0.1 Genomic technologies to assess genetic variations in livestock

Cattle is an important livestock species for producing animal-based protein. The global cattle population is highly diverse due to intense selection for specific breeding goals, such as for 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 State 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 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 covered 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-out in segments of few hundred bases (Fig. 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 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 Genome Project was launched to coordinate global sequencing efforts and compile huge datasets. In their latest (8^{th}) run, the consortium has already catalogued more than 150 million variants

from more than 4000 cattle acrossf 200 breeds [10]. 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 improve the prediction accuracy of genomic selection [11]. Recently, low-pass sequencing (<1x) coupled with genotype imputation techniques were proposed as a cost-effective strategy to enable population-scale whole genome sequencing variant analysis [12].

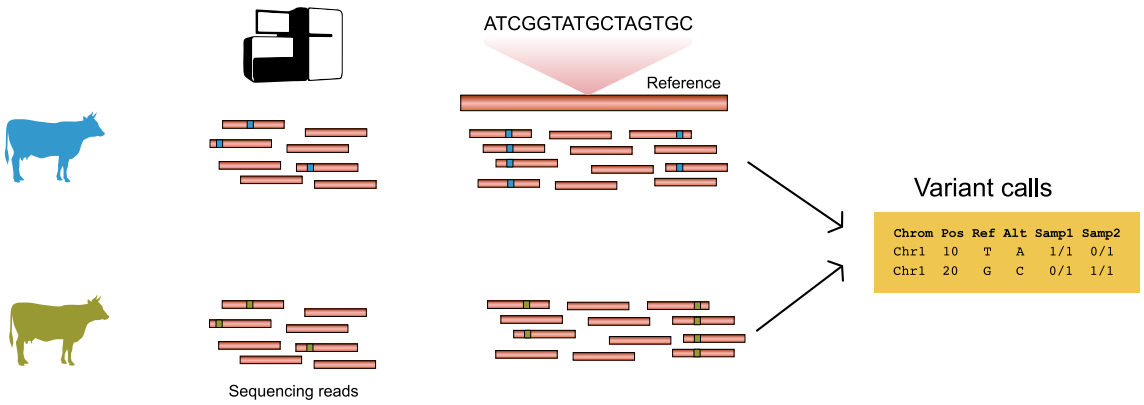

Figure 1: **Identification of genetic variants through re-sequencing**
Whole-genome sequences were fragmented into billions of short fragments which were then read by DNA sequencer in a massively parallel manner. The sequencing reads were compared (aligned) to the reference genome. Genetic variants were identified as nucleotide discordances relative to the reference sequences.

0.2 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 alignments, 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 the 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) sequencing [14]. The contig and scaffold N50 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 sophisticated 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 now enable the construction of assemblies of high continuity, correctness and completeness [13]. The recently generated assemblies exceed in quality the current bovine reference genome with contig N50 of larger than 70 Mb and could resolve complex genomic regions, e.g. major histocompatibility regions [17]. Trio binning takes advantage of the high heterozygosity 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], but now has been applied to broad range 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 the wild-relatives and under-represented breeds.

0.3 One reference genome is not enough

0.3.1 A single linear genome cannot fully represent species diversity

Despite recent spectacular quality improvements, the linear reference genomes still poorly represent the full genomic diversity within a species. A linear reference genome typically represents a mosaic haplotype of 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 approch, e.g., Duroc breed used for Sscrofa11.1 pig reference [21], San Clemente breed for domestic goat reference [22], and boxer breed for CanFam 3.1 dog reference [23]. While the selection of reference animals seems to be trivial, the resulting reference sequences do not necessarily reflect the most common allele in the population or from samples with the most ideal phenotypes [24]. Reference-guided variant discovery might reflect some properties of the reference

animal rather than the population; e.g., variant calling will output more variants when the reference contains rare alleles. Low et al. [25] found a striking difference in the number of 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. [26, 24] estimated that 2 million bases in the human reference genome are minor alleles.

