.39
SAMEA5159789	OBV	x				10.27
SAMEA5159790	OBV	x				10.52
SAMEA5159794	OBV	x				11.46
SAMEA5159776	OBV	x				9.71
SAMEA5159767	OBV	x				10.17
SAMN05216093	OBV	x				10.85
SAMN05216095	OBV	x				11.12
SAMN05216094	OBV	x				10.64
SAMN05216096	OBV	x				11.51
SAMEA6272131	FV	x				13.4
SAMEA6272130	FV	x				10.41
SAMEA4644727	BSW	x				14.86
SAMEA4644728	BSW	x				14.86
SAMEA19864918	BSW	x				9.23
SAMEA4644765	BSW	x				12.14
SAMEA4644766	BSW	x				16.48
SAMEA4644768	OBV	x				13.41
SAMEA4644769	BSW	x				16.04
SAMEA19312918	BSW	x				4.43
SAMEA19313668	BSW	x				7.13
SAMEA19314418	BSW	x				10.99
SAMEA19315168	BSW	x				9.7
SAMEA19318918	BSW	x				6.9
SAMEA19323418	BSW	x				18.83
SAMEA4644754	BSW	x				15.25
SAMEA4644755	BSW	x				13.58
SAMEA4644756	BSW	x				13.88

Continuation of Table S3.4

Accession	Breed	Graph construction	Simulation	Validation genotyping	Validation Mapping bias	Coverage
SAMEA4644730	OBV	x				14.85
SAMEA4644734	OBV	x				15.3
SAMEA4644735	BSW	x				9.43
SAMEA4644757	BSW	x				11.36
SAMEA4644739	BSW	x				14.13
SAMEA4644740	OBV	x				15.73
SAMEA4644741	BSW	x				15.57
SAMEA4644742	BSW	x				15.68
SAMEA4644758	BSW	x				13
SAMEA4644743	BSW	x				15.46
SAMEA4644749	OBV	x				13.85
SAMEA4644750	OBV	x				15.25
SAMEA4644762	BSW	x				13.92
SAMEA4644763	BSW	x				11.62
SAMEA4644764	OBV	x				10.57
SAMN07692225	BSW	x				10.72
SAMN02671625	FV	x				5.06
SAMN02671626	FV	x	x			23.24
SAMN02671627	FV	x				6.32
SAMN02671628	FV	x				4.95
SAMN02671629	FV	x				8.41
SAMN02671630	FV	x				4.88
SAMN02671631	FV	x				4.77
SAMN02671632	FV	x				7.64
SAMN02671633	FV	x				3.59
SAMN02671634	FV	x				7.67
SAMN02671635	FV	x				6.37
SAMN02671636	FV	x				6.26
SAMN02671637	FV	x				3.79
SAMN02671638	FV	x				3.95
SAMN02671639	FV	x				7.21
SAMN02671640	FV	x				8.62
SAMN02671641	FV	x				6.08
SAMN02671642	FV	x				5.47
SAMN02671643	FV	x				5.03
SAMN02671644	FV	x				4.35
SAMN02671645	FV	x				5.06
SAMN02671646	FV	x				5.79
SAMN02671647	FV	x				5.2
SAMN02671648	FV	x				5.81
SAMN02671649	FV	x				5.32
SAMN02671650	FV	x				5.34
SAMN02671651	FV	x				4.51
SAMN02671652	FV	x				7.48
SAMN02671653	FV	x				7.5
SAMN02671654	FV	x				7.6
SAMN02671655	FV	x				7.19
SAMN02671656	FV	x				5.4
SAMN02671657	FV	x				5.61
SAMN02671658	FV	x				4.91
SAMN02671659	FV	x				4.83
SAMN02671661	FV	x				5.58
SAMN02671662	FV	x				6.08
SAMN02671663	FV	x				5.06

Continuation of Table S3.4

Accession	Breed	Graph construction	Simulation	Validation genotyping	Validation Mapping bias	Coverage
SAMN02671664	FV	x				7.95
SAMN02671665	FV	x				6.53
SAMN02671666	FV	x				6.06
SAMN02671667	FV	x				8.13
SAMN02671572	HOL	x				6.79
SAMN02671574	HOL	x				10.25
SAMN02671576	HOL	x				5.02
SAMN02671578	HOL	x				19.78
SAMN02671580	HOL	x				10.52
SAMN02671582	HOL	x				15.22
SAMN02671584	HOL	x	x			29.97
SAMN02671586	HOL	x				17.21
SAMN02671588	HOL	x				16.99
SAMN02671590	HOL	x				13.79
SAMN02671592	HOL	x				16.31
SAMN02671594	HOL	x				19.56
SAMN02671596	HOL	x				16.43
SAMN02671455	HOL	x				9.23
SAMN02671457	HOL	x				10.28
SAMN02671459	HOL	x				8.4
SAMN02671461	HOL	x				9.47
SAMN02671463	HOL	x				6.36
SAMN02671465	HOL	x				10.61
SAMN02671467	HOL	x				9.78
SAMN02671469	HOL	x				9.13
SAMN02671471	HOL	x				6.49
SAMN02671473	HOL	x				8.71
SAMN02671475	HOL	x				9.57
SAMN02671477	HOL	x				10.89
SAMN02671479	HOL	x				8.81
SAMN02671481	HOL	x				8.59
SAMN02671483	HOL	x				10.79
SAMN02671485	HOL	x				9.18
SAMN02671487	HOL	x				10.1
SAMN02671489	HOL	x				10.06
SAMN02671491	HOL	x				9.83
SAMN02671493	HOL	x				10.1
SAMN02671495	HOL	x				8.58
SAMN02671613	HOL	x				23.58
SAMN02671615	HOL	x				20.36
SAMN02671617	HOL	x				20.36
SAMN02671619	HOL	x				12.54
SAMN02671621	HOL	x				12.86
SAMN02671623	HOL	x				4.73
SAMN02671668	HOL	x				11.92
SAMN02671670	HOL	x				11.35
SAMN02671672	HOL	x				10.21
SAMN02671674	HOL	x				10.4
SAMN02671676	HOL	x				11.21
SAMN02671725	HOL	x				11.54
SAMN02671727	HOL	x				5.43
SAMN02671729	HOL	x				13.68
SAMN02671731	HOL	x				13.58
SAMEA6272085	OBV	x				8.01

Continuation of Table S3.4

Accession	Breed	Graph construction	Simulation	Validation genotyping	Validation Mapping bias	Coverage
SAMEA6272091	OBV	x				9.55
SAMEA6272090	OBV	x				10.74
SAMEA6272089	OBV	x				8.25
SAMEA6272088	OBV	x				10.97
SAMEA6272093	OBV	x				11.3
SAMEA6272087	OBV	x				11.62
SAMEA6272086	OBV	x				12.58
SAMEA6272092	OBV	x				9.38
SAMEA6272094	OBV	x				8.31
SAMEA6272115	OBV	x				8.65
SAMEA6272114	OBV	x				8.06
SAMEA6272112	OBV	x				9.51
SAMEA6272113	OBV	x				10.61
SAMEA6272110	OBV	x				7.99
SAMEA6272103	OBV	x				9.09
SAMEA6272109	OBV	x				7.97
SAMEA6272107	OBV	x				10.34
SAMEA6272102	OBV	x				7.25
SAMEA6272100	OBV	x				8.55
SAMEA6272133	FV	x				12.73
SAMEA6272134	FV	x				10.25
SAMEA6272128	FV	x				11.09
SAMEA6163196	BSW	x				11.48
SAMEA6163197	BSW	x				9.86
SAMEA6163198	BSW	x				11.63
SAMEA6163199	BSW	x				13.68
SAMEA6272129	FV	x				14.9
SAMEA6272132	FV	x				15.25
SAMEA6272119	OBV	x				19.58
SAMEA6272123	OBV	x				16.93
SAMEA6272118	OBV	x				18.66
SAMEA6272120	OBV	x				18.5
SAMEA6272121	OBV	x				16.58
SAMEA6272126	OBV	x				61.9
SAMEA6272124	OBV	x				18.82
SAMEA6272122	OBV	x				18.33
SAMEA6272127	OBV	x				53.65
SAMEA6272125	OBV	x				23.01
SAMEA6272084	OBV	x				11.78
SAMEA6272083	OBV	x				31.95
SAMEA6272082	OBV	x				23.39
SAMEA6272095	BSW	x				25.36
SAMEA6272096	BSW	x				20.6
SAMEA6272097	BSW	x				10.68
SAMEA6272098	BSW	x				15.25
SAMEA6272099	BSW	x				12.32
SAMEA6272101	BSW	x				10.4
SAMEA6272104	BSW	x				12.63
SAMEA6272105	BSW	x	x			33.7
SAMEA6272106	BSW	x				15.76
SAMEA6272108	BSW	x				20.46
SAMEA6272111	BSW	x				28.82
SAMEA6272116	BSW	x				70.04
SAMEA5159861	BSW	x				24.84

Continuation of Table S3.4

Accession	Breed	Graph construction	Simulation	Validation genotyping	Validation Mapping bias	Coverage
SAMEA5159863	BSW	x				23.64
SAMEA5159864	BSW	x				24.92
SAMEA5159865	BSW	x				25.99
SAMEA5159866	BSW	x				25.11
SAMEA5159867	BSW	x				26.28
SAMEA5159868	BSW	x				26.73
SAMEA5159869	BSW	x				27.62
SAMEA5159870	BSW	x				32.64
SAMEA5159871	BSW	x				34.49
SAMEA5159872	BSW	x				27.96
SAMEA5159873	BSW	x				24.08
SAMEA5159874	BSW	x				33.8
SAMEA5159875	BSW	x				22.66
SAMEA5159885	BSW	x				23.1
SAMEA5159837	OBV	x				28.12
SAMEA5159843	OBV	x				22.81
SAMEA5159848	OBV	x				22.5
SAMEA5159849	OBV	x				26.32
SAMEA5159850	OBV	x				27.69
SAMEA5159886	OBV	x				35.51
SAMEA6163185	BSW			x	x	39.88
SAMEA6163188	BSW			x		25.74
SAMEA6163187	BSW			x		20.29
SAMEA6163177	BSW			x		8.26
SAMEA6163178	BSW			x		5.74
SAMEA6163176	BSW			x		9.29
SAMEA6163179	BSW			x		6.93
SAMEA6163183	BSW			x		7.86
SAMEA6163181	BSW			x		7.97
SAMEA6163182	BSW			x		8.36

Supplementary References

[1] Wolfgang Beyer, Adam M Novak, Glenn Hickey, Jeffrey Chan, Vanessa Tan, Benedict Paten, and Daniel R Zerbino. Sequence tube maps: making graph genomes intuitive to commuters. *Bioinformatics*, 35(24):5318, 2019.

[2] Jacob Pritt, Nae-Chyun Chen, and Ben Langmead. FORGe: prioritizing variants for graph genomes. *Genome biology*, 19(1):1–16, 2018.

[3] Ivar Grytten, Knut D Rand, Alexander J Nederbragt, and Geir K Sandve. Assessing graph-based read mappers against a baseline approach highlights strengths and weaknesses of current methods. *BMC genomics*, 21:1–9, 2020.

[4] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M Stütz, Vladimir Benes, and Jan O Korbel. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28 (18):i333–i339, 2012.

Supplementary Material
Chapter 4

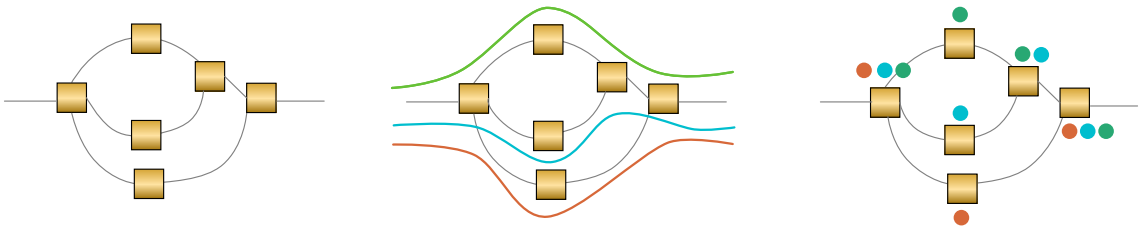


Figure S4.1: **Labelling of the nodes in the multi-assembly graph.**
To determine the support for the nodes in the graph, we aligned each individual assembly back to the multi-assembly graph and labeled nodes according to the assembly paths that traversed them with different colors. The left panel represents a schematic graph. Rectangles and lines represent nodes and edges, respectively. The middle panel represents the paths of three assemblies traversing the nodes. The right panel displays how each node that was traversed by an assembly receives a label (colored dots).

