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 through variation-aware genetic analyses that individual DNA sequences are compared to a set of haplotypes observed in the population rather than to the linear reference genome. Graph-based approaches have never been investigated in species with gigabase-sized genomes other than human. 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 1).

Using three different variation-aware genome graphs, this thesis has demonstrated 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 readily reveal a more complete pangenome, including 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 for genomic analysis. Chapter 5 also provides an example how these hitherto neglected sequences enable a better biological understanding of the molecular underpinnings of phenotypic variation.

0.1 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 based on a modified version of the *Graphtyper*

Table 1: **Comparison of the three genome-graph approaches implemented in this thesis**

	Chapter 2	**Chapter 3**	**Chapter 4**
Graphs	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 the graph	• Include large segments (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 best practice pipeline that also performs local read re-alignment. Therefore, the application of a region-specific graph-based method is competitive with the 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, it turned out that 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 large 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].

0.2 Variant prioritization to include in the graphs

Instead of only using genetic variants from the same cohort, informative graphs may 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 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 any benefits on read mapping accuracy. However, variant prioritization based on allele frequency increases the read mapping accuracy. Adding more variants prioritized based on allele frequency increases the mapping accuracy. However, the addition of variants with frequencies between 0.01 to 0.1 did not further improve the mapping accuracy. Thus, Chapter 3 indicates there seems to be 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. Analysis presented in Chapter 3 revealed that this threshold is population-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 pangenome graphs performed similar to 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 possible 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.

0.3 Investigation of the inaccessible genetic variations with multi-assembly graphs

Multi-assembly graph provides a platform to investigate complete genetic variations

Beyond integrating small variations as in the Chapter 2 and 3, graph genomes provide a more powerful framework to investigate large variations between individuals [2, 10, 11]. So far, there are 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.However, these studies were mostly 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 thousands of structural variations containing more than 70 megabases that are novel when compared to the ARS-UCD1.2 *Bos taurus* reference genome. An independent alignment of long-read sequencing data validated up to three-quarters of the structural variations in taurine and indicine breeds suggesting 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 underestimate the full genetic diversity between the six individual genomes. However, Chapter 4 already demonstrated that even with a simplified graph, biologically relevant information has readily been possible retrieved from the pangenome, suggesting that the existing cattle reference is not fully representative with an enormous potential of applying pangenome graph approaches in the cattle population.

Genetic variations in the 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 breeds 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 the theoretical expectation estimated for a diverged single human genome that differs at about 16 Mb from the reference [19]. Intriguingly, the pangenome revealed 4.4 megabases that were found in all assemblies but not in the 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.

0.4 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.

Repeat elements were enriched in the non-reference sequences

Repetitive elements account for the more than three-quarters of the non-reference sequences. Specifically, more than half of these repeat sequences belong to LINE/L1. LINE/L1 is still active in the bovine genomes and that 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 contribute 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 the breeds of cattle but not in the reference animals. Functional analysis indicates over-representation of non-ref genes related to immune response. Immune genes are highly polymorphic and contribute to genetic divergence and speciation [23]. Specifically, the Major Histocompatibility Locus at 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. In this regard, 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 considered to construct 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 and its expression is higher in disease-resistant animals. Previous studies have also observed 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.

0.5 Construction of the comprehensive and informative pangenome graphs for bovine population

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 their genomes are affected by polyploidization, higher repeat content, and larger effective population size [36]. However, the size of the bovine pangenome still grows when more genomes were added which indicates 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 increased the size of the pangenome by 20 Mb including 13 megabases private to gaur (Fig. 1). Yet, it needs to be seen whether this continued increase in pangenome size with still tractable computation still holds 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 prohibitively expensive, an initial assembly-based strategy might be followed by enriching multi-assembly graphs 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 in-