0.3.2 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 cannot be mapped (Fig. 2) [27]. Together, this limitation is referred to as **soft reference bias**, which hampers genomic analysis that depends on the allelic balance such as heterozygous variant calling [28], allelic-specific expression [29], or analysis in the highly polymorphic regions [30]. Wu et al. [31] observed the impact of reference bias affecting 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. 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 has been observed in many species. Ameur et al. [33], Audano et al. [34] estimated that each human genome on average carries about 10 Mb non-reference bases. Long read data analysis across global ancestries discovered 8.5 Mb insertions observed in majority of the human population Audano et al. [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 (Sherman et al. 2019).

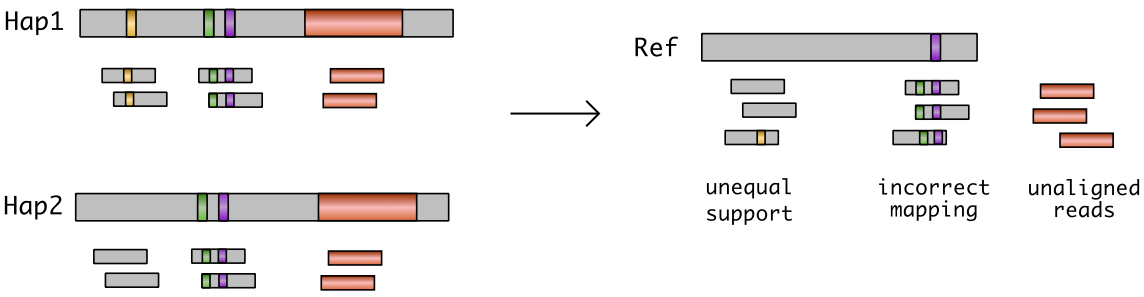

Figure 2: **Illustration of the reference allele bias.**
The origin of short sequencing reads of the sample (hap1 and hap2) are determined by alignments to the reference nucleotides. Thus, the comparison will always be biased towards nucleotides in the reference. Alignment of reads with alleles differing from reference might receive lower support than allele matches to the reference nucleotides (yellow stripe), results in incorrect alignments with multiple variations (green and purple stripes), or remain unmapped if the regions not present in the reference (e.g., large insertion, orange box). Grey background denotes reference sequences.

0.3.3 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. Genetic architecture of the bovine genomes has been shaped by various processes related domestication, admixture, introgression, local adaptation, and human-directed selection [1], resulting in the creation of more than 600 subpopulation (known as breeds) adapted for a variety of environmental conditions and selected for various breeding goals. Genetic diversity is higher in cattle than human populations [35]. 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 zebu [36]. 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 [37, 38]. Series of introgressions and hybridizations created specialized breeds with mosaic genomes, such as Brahman, composed of 10 % taurine and 90% indicine origin [39]. African cattle are generally admixture between *Bos taurus* x

Bos indicus, where the introgressed regions are selected for African pastoralism [40]. On average, each individual cattle carry more than 5 million variants relative to Bos taurus reference, which is higher than variations reported in the human population at about 3-4 million variants [11, 41]. The number of variants is higher in more diverged, indicine [39] or under-studied African cattle [40, 42]. Additionally, this amount likely underestimates the actual genetic diversity as it does not consider the structural variations, which are poorly characterized with short-read sequencing technology [43, 44].

0.4 Strategies to mitigate reference bias

0.4.1 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 analysis 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 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 [45]. Chen et al. [46] extended this idea into a so called reference flow approach, whereby it re-aligned sub-optimally mapped reads into a set of genomes from multiple population, that could reduce strongly heterozygous sites by 22%. Another effort, as in the human genome, is by continually expanding reference with alternative contigs in the polymorphic regions that are impossibly represented with a single haplotype. There were currently 13 updates with 261 alternate patches that add 109 Mb total length. 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 [47].

0.4.2 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 complete genome assemblies [48, 49]. Thus, more studies have now shifted from species-level references into population-

specific reference genomes, effectively creating more personalized genomes. Large genomic initiatives such as Vertebrate Genome Project (VGP, https://vertebrategenomesproject.org/), Darwin Tree of Life (https://www.darwintreeoflife.org/), or Earth Bio-genome Project [50] contributes to the explosion the number of genome assemblies across the tree of life accessible 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 the strict high-quality criterias [51]. On the other hand, some genomic initiatives focus to deeply characterize 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/). A similar internationally coordinated effort was also initiated for cattle with the Bovine Pangenome Consortium [20] that aim to generate reference-quality assemblies across global cattle breeds. There are already dozens of genomes from livestock species publicly available in the public repository. 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.