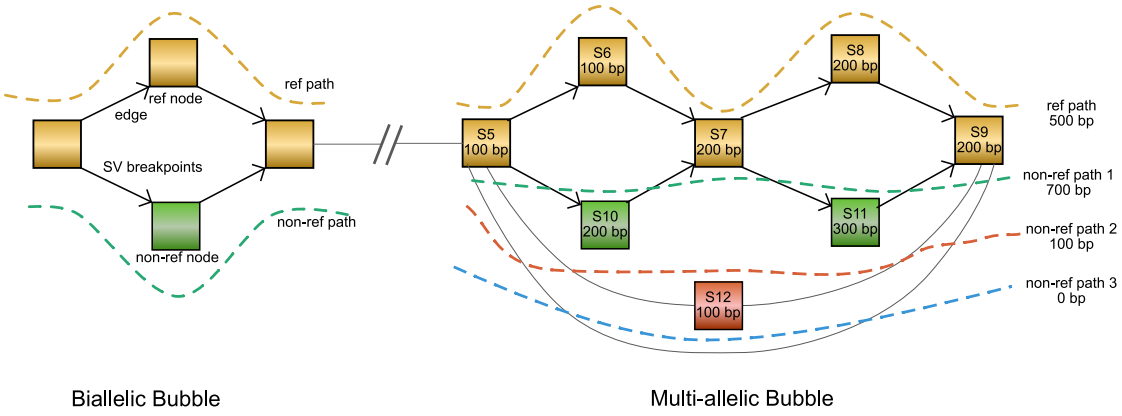


Figure S4.2: **Graphs and structural variants terminology used in the paper.**
(left) A node contains a sequence of nucleotides (S1-S12). Reference nodes (S1, S2, S4) are derived from the backbone assembly used to construct the graph. Non-reference nodes (S3) contain sequences from additional assemblies that are not present in the backbone. Nodes are connected by directed edges from parent to child where the underlying sequences are contiguous. Edges between reference and non-reference nodes are breakpoints of structural variations. Bubbles are branching regions in the graph which start and end at reference nodes. *(right)* Paths in the bubbles represent different alleles of structural variations, which are biallelic if a bubble contains two paths or multiallelic if it contains more. Nodes within biallelic bubbles represent alleles. Within multi-allelic bubbles, multiple nodes may be part of the same path and thus allele. It is worth noting that not all combinations of nodes within bubbles are real paths found in individual assemblies (e.g., S10-S7-S8). As such, color-consistent nodes within a bubble are stitched together to represent true paths. By comparing reference and non-reference paths, it is possible to determine the type of the structural variations (e.g., non-ref path 1: alternate insertion, path 2: alternate deletion, path 3: complete deletion).

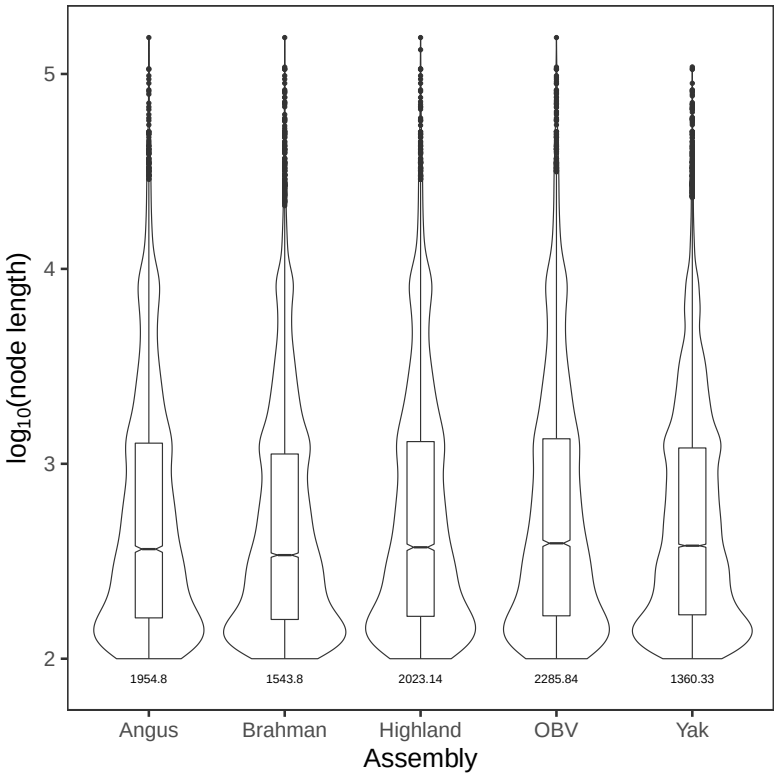


Figure S4.3: **The size of non-reference nodes labelled with each of the five assemblies.** The Y-axis is log10-scaled. Numbers below each plot refer to the average non-reference node length from each assembly.

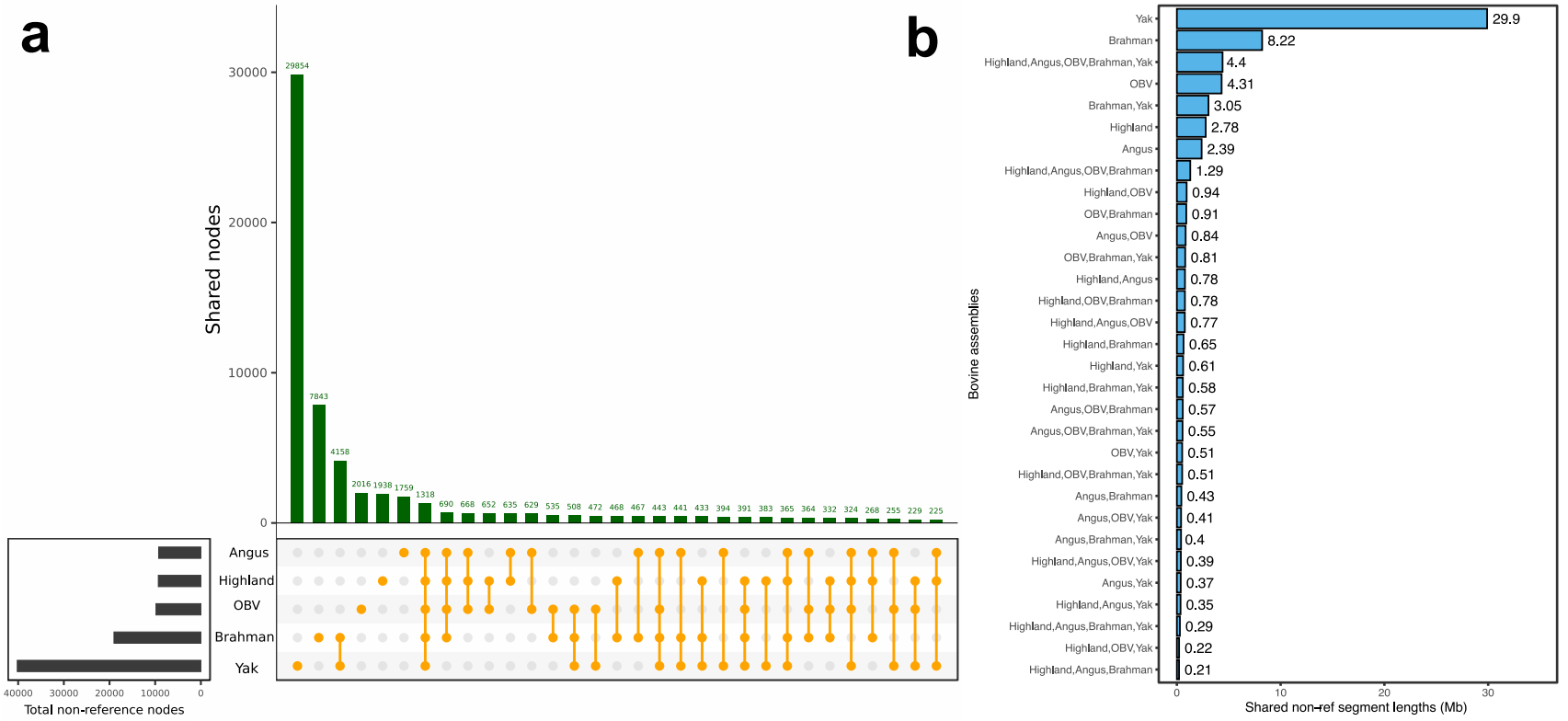


Figure S4.4: **Non-reference nodes detected across assemblies.**
Intersection of non-reference nodes (a) and cumulative length of non-reference sequences (b) found in five assemblies when compared to ARS-UCD1.2. OBV = Original Braunvieh.

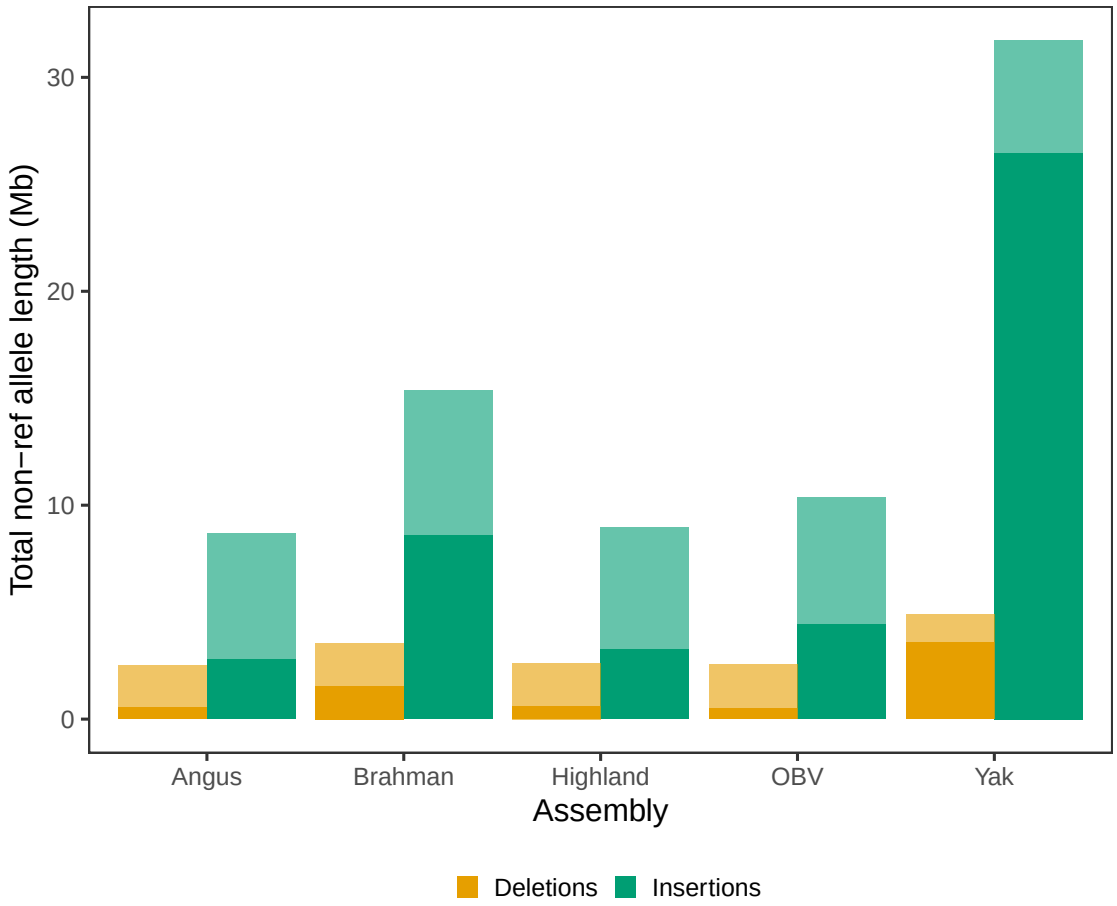


Figure S4.5: **Deletion and insertion polymorphism detected from each assembly in the pangenome graph.**
Transparent and solid bars indicate the total and private length of non-reference alleles respectively. OBV – Original Braunvieh.