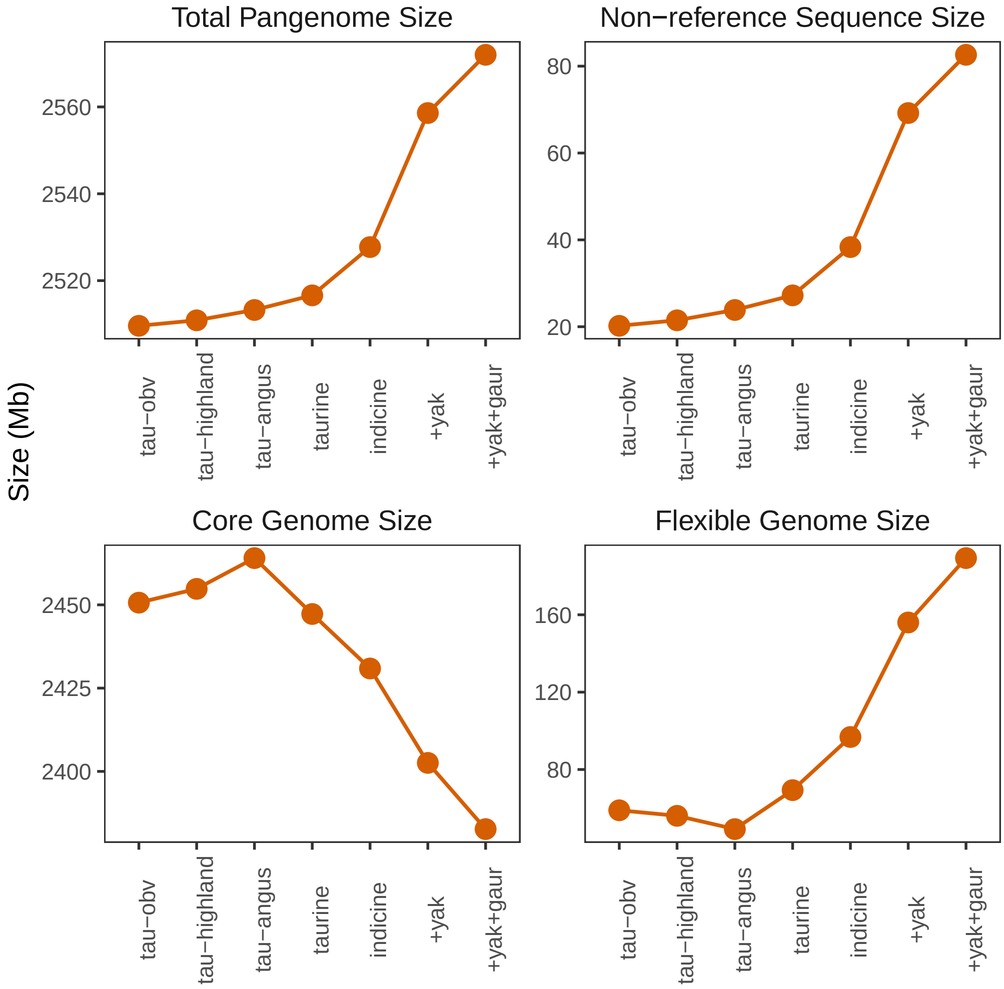

Figure 1: **Profile of the pangenome graph**
Pangenome graphs were constructed as in Chapter 4 (4 taurine breeds, 1 indicine breed, 1 yak) and complemented with a recently available gaur assembly. tau-*X* denotes a graph with taurine assemblies but excluding breed *X*. Taurine indicates a graph with 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.

tegrate haplotype information into the graph in order to indicate biologically plausible allele combinations. The recent development of the so-called dynamic genome graphs are appealing as they can iteratively update existing graphs once additional genomes become available, and subdivide a large graph 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 to be as static entity containing merely 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 epi-genomic data. As a proof of concept, analysis presented in Chapter 4 showed that labelling the nodes to track sample information is useful to characterize 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 already 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 the graph significantly 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 for as a reference structure for multiple layers of epi-genomics data. This approach is readily feasible in many livestock species for which vast amounts of functional omics data have been generated recently [45]. Overall, these graph resources will be highly valuable for the livestock genomics community to catalogue the global cattle diversity in order to perform comprehensive comparative genomics or even to identify beneficial alleles that are relevant for adaptation to future environmental changes.

0.6 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 assembly chosen as a backbone is critical for accurate and complete pangenome representation, especially to retrieve the real variable sequences diverged across animals rather than technical artifacts due to the incomplete assembly. Chapter 4 and Fig. 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].