0.5 Transition from genomics to pangenomics

0.5.1 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 [52, 47, 53]. The term pan-genome (pan – whole, Greek) was first introduced by Tettelin et al. [54] to describe complete gene repertoire across Streptococcus agalactiae strains where 20% of the genes are variable across isolates. This concept was quickly adopted across the tree of life, including the agriculturally important plant and animal species, such as pig [55, 56], goat [57], and human [58, 59]. There has been rapid growth in the number of pangenome publications across years [52], with close to 8000 studies indexed by PubMed, although most currently focus on bacterial pangenomes.

0.5.2 Categorization of the pangenome

The content of a pangenome may be divided into the core and flexible genome (also known as dispensable or accessory genome, Fig. 3a). Core genome is common sequences across all individuals that is responsible for maintaining essential function (e.g., DNA replication, cellular homeostasis and cellular processes). This part of genomes is under purifying selection, thus having less diversity. Dispensable genomes are segments that vary across individuals. They are under less evolutionary constraint, 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 [60]. Thus, dispensable genomes are of particular interest for the studies of adaptive traits that might drive genetic differentiation and give population their distinguishing characteristics. In animals, the pangenome is largely dominated by core component (e.g., 96.67% of genes in the human) [58]. However, 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 [61]. Pangenomes have been extensively characterized in plants, among them are in rice [62], tomato [63], wheat [64]. They reported larger proportion of accessory genomes (>20%), particularly in polypoid, outcrossing, or species history of whole-genome duplications [65]. Higher ratio of flexible to core genome indicates a species with higher adaptability [66].

It is important to consider whether the pangenome is of either closed or open type. 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 included [60]. Many plant and animal pangenomes are a closed type in terms in the number of genes but open in terms of total sequence content [58, 60], which also suggests that the non-coding segments primarily drive the sequence variability across samples. Bacterial pangenomes are generally open type due to prevalence of horizontal gene transfer [67]. Sampling bias of underrepresented diversity (such as genetically related samples) could lead to the falsely concludingthe pangenome is complete [66]. With additional, sufficiently diverged samples, the pangenome would continue to grow. Thus, sampling strategy in a pangenome study should maximize diversity to fully retrieve the complete pangenome.