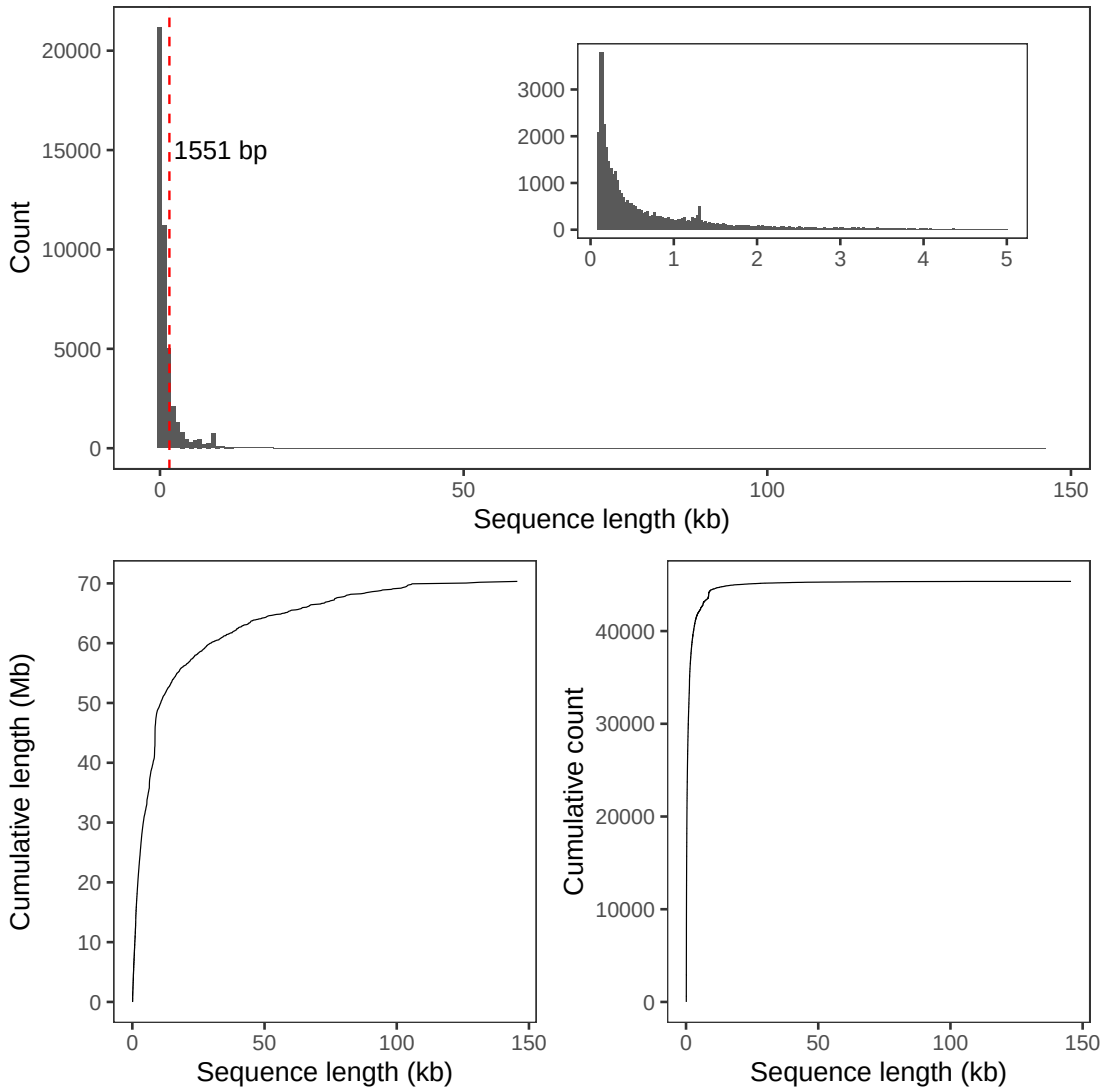


Figure S4.6: **Length of the non-reference sequences that were added linearly to the ARS-UCD1.2 reference.**
Length distribution of the non-reference alleles (upper panel) and their cumulative length and count (lower panels). The inset in the upper panel displays the distribution of non-reference alleles shorter than 5 kb. The dashed-red line indicates the average length (1551 bp) of the non-reference alleles.

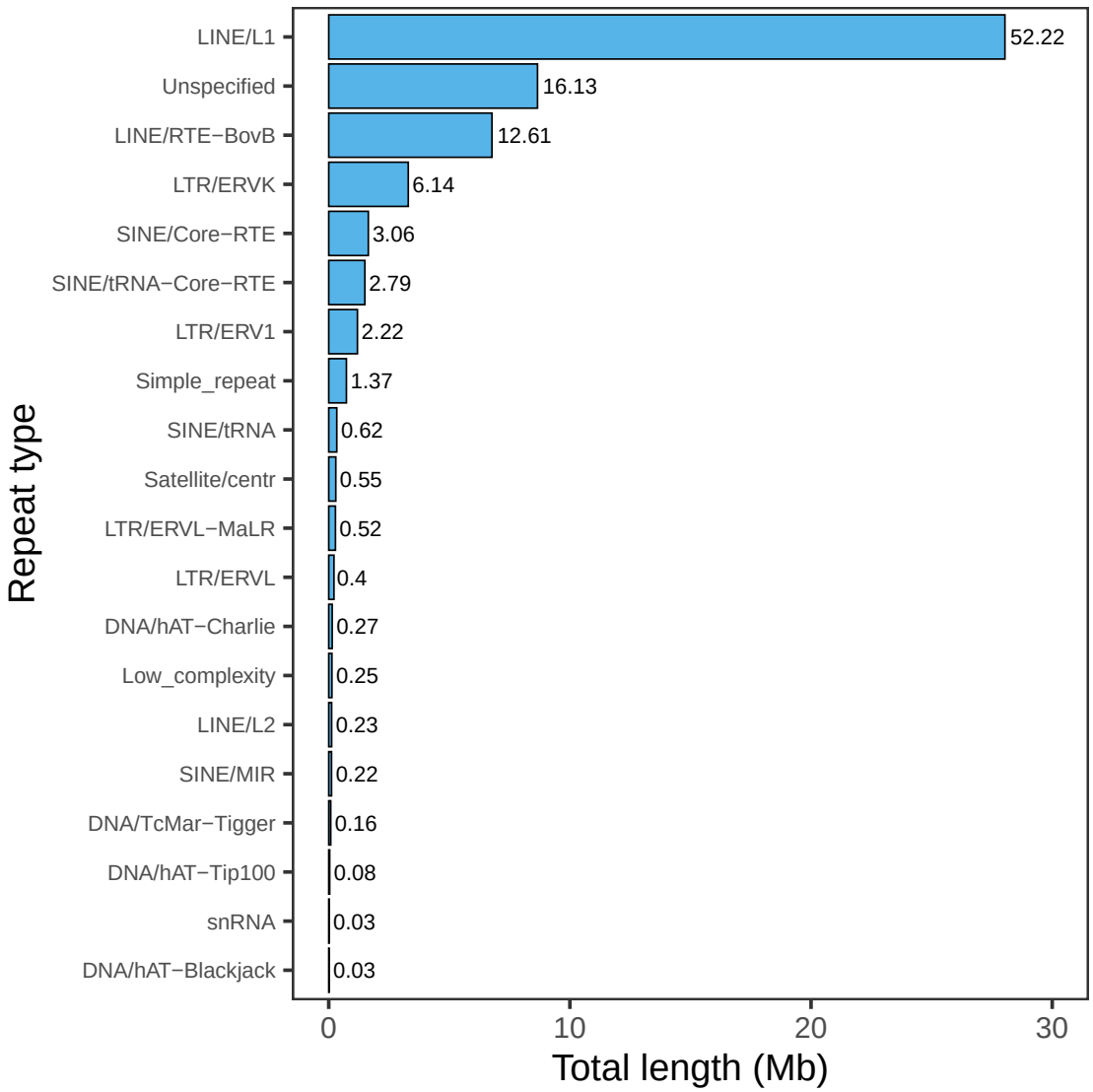


Figure S4.7: **Prevalence of repetitive elements in the non-reference sequences.** The 20 most prevalent repetitive elements account for 99.9% of the repetitive elements detected in the non-reference sequences. The X-axis indicates the summed sequence length (in Mb) spanned by the repetitive elements, with text labels indicate the proportion (%) of a repetitive element contributing to the total repeat length.

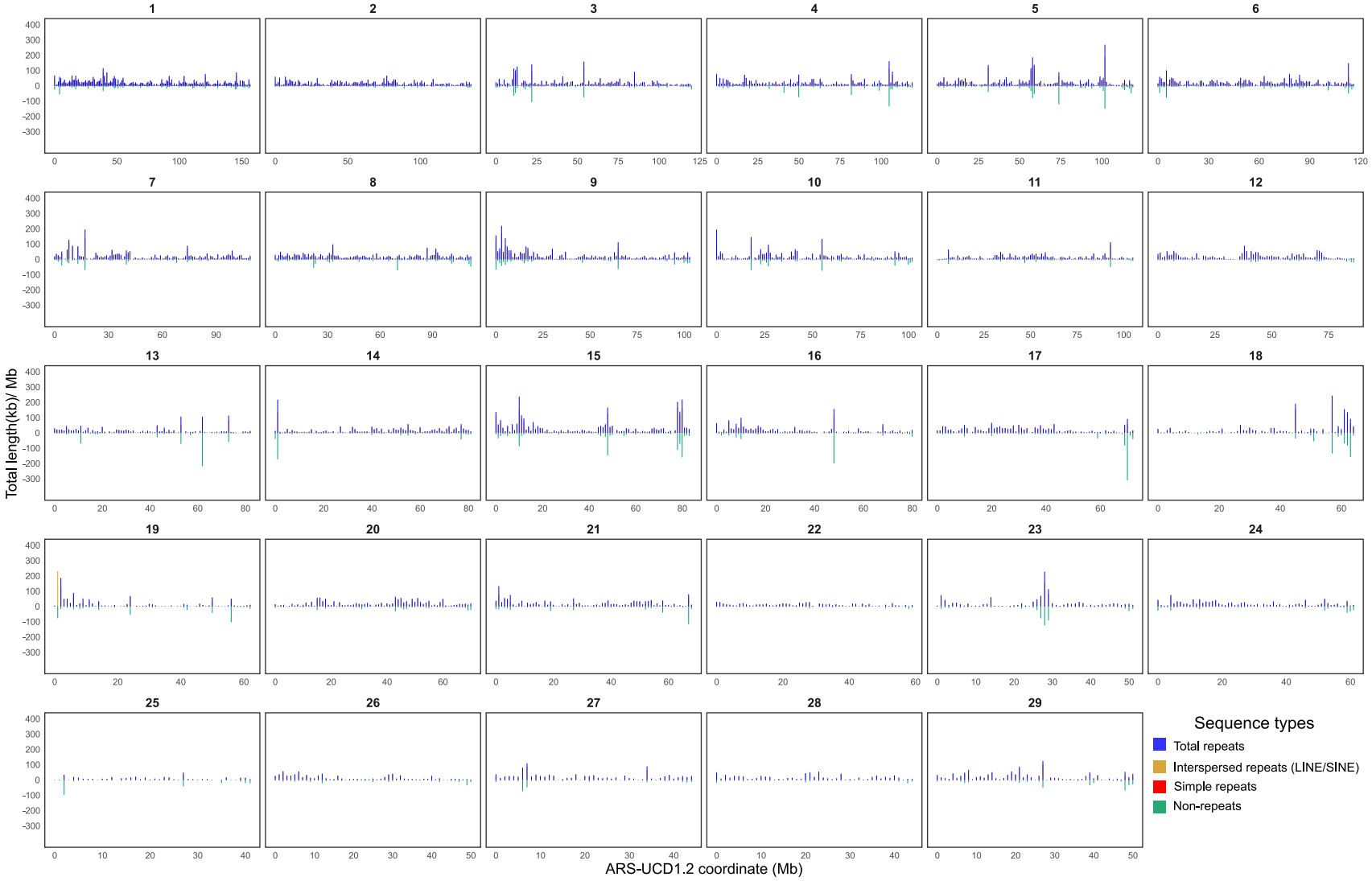


Figure S4.8: **The distribution of repetitive element (interspersed and simple repeats), and non-repetitive elements** found in non-reference sequences based on the ARS-UCD1.2 coordinate system. To aid visualization, the distribution of non-repetitive segments is mirrored to the negative Y-axis. The numbers above the individual panels are chromosome identifiers.

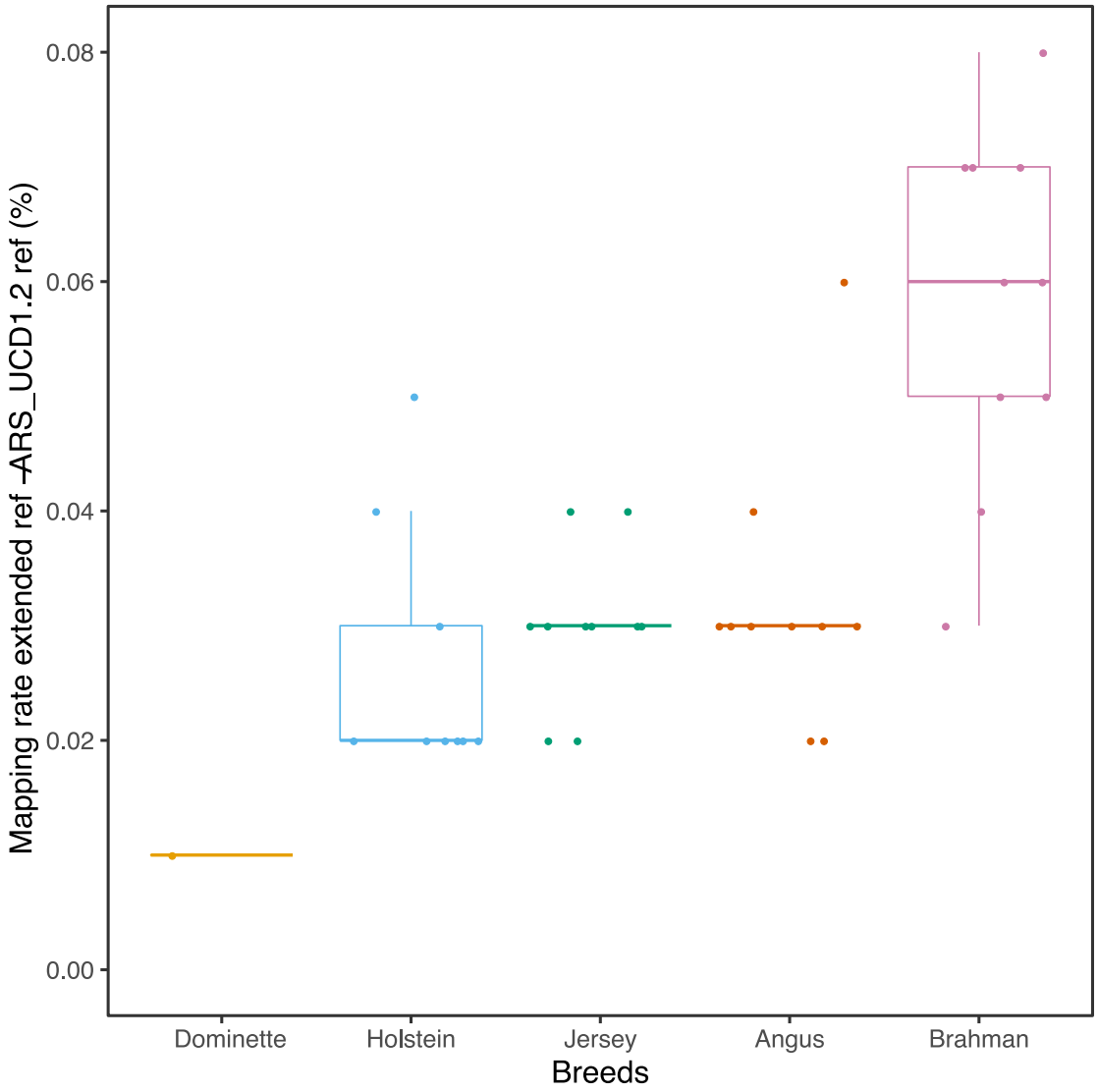


Figure S4.9: **Transcriptome mapping rate improvements in five breeds using the extended reference sequence over ARS-UCD1.2.**
Values along the Y axis represent the difference in mapping rate between the extended and the original ARS-UCD1.2 reference (%) as reported by HISAT2. Positive values indicate that more reads aligned to the extended than original reference. Dominette is the Hereford animal used to construct ARS-UCD1.2.