Additionally, the pangenome will benefit from the use of haplotype-resolved assemblies. The mapping algorithm in *vg* (Chapter 3) utilized 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 the 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

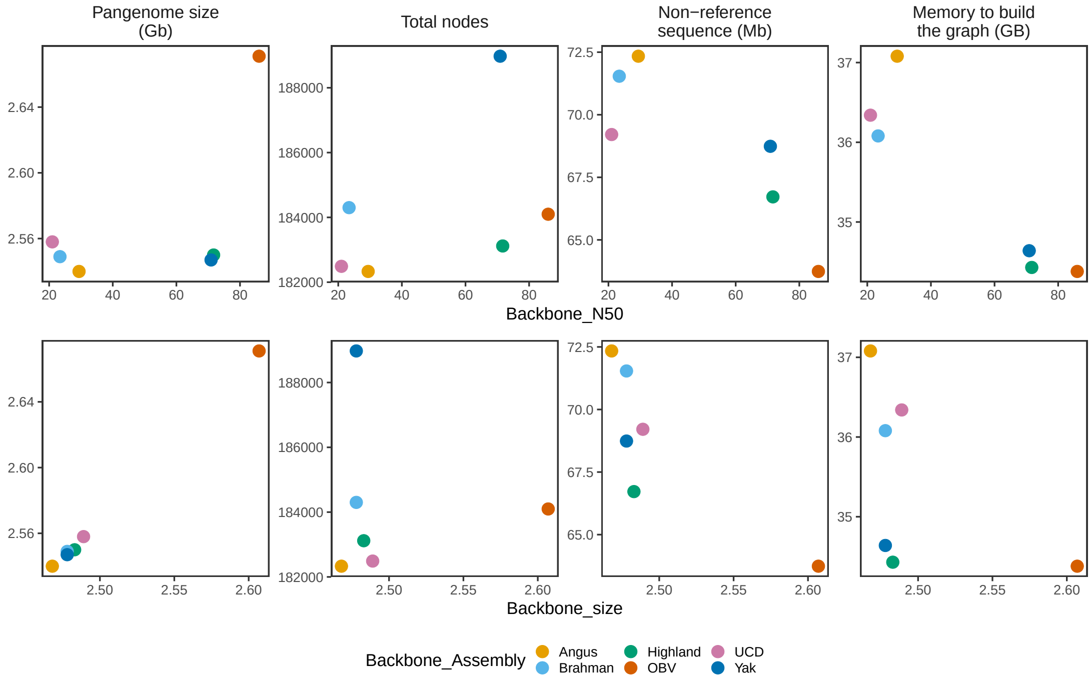

Figure 2: **Correlation between the backbone assembly size and quality with the profile of the pangenome graph**
A colored dot represents the backbone assembly from which that the graph was built from. N50 represents assembly contiguity with a higher number reflects a more contiguous assembly.

to the autosomes. The high quality of the novel HiFi-based assemblies now provide 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 was limited to structural variations of longer than 50bp (Table 2). Thus, to exploit the full potential of the pangenome, full graph models that can account all haplotypes of the individuals in the population and including their sites of variations, are required. Development a more comprehensive genome graph such as pggb (https://github.com/pangenome/pggb) or cactus [52] pipeline is promising, that can perform reference-free multi-genome alignment to generate a full graph containing complete variations. Utilizing the full graph models as implemented in pggb or cactus may uncover more variable and non-reference sequences, as an initial analysis revealed (Table 2). However, this approach is computationally demanding for whole-genome applications, likely because the resulting graph is more complex with many small nodes (Table 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.

Table 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.

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. 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 is supposed to 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 incomplete a single reference genome. The graph-based approaches presented in this thesis could provide a starting point for many analyses that have either not been possible so far 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 site of variations) that is on the one hand relevant for genomic 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 across species or diverged that might underlying the adaptive traits, domestication, or evolution of livestock species which can then be exploited to accelerate genetic progress [45, 62].

Some areas in livestock genomics with potential applications of the genome graphs are discussed as below

Unbiased genomic analyses using genome graphs

Genome graph approaches provide an immediate application 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 has been 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, limit alterations 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 explain 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 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.

Multiple studies in humans [2, 10, 75] have attempted to integrate sequence-resolved large structural variations detected using long read sequencing into 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 re-analysis the vast amount of whole-genome short-read re-sequencing data that have been collected in cattle to detect and genotype comprehensive structural variants. 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 variation markers detected from a linear reference genome representation. 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 re-sequencing data. Pangenome graphs offer the ability to catalogue more accurate and unbiased variations from the population, particularly in 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, graph genomes might be used to integrate diverse functional omics data for prioritizing variants used in genomic prediction. MacLeod et al. [79] showed that stratification of variants with functional omics data improves prediction accuracy than 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. [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 stresses. For example, successful 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 the undomesticated Bovinae 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. 1). Additionally, the trio binning assembly technique perform better in the hybrid cattle of a more diverged parents [46, 89]. This provides exciting opportunities to generate assemblies using the understudied or undomesticated cattle relatives as one of the parental that will generate diverse collections cattle assemblies that become ideal starting materials for 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*, pages 1–24, 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.