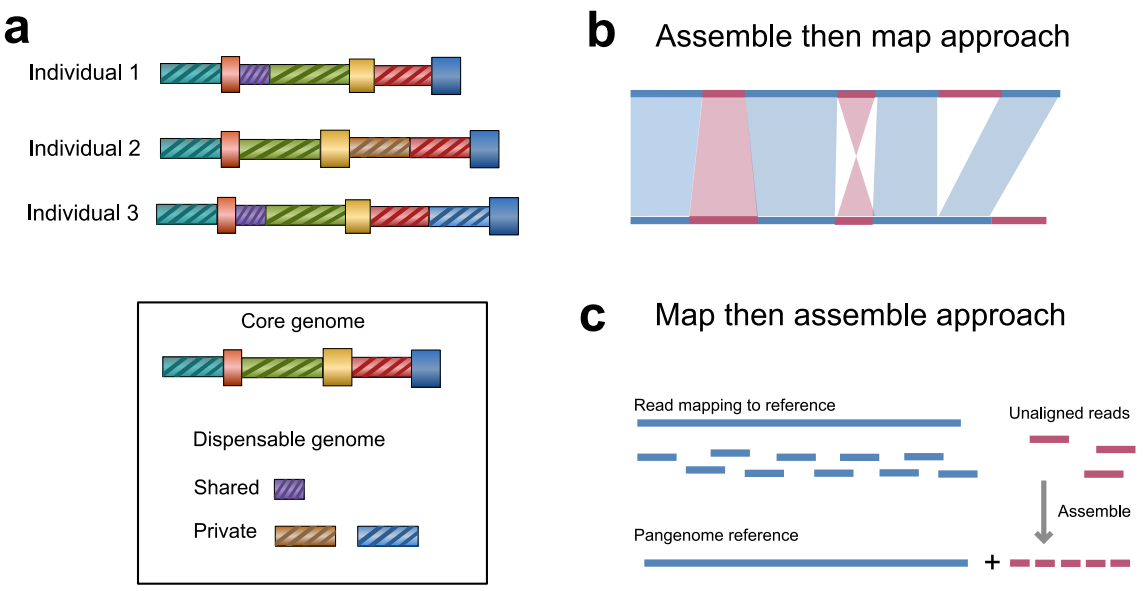

Figure 3: **The concept of pangenomes.**
(a) Pangenomes refers to a collection of individual genomes in the populations, which is further divided into core (shared by all members of populations) and flexible parts that the presence varies across individuals. Different strategies to build the pangenome **(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 into the reference. Unmapped sequences assembled and added as additional contigs to the reference sequences. Figures are adapted from [47] and [52].

0.5.3 Approaches to build a pangenome

There are two commonly used approaches to build a pangenome (Fig. 3bc): “assemble-then-map" and “map-then-assemble" (also known as map-to-pan) [60]. 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 [58, 57, 68]. 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 erroneous identification of the structural variations. Additionally, genome assemblies are still too expensive to be performed on the population-scale, limiting analysis only on a subset of individuals. To take advantage the massive amount of population-scale of the short-read sequencing data, the majority of recent pangenome studies utilize the “map-then-assemble"-approach [69, 70, 59]. Sequencing reads from each sample are independently mapped to the reference genome. The unmapped (or poorly mapped) reads are subsequently assembled to obtain the non-reference sequences. However, due to the nature of short-read-based assembly, most of the resulting contigs are fragmented,

making it difficult to locate the breakpoints' origins in the reference genome [59].

0.6 Graph-based pangenomics

0.6.1 Graphs as richer reference structures to integrate the genetic diversity

The pangenome approaches based on unmapped reads or assembly comparison, as discussed above, rely on collections of linear genomes and do not attempt to provide coherent representation that relates all genomes. Considering the prevalence of genetic variations across individuals in the population and availability of abundant 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 collective efforts led to a new genomic discipline known as Computational Pangenomics, see review [71, 72, 73].

Graph-based models (also known as genome graphs or sequence graphs) are currently proposed as data structures that unify a collection of related sequences in a compact way (Fig. 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 sample differs from each other form bubbles, with alternate paths representing different alleles [74]. Traversing (or walk through the graphs) recovers the initial input sequences as well as all possible recombinations.

0.6.2 Graph genomes implementations

The first pangenome graph implementation was based on the DBG (*De Bruijn Graphs*). Sequencing reads from all samples were fragmented into k-mer length k, and 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 map to the origin of the samples. Iqbal et al. [75] 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