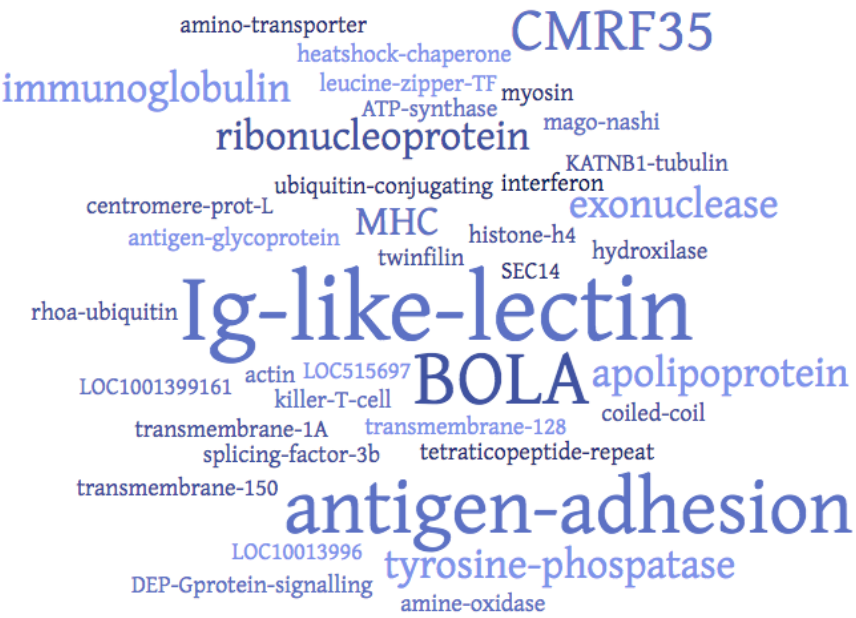


Figure S4.10: **Word cloud of the top blast hits from 142 putatively novel genes assembled from RNA sequencing reads mapping to non-reference sequences.**
The BLAST query was performed against a protein database containing sequences from *Bos* and related species. Word size reflects the frequency of the hits.

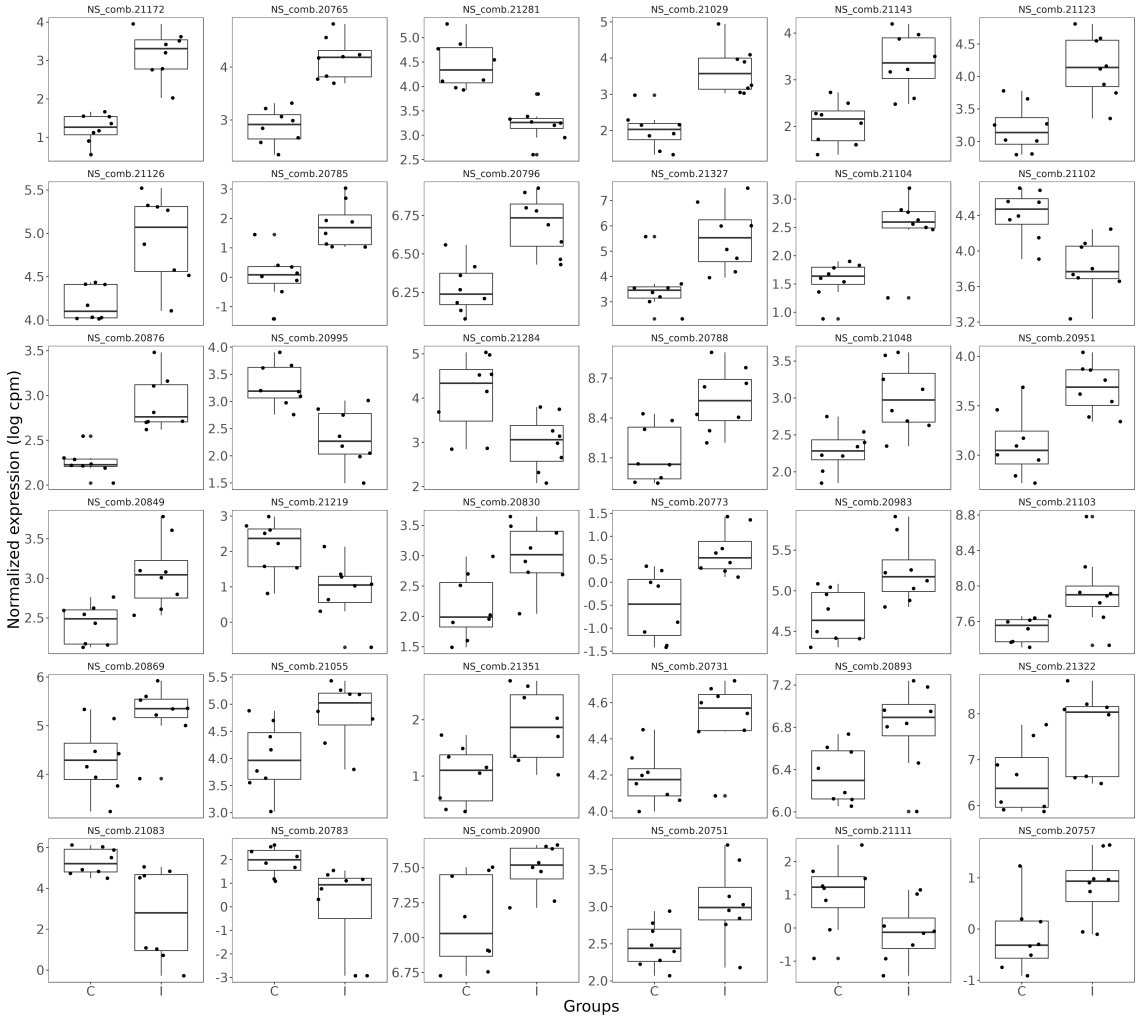


Figure S4.11: **Differential expression of 36 non-reference genes in Mycobacterium bovis-infected cattle.**
Control (C) and *Mycobacterium bovis*-infected (I) cattle are grouped separately for each gene. Y axis indicates the normalized transcript abundance expressed as $\log_2$ CPM as reported by EdgeR.

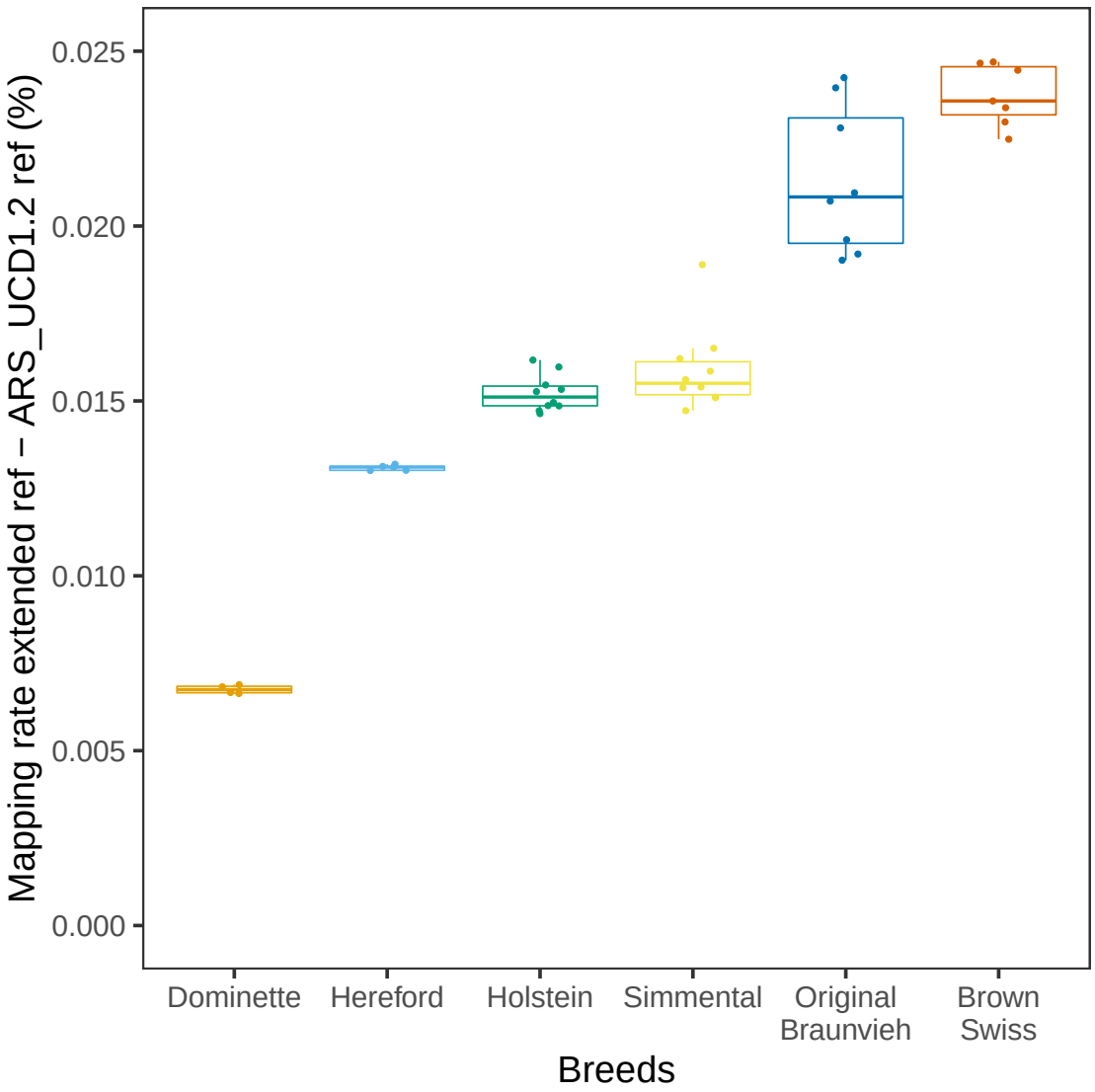


Figure S4.12: **Mapping rate of whole-genome short sequencing reads to the extended linear reference genome.**
The Y-axis reflects the difference (in %) in mapping rate between the extended reference and the original ARS–UCD1.2 reference sequences. Positive values indicate that the mapping rate is higher for samples aligned to the extended than original ARS-UCD1.2 reference sequences. Short sequencing reads of 45 cattle from five breeds were considered. Dominette is a Hereford cattle, but is separated as she is the animal used to construct ARS-UCD1.2.

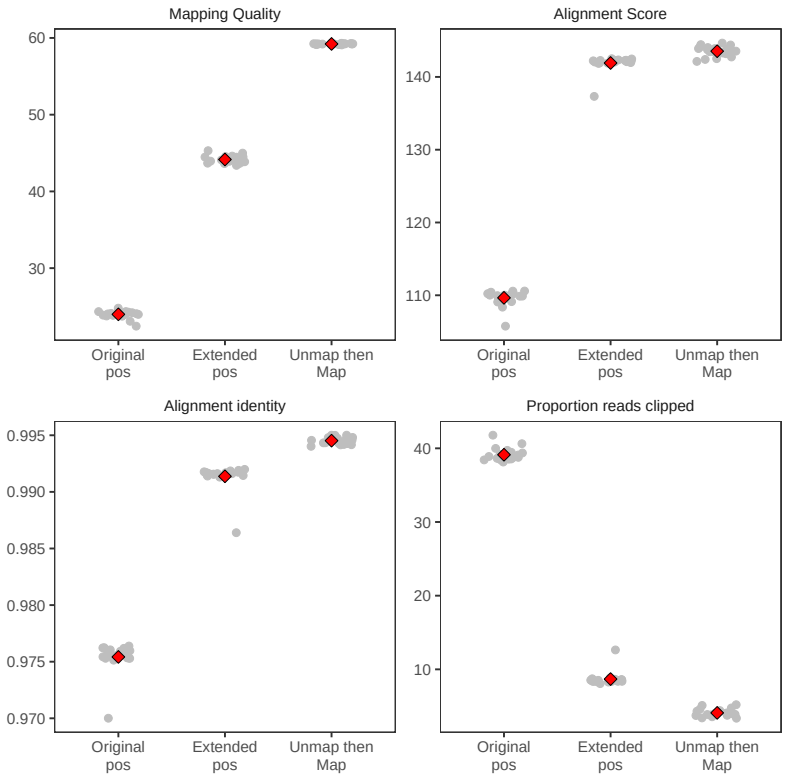


Figure S4.13: **Accuracy of read mapping to non-reference sequences.**
Four mapping statistics (mapping quality, alignment score, alignment identity, proportion of clipped reads) were assessed for short sequencing reads from 45 samples across 5 breeds. First, we consider reads that mapped to autosomal sequences of the ARS-UCD1.2. The mapping statistics of these reads are compared between the ARS-UCD1.2 reference sequence (Original pos) and their mapping position at the novel non-reference sequences of the extended reference genome (Extended pos). Second, we consider reads that were unmapped against the ARS-UCD1.2 reference genome but received a mapping position against the extended reference genome (Unmap then map). Each grey point indicates the average mapping statistics for one DNA sample and red diamond indicates the average across all animals.