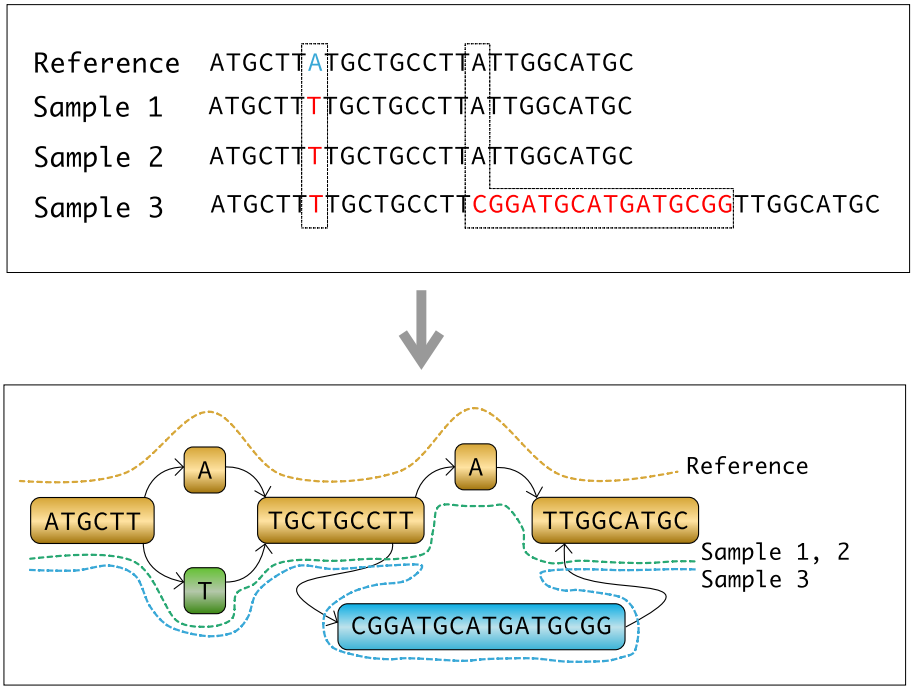

Figure 4: **Graph-based pangenome approach.**
(a) The majority of the pangenome studies follow the classical pangenome approach, where multiple linear genomes are compared without compressing redundant information and might lack orthology relationships. **(b)** Graph-based pangenome approach offers unified and richer multiple genomes representation. Nodes contain DNA sequences and nodes with continuous sequences 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 retrace the original sets of sequences from which it was built (dashed line).

reads, DBG-based approaches are not suitable for resequencing study, although a recent study attempts to embed a long-range path information into the graph [76].

Current well-established graph genome implementations establish a variation graph as an extension of the linear reference genome [77, 28, 78, 79, 80]. 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 graphs. 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].

Graphtyper is the first open-source variation graph-based software designed for genotyping from a local (region-specific) graph [77, 81]. It uses a variant file (*VCF*) as input source of variant sites and a reference assembly as backbone of the graph. Because of

the limited variations modelled by a VCF file, the output graph is directed and acylic containing insertions and deletions but not necessarily complex variations (Fig. 5b). *Graphtyper* applies a two-step genotyping proces. The "discovery step" is similar to linear reference-guided variant analysis. Sequencing reads are mapped to the linear genome and variants are identified from the alignments. This step is then followed by read realignment towards local graphs. To this end, *Graphtyper* first constructs small regional graphs of 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 does, however, not fully eliminate reference bias because it relies on the global read placement by a linear mapper. However, this design makes it highly efficient as evidenced with scalable joint genotyping of close to 50,000 Icelander samples [81]. Additionally, Graphtyper outperformed current state-of-the-art linear genome-based tools (e.g., *SAMtools* and *GATK*), particularly from more refined variants surrounding Indels with considerably reduced Mendelian errors [77].

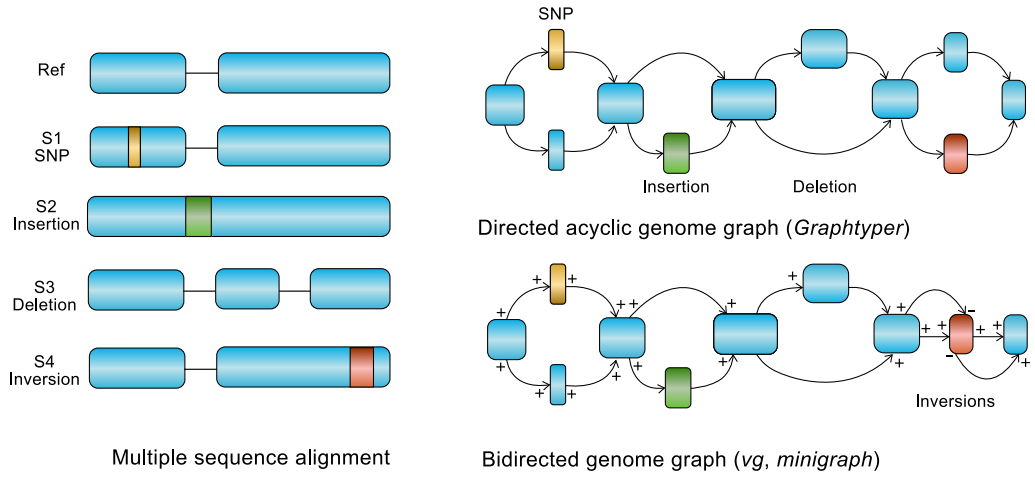

Figure 5: **Various genome graph implementations and representations of variations in the graphs**
(a) multiple sequence alignments capturing sequence relationships. **(b)** directed genome graphs underlying the data structure of *Graphtyper*, similar to multiple sequence alignments but with compressing 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 [73].

0.6.3 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 a gigabase-sized

genome [28]. The basic structure of *vg* is a bidirected sequence graph that can express the strand-ness of the input sequences (Fig. 5c). Each edge endpoint has an independent orientation to indicate whether the forward or reverse sequences are spelled out when visiting the node [71]. Therefore, *vg* can represent variations with complex topology e.g., inversion or translocation. Haplotype information from the sample are stored in an index so that analysis from the graph can consider haplotype information [82]. 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 [83]. The index construction is the computationally most demanding step because all *k*-bp paths in the graphs need to be enumerated, which is intractable in complex regions with high variant density. In practice, *vg* can handle complex region by indexing on a simplified graph e.g., retaining only biologically plausible paths informed by the haplotype index [83]. 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 minimizer-based mapping paradigm and restrict the mapping that conforms the haplotype paths. It can achieve the same mapping speed as the *BWA* linear mapper with more accurate alignment performance, especially for application related to structural variant genotyping [84].

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

0.7.1 Multi-assembly graphs as a platform 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 [85]. 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) [86]. Building a graph directly from a collection of genome assemblies is a better approach to capture more comprehensive genetic variation. 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 across these resources.

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

0.7.2 Strategies to build multi-assembly graphs

Accurate multi-genome alignment is the key for the multi-assembly-based graph approach. However, multiple genome alignment 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 [87] 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 species using Cactus facilitate a thorough comparative genomics across vertebrate phylogeny [88, 89]. Hickey et al. [90] applied the *vg toolkit* to induce graphs from *Cactus* alignment of 12 yeast strains. 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 generation. It uses a selected genome as a backbone of the graph which is then iteratively augmented with unaligned or poorly mapped segments from the other assemblies. *Minigraph*simplifies the general bidirected sequence graph data model resulting

in a faster and a more straightforward graph analysis. For example, it enforces linearity of the input genomes that produces graph containing insertions and deletions between genomes but ignoring events that breaks the linearity, such as translocations. Constraining alignment to an anchor genome also ensures that the graphs devoid of complex and highly tangled parts which are difficult to interpet [93]. Comparative genomics using a pangenome graph built with minigraph containing human and its closely related ape species revealed important biological insights, including the evolution of repeat-rich regions in primates [91], inaccessible with a linear genome. An unpublished graph pipeline (Pangenome Graph Builder, https://github.com/pangenome/pggb) aims to build 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 the infancy and requires further testing.

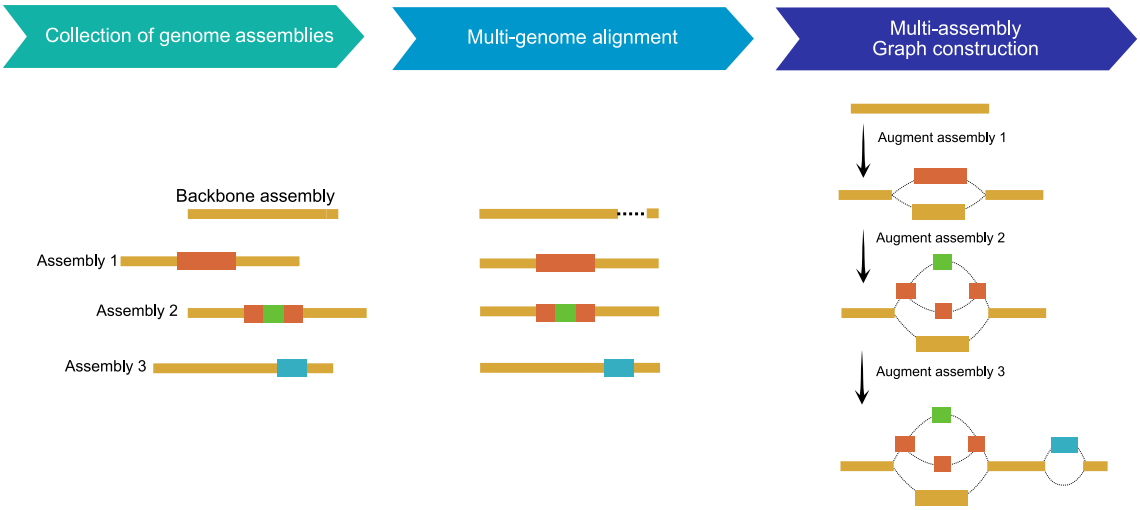

Figure 6: **Construction of the multi-assembly graphs**
Multi-assembly graph is built based on multi-genome alignment from the collections of genome assemblies (*left*, *middle*). In the *minigraph* approach (*right*), the graph is built iteratively from alignment of the genome to the backbone or to the existing graphs, which is then augmented with diverged sequences from the alignment.