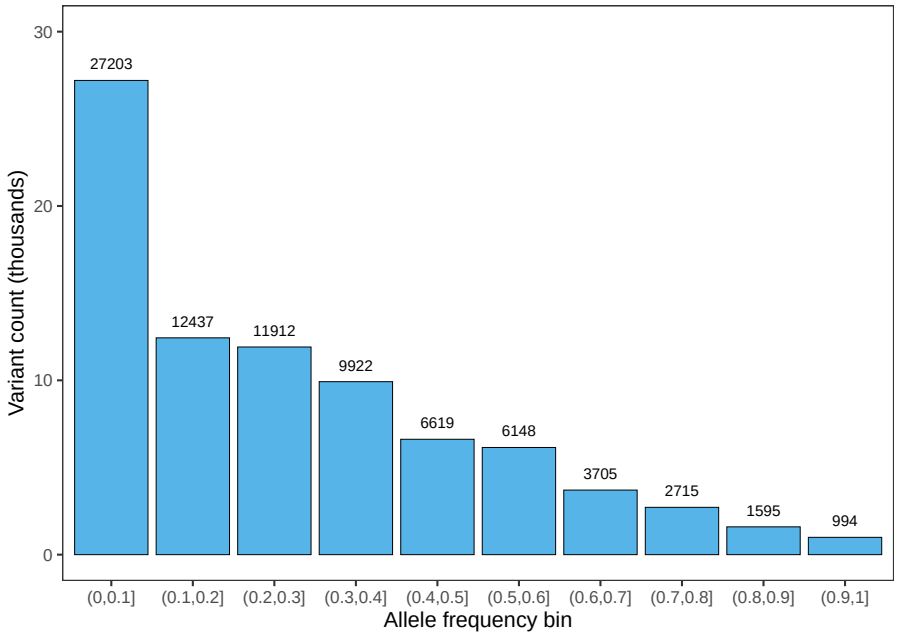


Figure S4.14: **Alternate allele frequency of 83,250 variants detected from non-reference sequences in 45 samples from 5 breeds.**

Table S4.1: **Different types of structural variations discovered from the multi-assembly graph.**
Variant length is calculated based on the absolute difference between reference and non-reference allele.

Mutations	Types	Count	Complete type	Alternate type	Non-ref allele length	Variant length
Insertions	biallelic	35748	20432	15316	40361474	37388222
Insertions	multiallelic	4621	4221	400	21116534	10303720
Deletions	biallelic	28476	15377	13099	2845080	28373582
Deletions	multiallelic	4661	1972	2689	10130841	11727721
Total		*73506*	*42002*	*31504*	*74453929*	*87793245*

Table S4.2: **Gene model prediction from repeat masked non-reference sequences.**
Total novel genes denote all gene models (including partial genes) predicted by Augustus. Complete gene models restricted to only full gene models (TSS, start codon, exon, intron, stop-codon present). Transcript, exon, and CDS statistics reported as mean (maximum-minimum) length from the full gene models.

Feature of the gene model	**Value**
Total novel genes (distinct SVs)	857 (768)
Complete novel genes (distinct SVs)	374 (328)
Transcript length (bp)	4742.14 (min: 314; max: 104024)
Exon length (bp)	942.30 (min: 15; max: 6725)
Exon length/gene (bp)	2050.89 (min: 314; max: 7762)
Exon count/gene (bp)	2.18 (min: 1; max: 20)
CDS length (bp)	396.64 (min: 5; max: 3059)
CDS length/gene (bp)	794.34 (min: 199; max: 6280)
protein length (aa)	264.78 (min: 66.33; max: 2093.33)

Table S4.3: **BLASTX hits of the transcripts from differentially expressed non-reference genes**
Log_2 FC is the difference in expression between Mycobacterium bovis-infected and non-infected control cattle (e.g., a positive value indicates that expression is higher in infected than control cattle), and Adj FDR is the adjusted false discovery rate determined using the Benjamini-Hochberg correction.

Hits	Mean (SD) expression in CPM		log_2 FC	Adj FDR
	Control	Infected		
Antigen WC1.1-like	2.43 (0.6)	9.54 (3.65)	2.0137	1.98E-05
Leukocyte immunoglobulin-like receptor subfamily A member 5 isoform X1	23.10 (8.30)	9.59 (2.54)	-1.2870	0.0001
PREDICTED: synaptobrevin homolog YKT6	4.39 (1.36)	11.19 (4.7)	1.3754	0.0008
PREDICTED: major vault protein isoform X1	21.70 (3.87)	14.23 (2.88)	-0.6140	0.0040
PREDICTED: heat shock 70 kDa protein 1B	10.18 (2.7)	5.33 (1.94)	-0.9511	0.0041
Elongation factor 1-alpha 1	282.86 (40.74)	374.22 (63.08)	0.4033	0.0093
Heterogeneous nuclear ribonucleoprotein R isoform 2	5.47 (0.94)	8.65 (2.73)	0.6740	0.0148
PREDICTED: prothymosin alpha isoform X2	22.02 (10.73	39.85 (12.84	0.8523	0.0243
Stathmin isoform a	17.56 (7.39)	30.3 (10.17)	0.7823	0.0271
Serine/arginine repetitive matrix protein 1 isoform X1	18.3 (1.74)	22.99 (3.29)	0.3293	0.0285
BOLA class I histocompatibility antigen, alpha chain BL3-7-like	109.32 (61.3)	223.94 (116.59)	1.0387	0.0302
Predicted gene, EG665562	141.66 (31.9)	179.9 (20.92)	0.3440	0.0400
PREDICTED: GTP-binding protein SAR1a	5.71 (1.22)	8.71 (3.28)	0.6231	0.0415

Table S4.4: **Comparison of read mapping accuracy between the extended and ARS-UCD1.2 reference.**
All metrics were extracted from BAM files using pysam v0.16.0.1 https://github.com/pysam-developers/pysam. Alignment identity reflects the proportion of bases from an aligned read that match the reference sequence. A read was considered to be clipped if the CIGAR string of the alignment contains tags for either hard- (H) or soft-clipped (S) bases. Supplementary alignments were reported for alignments with an XS tag. Criteria for perfect and unique alignments were based on those reported by Crysnanto and Pausch [1]. Specifically, reads were considered to align perfectly if the edit distance was zero along the entire read (NM:0 tag), and when the CIGAR did not include H or S tags. Unique alignments are reported for reads that either have a single primary alignment or reads that have a secondary alignment (XA tag) but one alignment has a maximum mapping quality score of 60. Reported values are averaged over n=45 samples. Paired one-sided t-tests were conducted with n-1 degrees of freedom. Parameters marked with '*' indicate the null-hypothesis that ARS-UCD1.2 would perform better than the extended reference, while those without marks indicate the reverse. All tests rejected the null hypothesis.

Parameter	**Extended reference**	**ARS-UCD1.2**	**Difference**	**Stdev**	**t-statistic & p-value**
Unmap (%) *	0.4291	0.4467	-0.0176	0.00461087	t =-24.12, p = 2.39e-25
Alignment identity 99% (%)	87.2716	87.1875	0.0841	0.00433417	t = 122.72, p = 2.19e-52
Alignment perfect (%)	68.5732	68.4687	0.1045	0.00667272	t = 99.04, p = 9.10e-49
Clipped alignment (%) *	2.1335	2.1923	-0.0588	0.00891613	t = -41.74, p = 2.81e-34
Supplementary alignment (%) *	0.2078	0.2219	-0.0141	0.00379671	t = -23.45, p = 6.69e-25
Unique alignment (%) *	83.2919	83.6016	-0.3017	0.03348539	t = -58.51, p = 6.50e-40

Table S4.5: **Functional consequences predicted for 83,250 non-reference variants.** Variant consequences were predicted using VEP (version 91.3) based on a custom annotation file from Augustus. Only the most severe consequence is shown for each variant.

Variant consequence	**SNPs**	**Indels**	**All**	**Proportion (%)**
splice_acceptor_variant	4	1	5	0.006
splice_donor_variant	2	0	2	0.0024
frameshift_variant	0	26	26	0.0312
inframe_insertion	0	1	1	0.0012
inframe_deletion	0	4	4	0.0048
splice_donor_variant	2	0	2	0.0024
stop_gained	17	0	17	0.0204
stop_lost	1	0	1	0.0012
start_lost	3	0	3	0.0036
missense_variant	700	0	700	0.8408
splice_region_variant	45	1	46	0.0553
synonymous_variant	374	0	374	0.4492
stop_retained_variant	1	0	1	0.0012
coding_sequence_variant	2	0	2	0.0024
5_prime_UTR_variant	86	2	88	0.1057
3_prime_UTR_variant	1253	149	1402	1.6841
intron_variant	5809	443	6252	7.5099
upstream_gene_variant	2559	277	2836	3.4066
downstream_gene_variant	1811	179	1990	2.3904
intergenic_variant	61040	8458	69498	83.481
moderate impact	701	5	706	0.848
high impact	27	27	54	0.0649

Note S4.1

Assembly of the Original Braunvieh (OBV) genome

The Original Braunvieh primary assembly was generated from PacBio HiFi CCS reads (study accession PRJEB42335 under sample accession SAMEA7759028), generated from subreads with minimum three passes and minimum predicted read quality of 20. The fastq data contained 86.9 gigabases, corresponding to nearly 30-fold coverage. The CCS reads were filtered by fastp [0.21.0] [2] with minimum average quality of Q20 and minimum read length of 1kb, with 99.99% of the data passing these thresholds. Hifiasm [v0.13-r308] [3] was then used to generate the assembly from the reads using the additional parameters "-r 4 -a 5 -n 5" on a computing cluster. Hifiasm yields the primary contigs in the GFA format, which were then converted using gfatools [0.4-r196-dirty] into a fasta sequence representation. These contigs were then scaffolded using RagTag [v1.0.1] [4] to the ARS-UCD1.2 reference, with custom parameters "–mm2-params "-c -x asm5" -r -m 1000000".

The contigs were validated for contiguity, completeness, and correctness by multiple independent tools, available in a Snakemake [5.26.1] [5] pipeline online at https://github.com/AnimalGenomicsETH/bovine-assembly. Basic contiguity was determined through the asmstat command of paftools [6]. Similarly, the NGA50 value was determined through mapping the contigs to the ARS-UCD1.2 reference, and subsequently considering the length of alignment blocks again with asmstat. These values are described in Table SN41.

Table SN41: Contiguity metrics of the primary Original Braunvieh assembly.
Size refers to the total number of bases in the chromosomes and unplaced contigs. NG50 was calculated for both the contig set and the scaffolded assembly with the expected genome size taken from the ARS-UCD1.2 reference. Similarly, NGA50 is the NG50 value for aligned blocks of the assembly to the ARS-UCD1.2 reference.

	Size	Contig NG50	NGA50	Scaffold NG50	L50	Contigs
assembly	3.17gb	86.0	68.9	96.3	15	765

Completeness of the assembly was determined through two independent approaches, BUSCO (8) and the asmgene command of paftools. The former relies on the OrthoDB datasets, specifically version 10 of the cetartiodactyla lineage. The latter uses cDNA libraries of annotated gene sequences from the ARS-UCD1.2 reference available from Ensembl. Both methods report a high completeness (>96%) with respect to predicted gene content, as shown in Table SN42.

Table SN42: Predicted single-copy gene completeness of the primary Original Braunvieh assembly.
Single-copy refers to genes that were correctly present once in assembly, while duplicates are genes which appeared more than expected. Fragmented genes are those which are only partially mapped, or fully mapped but split into multiple pieces. Missing genes are either not found or mapped below 10% of the expected gene.

	Single copy	duplicates	fragmented	missing	total
Busco	12533	283	166	353	13335
asmgene	18503	166	68	136	18873

Correctness was likewise determined by two k-mer based approaches, yak [r58] [3] and Merqury [7]. Yak uses an approximate hash-table approach, while Merqury can be run in an exact mode. Both used short read sequences (2x150 bp) from the primary animal, which importantly were not used in generating the assembly, allowing for an independent evaluation. In addition, short read sequences from both parents enabled a quantification of the switch error rate. Only yak provided an estimate of the Hamming error rate, while only Merqury provides phased block statistics. An overview of these statistics is shown in Table SN43.

Table SN43: K-mer based, reference-free validation of the primary Original Braunvieh assembly.
Assembly quality value (QV) is given as a Phred quality score. Completeness estimates how many k-mers present in the short reads are found in the assembly contigs. The switch error rate is calculated differently by yak and Merqury, measuring the percent of wrongly phased adjacent SNPs in yak while in Merqury it measures the percent of wrongly phased haplotype-specific k-mers ("hap-mers"). There are more than 100 phase switches within a 20kb window (long-range switch). The Hamming error is the percent of SNP sites that are phased wrongly. The phase block statistic is the N50 after contigs have been broken at long-range switches, defined as more than 100 wrongly phased hap-mers per 20kb window.

	QV	Completeness	Switch error	Hamming error	Phased N50
Yak	48.76	100	0.012	0.37	-
Merqury	50.85	93.46	0.08	-	2.5 mb

Furthermore, the assembly was validated by comparing structural variants called by pbsv [2.4.0] between the reads and the ARS-UCD1.2 reference and those called by mumandco [v2.4.2] [8] between the assembly and the reference. There was good concordance between these approaches, for example an 8kb inversion identified in chromosome six of the assembly matched an 8kb inversion predicted by the read mapping.

The repeat content of the assembly was also in line with expectations, with approximately 48% of the assembly consisting of repeat elements or low complexity regions according to RepeatMasker version 4.1.1 [9] using the Repbase repeat database (release 20181026) [10]. Several bovine-specific repeats were identified, along with telomeric or centromeric sequences not present in the existing ARS-UCD1.2 reference, indicating that several contigs are approaching chromosomal-scale and completeness.

Note S4.2

Determination of the core and flexible parts of the pangenome graph

To investigate if the order of assemblies used to establish the multi-assembly graph impacts the core and flexible parts, we added the assemblies randomly to the graph. Core genome represents bases shared across all assemblies in the graph, while flexible genome represents number of bases that are variable across assemblies (i.e., not found in all assemblies) [11]. The pangenome increased gradually with the number of assemblies added, driven by an increase in the flexible genome. The core genome size decreased from 2480 Mb to 2400 Mb in the full graph (Figure SN41), indicating that more genomic segments are variable across bovine species as we add more assemblies into the graph.

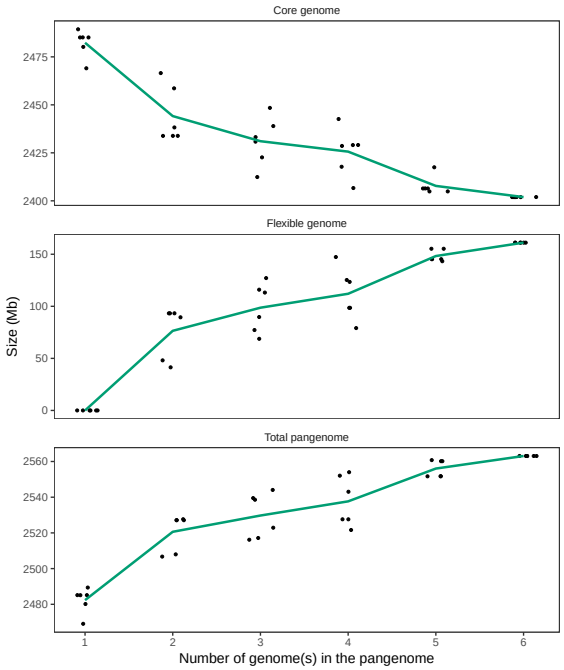


Figure SN41: Profile of the multi-assembly graph with an increasing number of genomes integrated into the graph.
We varied the order and number of genomes added to the graphs, and calculated the number of bases in the pangenome, number of bases that are shared across all assemblies in the graph (core genome), and the number of bases that are variable across assemblies (i.e., not found in all assemblies, flexible genome). Points and lines indicate individual and average values.

Next, we investigated the profile of a multi-assembly graph that gradually increases in complexity. We built taurine-only graphs that contained either all or all but one taurine assemblies, a TauInd (four taurine and one indicine), and a full graph (four taurine, one indicine, and yak). The profile of the pangenome changed markedly as more distant assemblies were added to the graph. For example, the flexible part declined substantially from 6.10% in the full graph to 3.83% and 2.76% for the TauInd and taurine-only graph, respectively. However, when an individual taurine assembly is

removed from the taurine-only graphs, the size of the flexible part changes only slightly Figure SN42).

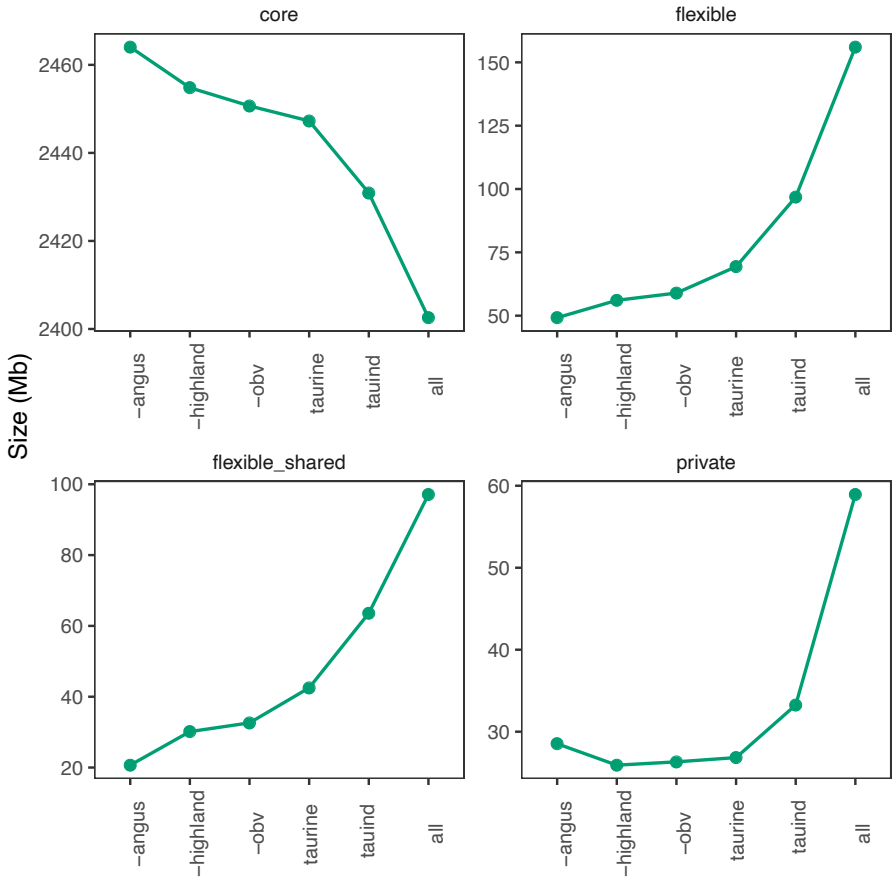


Figure SN42: Pangenome profile as more distant assemblies are added to the graph. The X-axis indicates the constructed graphs (- denotes the taurine assembly that was removed from the taurine-only graph, taurine denotes a graph with four taurine assemblies, the indicine graph contains all taurine assemblies and the assembly of Brahman, and all reflects a multi-assembly graph that contains all six assemblies (taurine, indicine, and yak). The core part is the size of the segments that are common to all assemblies, flexible_shared indicates the size of segments shared by at least two but not all assemblies, and private denotes the size of segments found only in a single assembly, thus flexible genome is composed of flexible_shared + private segment.

Note S4.3

Construction of bovine multi-assembly graphs with different backbones

To investigate if the choice of the backbone assembly influences the properties of the bovine multi-assembly graph, we constructed six graphs, one for each possible assembly backbone. The remaining five assemblies were added according to their Mash-distance to the chosen backbone. Larger assembly backbones tended to result in larger multi-assembly graphs (see Table 4.1 in the main paper, Table SN44). The total number of non-reference bases detected varied between 63,745,420 bp and 72,349,303 bp for the OBV and Brahman backbone, with a mean value of 68.72 Mb and a standard deviation of 3.17 Mb. Fewer non-reference bases were detected when the OBV and Highland assemblies were used as backbones.

Table SN44: Properties of bovine multi-assembly graphs with different backbones

Parameter	Unit	Backbone Assembly					
		Hereford	Angus	Highland	OBV	Brahman	Yak
Nodes	n	182,940	182,332	183,118	184,098	184,301	188,975
Size	bp	2,558,596,439	2,540,507,180	2,550,176,720	2,671,491,862	2,549,613,449	2,547,048,782
Ref nodes	n	123,483	123,116	124,220	125,088	124,410	126,883
Ref length	bp	2,489,385,779	2,468,157,877	2,483,452,092	2,607,746,442	2,478,073,158	2,478,308,164
Nonref nodes	n	59,457	59,216	58,898	59,010	59,891	62,092
Non-ref length	bp	69,210,660	72,349,303	66,724,628	63,745,420	71,540,291	68,740,618
Edges	n	258396	257531	258608	260044	260209	266139
Edges/nodes	ratio	1.4125	1.4124	1.4122	1.4125	1.4119	1.4083
R-R edges	n	141,086	140,742	142,133	143,133	141,978	144,442
R-NR edges	n	113,332	112,669	113,116	114,058	114,064	114,837
NR-NR edges	n	3,978	4,120	3,359	2,853	4,167	6,860
core count	n	67,482	67,499	67,616	67,619	67,763	68,614
core length	bp	2,402,561,410	2,394,756,562	2,402,656,874	2,414,762,810	2,398,150,572	2,397,494,177
core prop	%	93.9	94.26	94.22	90.39	94.06	94.13
flexible count	n	115,458	114,833	115,502	116,479	116,538	120,361
flexible length	bp	156,035,029	145,750,618	147,519,846	256,729,052	151,462,877	149,554,605
flexible prop	%	6.10	5.74	5.78	9.61	5.94	5.87
CPU time	min	290.43	276.33	274.52	210.46	282.59	299.01
Max mem	Gb	55.03	58.88	55.6	58.31	56.67	56.96
Average mem	Gb	36.34	37.08	34.43	34.38	36.08	34.64
Run Time	min	41.78	39.4	41.23	30.6	39.35	42.7

The choice of the backbone had, as expected, a major impact on the amount of non-reference bases detected from each of the remaining assemblies (Table SN45). A multi-assembly graph with a *Bos taurus taurus* backbone contains between 10.14 and 19.48 million non-reference bases from the remaining three taurine assemblies. Using the Hereford or Angus assemblies as the backbone resulted in more total non-reference sequences than using the OBV or Highland assemblies. Regardless of backbone choice, the OBV and Highland assemblies also contribute more non-reference sequences to the multi-assembly graph than the Hereford and Angus assembly. These two observations suggest that the OBV and Highland assemblies represent a more comprehensive *Bos taurus taurus* genome, agreeing well with their high completeness, continuity, and correctness (see Note S4.1 and [12]). Although selecting a more distant assembly as the backbone identifies more non-reference sequences from each remaining assembly on

average (~40 Mb with Yak compared to ~20 Mb for taurine), this appeared to be a smaller effect compared to the backbone completeness.

Table SN45: : Non-reference bases detected (Mb)

Backbone	Assembly						
	Hereford	Yak	Brahman	OBV	Angus	Highland	Total1
Hereford	-	43.34	23.64	18.2	14.45	15.54	69.21
Yak	40.88	-	40.12	44	38.94	39.48	68.74
Brahman	23.09	42.11	-	24.62	21.23	21.7	71.54
OBV	11.91	39.64	17.69	-	10.14	11.18	63.75
Angus	17.99	44.02	23.40	19.48	-	16.94	72.35
Highland	13.03	40.09	20.27	14.21	11.64	-	66.72

Furthermore, the total length of each assembly's private non-reference nodes, was barely affected by backbone choice. This suggests that our approach to building multi-assembly graphs with minigraph and labelling non-reference nodes work well regardless of choice of the initial backbone assembly (Table SN46).

Table SN46: Total length in Mb of private non-reference nodes.

Backbone	Assembly					
	Hereford	Yak	Brahman	OBV	Angus	Highland
Hereford	-	29.9	8.22	4.61	2.39	2.78
Yak	4.36	-	8.62	5.05	2.47	2.93
Brahman	4.69	30.26	-	4.85	2.69	3.19
OBV	4.33	30.02	8.34	-	2.67	3.02
Angus	4.60	29.85	8.20	5.01	-	2.76
Highland	4.40	29.85	8.14	4.75	2.38	-

Assessment of the sequences not included in the graphs

Minigraph might fail to align and include input sequences into the graph. We assessed bases not included in the graph by comparing the total realignment size of the assembly to the graphs with the total pangenome size. All bases in the backbone are included in the multi-assembly graph, but this is not the case for the additional assemblies which subsequently augment the graph (Table SN47). We again found that the use of the Original Braunvieh backbone led to fewer non-reference bases not included in the graph from each remaining assembly.

Table SN47: Total assembly sequences (bp) not included in the graphs.

Backbone	Assembly						Total
	Hereford	Yak	Brahman	OBV	Angus	Highland	
Hereford	-	5,388,777	6,369,649	106,905,680	7,442,463	8,165,867	134,272,436
Yak	15,420,157	-	9,466,221	116,311,772	8,286,464	10,732,937	160,217,551
Brahman	17,427,243	9,103,978	-	111,856,939	12,067,771	12,108,523	162,564,454
OBV	9,716,932	2,690,446	3,339,031	-	882,571	4,659,329	21,288,309
Angus	26,973,081	18,470,665	20,032,971	115,587,224	-	20,796,186	201,860,127
Highland	13,401,923	6,069,608	7,053,961	104,868,806	6,680,628	-	138,074,926

The Original Braunvieh assembly has between 105 and 116 Mb of sequences which are not augmented into the different multi-assembly graphs. We investigated which parts of the OBV assembly were not included in the Hereford-backbone graph using the reverse mapping approach enumerated below:

1. Extract nodes in the graph covered with OBV alignment
2. Map the sequence in the node to the OBV assembly using minimap2
3. Collect region longer than 10kb with no coverage from the alignment
4. Visualize the region across OBV genome region

As shown in Figure SN43, many sequences not included in the graph are located at the start or end of chromosomes, which might indicate that the HiFi reads enabled a better (more complete) assembly of telomeric or centromeric regions. This hypothesis is further supported by a repeat masker analysis revealing that these regions contain many DNA satellite (21,139,818 bp) and retroelements (7,774,559 bp).

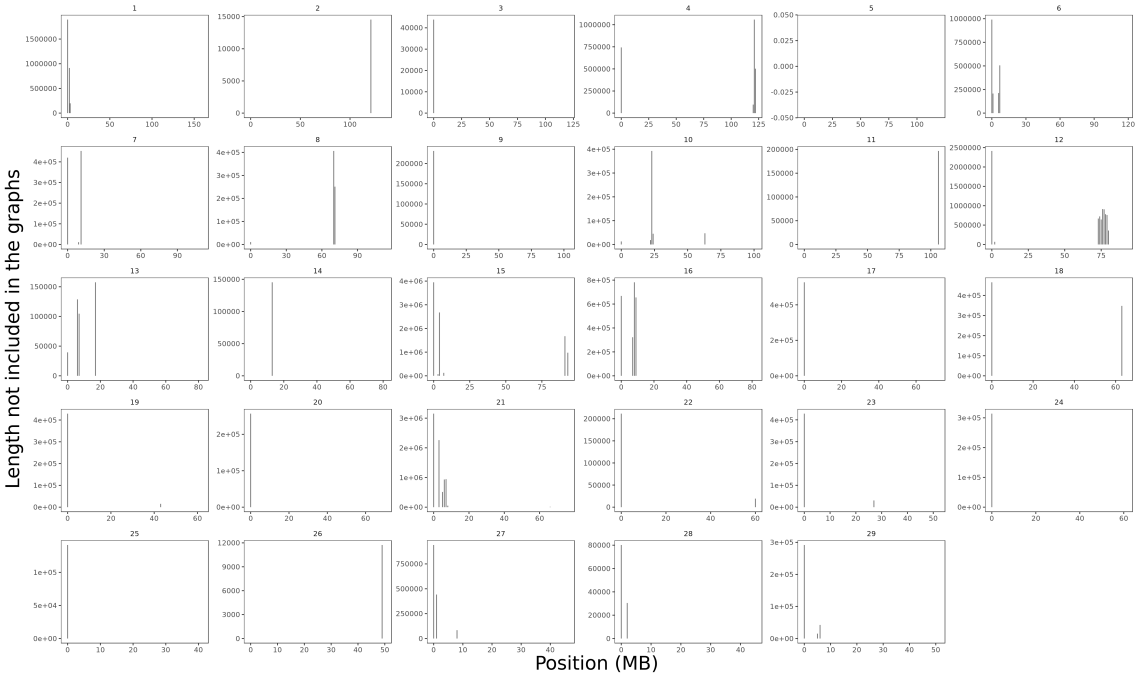


Figure SN43: The location of sequences from the Original Braunvieh assembly not included in the graph. Numbers above the plot denote chromosomal identifiers.

Note S4.4
Differential expression analysis

We tested 13,085 genes that were expressed >= 1 CPM in at least eight samples (sample size from each group) for differential expression between *Mycobacterium bovis*-infected and non-infected control animals. We detected (adjusted FDR $\leq$ 0.05) 1,769 and 1,877 genes that were up-and down-regulated respectively in peripheral blood leukocytes of *Mycobacterium bovis*-infected cattle (Figure SN44). Of 12,813 genes of the Ensembl ARS-UCD1.2 genome annotation that were expressed at >=1 CPM in at least eight samples, 3610 (28.17%) were differentially expressed. Of 272 putatively novel genes that were expressed >=1 CPM in at least 8 samples, 36 (13.23%) were differentially expressed.

We found that genes relevant for the immune response were among the top differentially expressed genes with the greatest mean log-fold change (e.g., DEFB10 -8.24-fold, CXCL10 -3.30-fold, IL12B -3.11-fold, CXCL5 7.11-fold, CTLA4 4.25-fold, and CXCL8 5.70-fold), matching observations on an older reference genome annotation by McLoughlin et al. [13] Table SN48. Multidimensional scaling (MDS) representations of transcript abundance estimates from either all 13,085 genes (Figure SN45b) or 3646 differentially expressed genes (Figure SN45a) separated Mycobacterium bovis-infected from healthy cattle. We discovered more differentially expressed genes of the Ensembl ARS-UCD1.2 genome annotation than McLoughlin et al. [13] (3610 vs. 3250), likely due to a vastly improved genome assembly (27,115 vs. 24,616 genes are included in build 101 (ARS-UCD1.2) and build 73 (UMD3.1), respectively). Using data from the supplement provided by McLoughlin et al. [13], we were able to compare the expression levels of 2678 (out of 3250) differentially expressed genes between different genome builds (UMD31, standard and extended ARS-UCD1.2) and annotations (Figure SN46). Six genes with the greatest fold-change increase in expression reported by [13] had a very similar expression pattern from all assemblies considered.

Table SN48: The expression of 6 immune genes reported by McLoughlin *et al.* across different assemblies.

Ensembl gene ID	Gene symbol	Log2FC UMD3.1 (McLoughlin *et. al*)	Log2FC Standard ARS-UCD1.2	Log2FC Extended ARS-UCD1.2
ENSBTAG00000019716	CXCL8	2.435	2.512	2.512
ENSBTAG00000009812	CXCL5	2.763	2.831	2.831
ENSBTAG00000013170	CTLA4	1.849	2.088	2.088
ENSBTAG00000004741	IL12B	-2.129	-1.841	-1.841
ENSBTAG00000048737	DEFB10	-2.850	-3.042	-3.042
ENSBTAG00000001725	CXCL10	-1.712	-1.722	-1.722

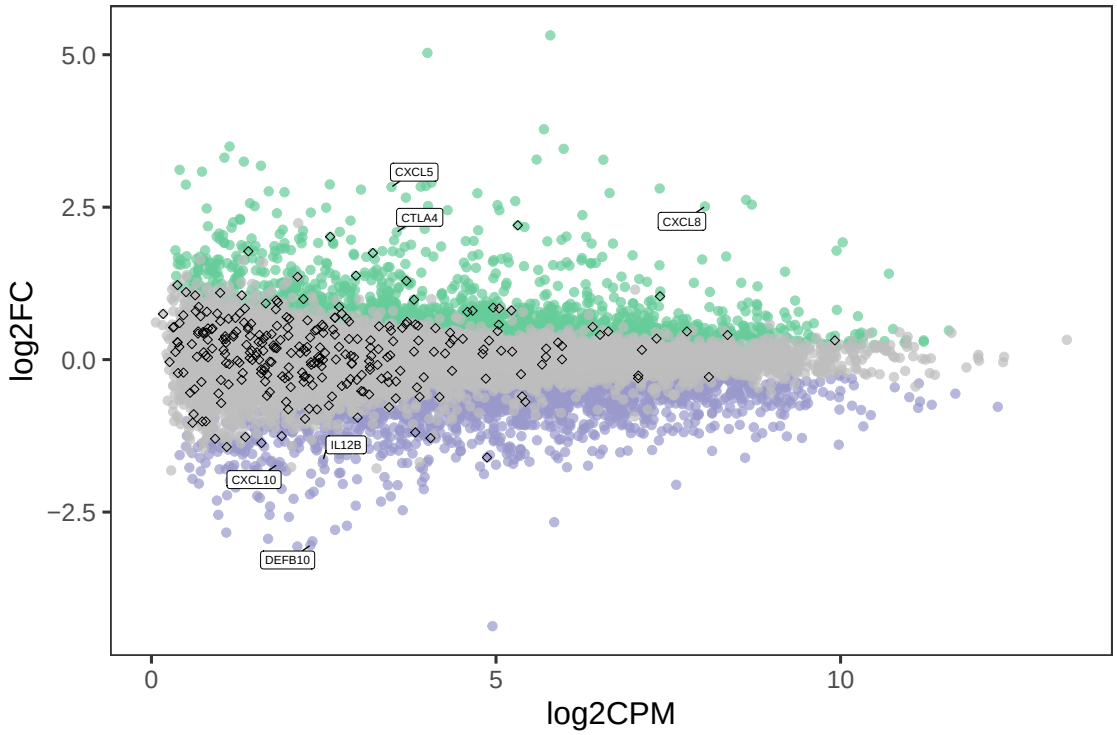


Figure SN44: Smear plot from the differential expression analysis. Grey, green, and purple color indicates genes with no expression difference, significant up-regulation, and down-regulation in peripheral blood leukocytes of *Mycobacterium bovis*-infected cattle. Diamonds indicate 272 putatively novel genes assembled from RNA sequencing reads mapping to non-reference sequences. Six genes reported by McLoughlin et al. [13] are indicated with the text labels.

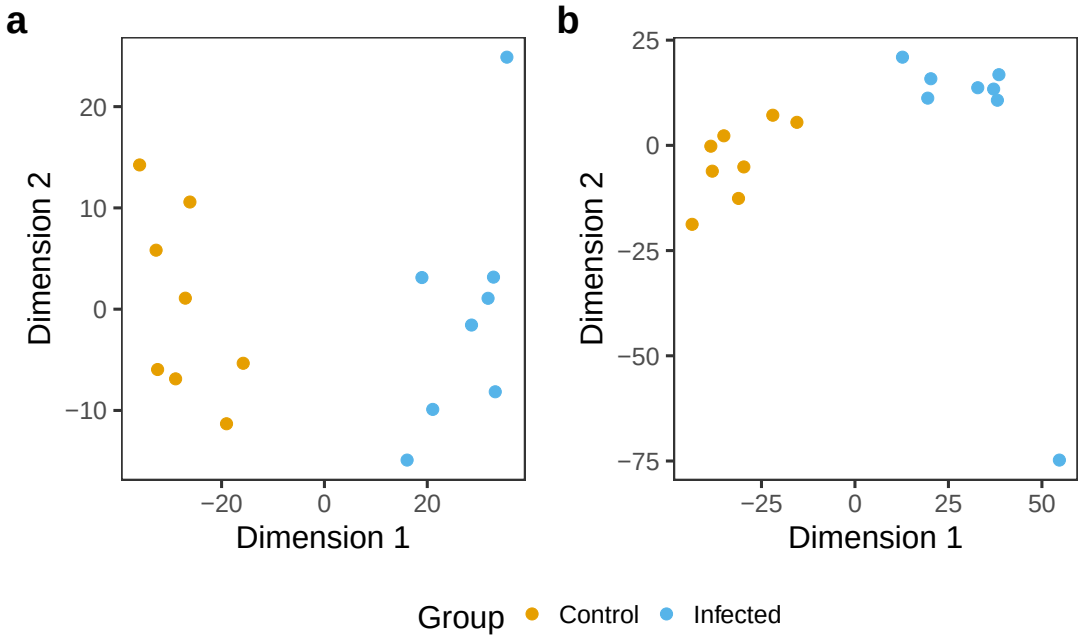


Figure SN45: Multidimensional scaling analysis (MDS) based on transcript abundance estimates of (a) 3646 differentially expressed, and (b) 13,085 genes with CPM>=1 in eight samples. Each point represents an individual *Mycobacterium bovis*-infected (blue) or control (orange) sample.

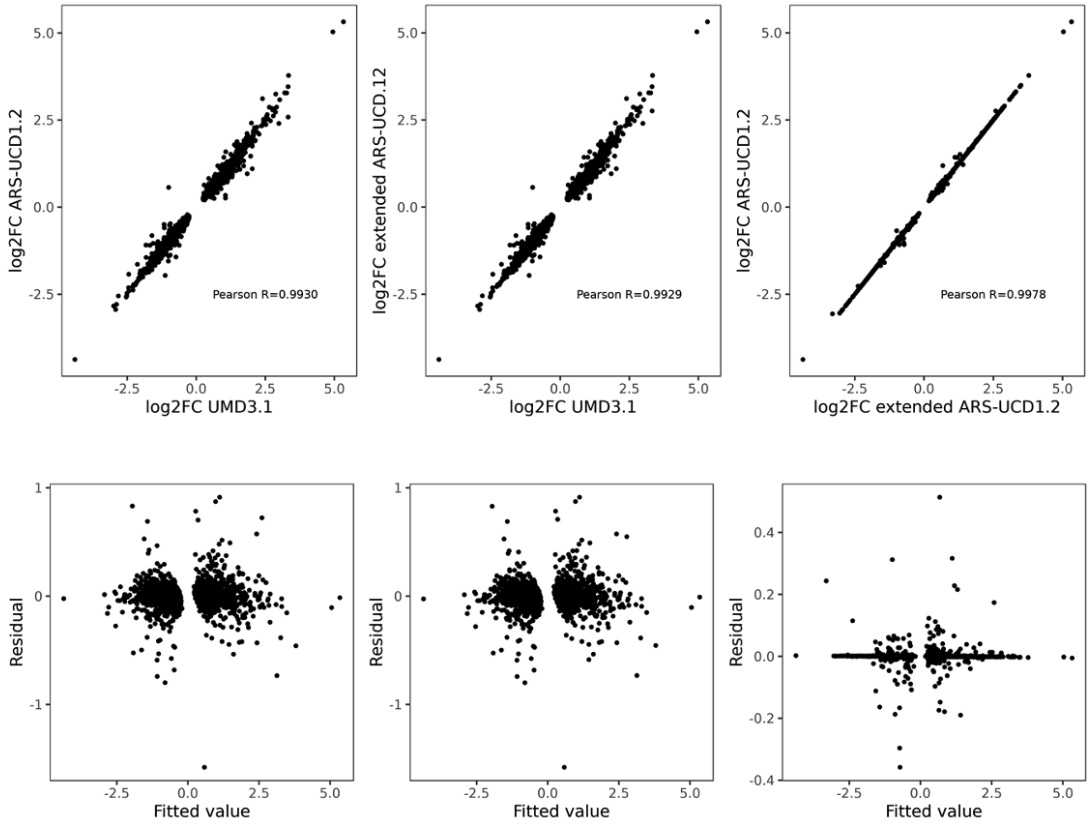


Figure SN46: Log2FC expression of 2678 genes between UMD3.1 (as reported in McLoughlin *et. al*), the standard ARS-UCD1.2, and extended ARS-UCD1.2 reference. Each point indicates the expression of a gene.

Note S4.5

Detailed description of the analysis workflow presented in the main paper

Step by step (manual) instruction to construct a multi-assembly graph from a collection of genome assemblies and characterize its structural variations. All steps can be automatically invoked with a workflow from a Github repository (https://github.com/AnimalGenomicsETH/bovine-graphs).

Pangenome graph construction

1. Estimate pairwise genetic distance between the assemblies.

```
# sketch assembly, done separately for each assembly
mash sketch -p {threads} -o {output} {input_assembly}

# combined all sketches
mash paste {output} {input_sketch1} {input_sketch2}

# estimate distance based on combined sketches
mash dist {input_combined_sketch} {input_combined_sketch} > {output_distance}

# visualize the genetic relationship as tree (optional)
scripts/phylo_tree_assembly.R
```

2. Graph construction
Assemblies are added to the graph based on their genetic distance to the backbone assembly. Less distant assemblies are added before the more distant ones.

```
# graph construction
minigraph --inv no -xggs -t {threads} {input_assemb1} {input_assemb2}  > {graph.gfa}

# subset graph file for easier access of nodes and edges information
# without the need for loading the sequences in nodes

# extract node information
awk '$1~/S/ { splt($5,chr,":"); split($6,pos,":"); split($7,arr,":");
            print $2,length($3),chr[3],pos[3],arr[3] }' {graph.gfa} > {graph_len.tsv}

# extract edge information
awk '$1 == "L"' {graph.gfa} > {graph_link.tsv}
```

3. Re-align each assembly to the multi-assembly graph
Separately realign each assembly to the multi-assembly graph to record the coverage for all nodes and edges in the graph.

```
minigraph -t {threads} --cov -x asm {graph.gfa} {assembly1.fa} > {graph_use_assembly1.gfa}
```

4. Combine node and edge coverage across assemblies.

```
# custom python script

scripts/comb_coverage.py -g {assemb1} {assemb2} -a {graph_name}

#will output node_coverage.tsv and edge_coverage_use.tsv
```

5. Use coverage data to label the nodes in the graph.

```
# custom R script

scripts/colour_node.R {assemb1} {assemb2} {graph_name}
```

6. Analyze the properties of the multi-assembly graph based on the node and edge labels.

```
# custom R script

scripts/run_core_nonref.R {graph_name}
```

Structural variations analysis

1. Identify bubbles in the graph
Bubbles are regions that diverged between assemblies which have a common start and stop node derived from reference sequences.

```
gfatools bubble {graph.gfa} > {bubble.tsv}
```

2. Identify the precise location of the structural variations from the multi-assembly graph
Paths in the bubble represent alleles of the structural variations. This step will enumerate all possible paths based on the start and stop node of the bubbles, done separately for biallelic (2 paths/alleles) and multi-allelic (>=3 alleles) bubbles. Finally, it labels structural variations according to the origin of the assemblies.

```
# custom Python scripts

# biallelic SV
scripts/get_bialsv.py -a {assemb1} {assemb2} > {output} #output: biallelic_sv.tsv

# multiallelic SV
scripts/get_multisv.py -a {assemb1} {assemb2} > {output} #multiallelic_sv.tsv

#trace path in each SV according to the origin of the assemblies
scripts/trace_path.py -g {assemb1} {assemb2} -a {graph_name} > {output}
```

3. Annotate the breakpoints detected in from the multi-assembly graph
Annotate the breakpoint of the structural variations using start and stop node coordinate for left and right breakpoints, respectively on the reference backbone coordinate. This step requires an annotation file from the backbone assembly.

```
# custom Python script

scripts/annot_breakpoints.py #output bubble_annot.tsv
```

4. Extract structural variation alleles in the bubbles
Extract non-ref alleles (excluding paths less than 100 bp, complete deletions, or paths without non-ref sequences) as a representative non-reference sequences of the pangenome. Sequences in nodes are not used directly, because multiple consecutive nodes might be part of the same allele, and they are representing a continuous sequences.

```
#custom Python scripts

# biallelic allele extraction
scripts/get_bialseq.py -a {assemb1} {assemb2} #output bialsv_seq.fa
# multiallelic allele extraction
scripts/get_multiseq.py -a {assemb1} {assemb2} # output multisv_seq.fa
# combined biallelic and multiallelic SV sequences as the representative of the non-ref sequences
cat {bialsv_seq.fa} {multisv_seq.fa} > {nonref_seq.fa}
```

Supplementary References

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

[2] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890, 2018.

[3] Haoyu Cheng, Gregory T Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, 18(2):170–175, 2021.

[4] Michael Alonge, Sebastian Soyk, Srividya Ramakrishnan, Xingang Wang, Sara Goodwin, Fritz J Sedlazeck, Zachary B Lippman, and Michael C Schatz. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome biology*, 20(1):1–17, 2019.

[5] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[6] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[7] Arang Rhie, Brian P Walenz, Sergey Koren, and Adam M Phillippy. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome biology*, 21(1):1–27, 2020.

[8] Samuel O'donnell and Gilles Fischer. MUM&Co: accurate detection of all SV types through whole-genome alignment. *Bioinformatics*, 36(10):3242–3243, 2020.

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

[10] Weidong Bao, Kenji K Kojima, and Oleksiy Kohany. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile Dna*, 6(1):1–6, 2015.

[11] Agnieszka A Golicz, Philipp E Bayer, Prem L Bhalla, Jacqueline Batley, and David Edwards. Pangenomics comes of age: from bacteria to plant and animal applications. *Trends in Genetics*, 36(2): 132–145, 2020.

[12] Edward S Rice, Sergey Koren, Arang Rhie, Michael P Heaton, Theodore S Kalbfleisch, Timothy Hardy, Peter H Hackett, Derek M Bickhart, Benjamin D Rosen, Brian Vander Ley, et al. Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. *Gigascience*, 9(4):giaa029, 2020.

[13] Kirsten E McLoughlin, Nicolas C Nalpas, Kévin Rue-Albrecht, John A Browne, David A Magee, Kate E Killick, Stephen DE Park, Karsten Hokamp, Kieran G Meade, Cliona O'Farrelly, et al. RNA-seq transcriptional profiling of peripheral blood leukocytes from cattle infected with Mycobacterium bovis. *Frontiers in immunology*, 5:396, 2014.

Acknowledgements

First, I would like to thank Prof. Hubert Pausch for having me as a doctoral student, supervising me over the years and providing a great environment for the research. I learned a lot about genomics, programming, problem solving, critical thinking, and scientific writing from your guidances. Also, thank you for giving me the freedom to explore research ideas and the trust to organize my time. I really appreciate all opportunities that I was given: including me in the other projects in the lab, providing funding to attend courses, and sending me to many international conferences. All these have become extremely valuable experiences.

Thanks to Prof Prof. Dr. Bernt Guldbrandtsen and Prof David MacHugh who have agreed to review this thesis.

I would like to thank the current and former members of the Animal Genomics Group for being very supportive for my day-to-day as a PhD student. A special mention to Dr. Alexander S. Leonard who has been extremely helpful in the last project and for dedicating time to proofread this thesis. Also to Maya and Meenu, who have become helpful peers since starting my PhD. I would also like to thank to staff at Agrovet Strickhof that have provided a great research facility. Thank you to Dorota Niedzwiecka for organizing all administrative tasks to ensure my smooth stay in Zurich.

Lastly, I would like to thank my families, especially my wife, who has accompanied me studying abroad.

Danang Crysnanto

08.01.1992
Citizen of Indonesia

Hoffeld 24 8057 Zurich
danangcrysnanto@gmail.com

Education

ETH Zurich — Dec 2017 – Present
Doctoral in Animal Genomics — *Zurich, Switzerland*

The University of Edinburgh — Aug 2016 – Aug 2017
*Msc in Quantitative Genetics and Genome Analysis with **Distinction*** — *Edinburgh, UK*

Bandung Institute of Technology — Oct 2010 – Oct 2014
*Bsc in Biology (Genetics) with **Cum Laude*** — *Bandung, Indonesia*

Publications

1. **Crysnanto D.**, A. S. Leonard, Z. H. Fang, and H. Pausch, 2021. Novel functional sequences uncovered through a bovine multi-assembly graph. *Proceedings of the National Academy of Sciences USA (PNAS)* 118:20
2. **Crysnanto D.**, and H. Pausch, 2020. Bovine breed-specific augmented reference graphs facilitate accurate sequence read mapping and unbiased variant discovery. *Genome Biology* 21:184
3. **Crysnanto D.**, C. Wurmser, and H. Pausch, 2019. Accurate sequence variant genotyping in cattle using variation-aware genome graphs. *Genetics Selection Evolution* 51:21
4. **Crysnanto D.**, and D. J. Obbard, 2019 Widespread gene duplication and adaptive evolution in the RNA interference pathways of the *Drosophila obscura* group. *BMC Evolutionary Biology* 19:1
5. Nosková A., M. Bhati, N. K. Kadri, and **D. Crysnanto** et al., 2021 Characterization of a haplotype-reference panel for genotyping by low-pass sequencing in Swiss Large White pigs. *BMC Genomics* 22:290
6. Nosková A., C. Wurmser, **D. Crysnanto**, A. Sironen, P. Uimari, et al., 2020 Deletion of porcine BOLL is associated with defective acrosomes and subfertility in Yorkshire boars. *Animal Genetics* 51: 945–949
7. Fang Z. H., A. Nosková, **D. Crysnanto**, S. Neuenschwander, P. Vögeli, et al., 2020 A 63-bp insertion in exon 2 of the porcine KIF21A gene is associated with arthrogryposis multiplex congenita. *Animal Genetics* 51: 820–823
8. Hiltpold M., G. Niu, N. K. Kadri, **D. Crysnanto**, Z. H. Fang, et al., 2020 Activation of cryptic splicing in bovine WDR19 is associated with reduced semen quality and male fertility. *PLoS Genetics*. 16:5
9. Bhati M., N. K. Kadri, **D. Crysnanto**, and H. Pausch, 2020 Assessing genomic diversity and signatures of selection in Original Braunvieh cattle using whole-genome sequencing data. *BMC Genomics* 21:27

Invited Talks

CIGENE Seminar Series — March 2021
Talk title: Bovine pangenomics — *NMBU Norway*

International Virtual Animal Breeding Journal Club — Sept 2020
Talk title: Bovine pangenomics — *USDA USA*

Plant and Animal Genome Conference (PAG) — Jan 2020
Talk title: Bovine pangenome graph enables unbiased genetic variants discovery — *San Diego USA*

Livestock Genomics session - Genome Informatics — Sept 2018
Talk title: Development of graph-based genotyping pipelines for bovine whole-genome data — *Cambridge UK*

Population Genetics (PopGroup) Conference — Jan 2018
Talk title: Widespread gene duplications in *Drosophila* immune system pathways — *Oxford UK*

Awards

Sir Kenneth Mather Memorial Prize — Jan 2018
An award for a MSc or PhD student of any UK University or Research Institutions
which shown an outstanding performance in the area of quantitative and population genetics. — *The Genetics Society*

The Douglas Falconer Prize — Oct 2017
An award for the Best Master Thesis in Quantitative Genetics and Genome Analysis — *The University of Edinburgh*

Bronze Medalist 21st International Biology Olympiad — Aug 2010
Won a medal on a highly prestigious International Bioscience Olympiad — *International Biology Olympiad*