0.8 Utilization of graph genomes for genomic analyses

Graph genome approaches have been developed for a wide range of genomic analyses, but largely have only been applied to 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 genotyping. Multiple studies [77, 28, 78, 79] then assessed the performance of graph-based methods on whole-genome variant discovery and genotyping. 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, particularly for the subset of reads that differs from the reference and substantial reduction in the bias of calling large indels.

Pritt et al. [27] estimated that with carefully selected variants, genome graphs could rescue 1.2 million incorrectly mapped reads from 30-fold coverage of human whole-genome sequencing data compared 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. The benefit of mapping to a graph translates to substantial improvement in calling indels with sufficient accuracy for population genomic inference. Grytten et al. [96] extended 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 motif from *vg* graphs, Tognon et al. [97] identified variations in regulatory regions affecting the gene expression that otherwise missed with a linear genome.

Graph genome approaches were also rigorously exploited to investigate large (structural) variations. 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 [43]. 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. [82] constructed a *vg* graph from 167 thousand structural variations detected from long-read data across diverse human ancestry. Re-genotyping of 5202 short-read sequencing data using this graph considerably improves the SV genotyping. Further analyses lead to 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 line assemblies enabled identification of a hitherto unknown 10 kb insertion that is associated with a seed phenotype.

0.9 Applications of the pangenome

0.9.1 Pangenome analysis in plant genomes

Pangenome studies in plants successfully identified a large number of genes not included in the reference and highlight the substantial contribution of large variations into the dynamic of the pangenome. For example, pangenome analysis on 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 from *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. Construction pangenomes from 725 tomato accessions, Gao et al. [63] revealed 4873 genes absent from the reference genome and discovered 2 kb promoter insertions regulating fruit flavour that lost during domestication but present in the ancestral accession. An increasing number of studies shift towards the graph-based approach, which is pioneered by construction of a graph-based soybean pangenome [98].

0.9.2 Pangenome analysis in human genomes

In human genomics, pangenome analyses following the large scale re-sequencing initiatives reveal several important insights. The 1000 Genomes project revealed that each genome carries more than two-fold regions affected by structural variations (8.9 Mb) than small variations (3.6 Mb) [101]. Importantly, they discovered 240 genes related to immunoglobulin and glycoprotein with homozygous (knock out) deletions in multiple populations, suggesting its dispensable role [41]. Pangenome analyses focusing on the Icelandic population, [102] found a common 766 bp insertions (allele frequency of 0.65) associated with decreased risk of myocardial infarctions, where the signals are stronger than the SNPs-based association. A follow-up study based on 3622 samples sequences using Nanopore (the largest long-read-based pangenome study to date) found that Icelanders carries on average large insertions covering 10.02 Mb genomic regions and identified a tandem repeat motif strongly associated with height [103]. Application of the customized pangenome pipeline in Chinese Han population detected 29.5 Mb non-reference sequence, including 185 genes missing from the reference genome [58]. Sherman and Salzberg [47] reported markedly larger non-reference sequences from African pangenome suggesting the substantial underrepresentation of the African diver-

sity in the reference.

0.9.3 Pangenome analysis in livestock genomes

Pangenome approaches have also been applied in the livestock species, although at a lower rate than the plants or humans. The most notable is the analysis of the 44 genomes spanning all extant Ruminant families reveal the genomic basis underlying the evolutionary innovations such as multi-chambered stomach, headgear, cursorial locomotion, and dentition [104]. Initial livestock pangenome analyses in animals rely on the assembly of unmapped reads. For example, Holden et al. [69] identified 4.6 Mb novel insertions from assemblies of non-aligned reads from three dog breeds including novel insertions in six known disease-associated loci. Analysis of unmapped reads from the reference individual in song bird (*Parus major*), Laine et al. [70] uncovered 1822 genes missing in the reference annotation, including *TRY1*, which is highly expressed in the reference bird. A similar effort to characterize the 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 analysis in animals now transition into a more direct assembly comparison. Comparison between Angus (*Bos taurus*) and Brahman (*Bos indicus)* haplotypes-resolved assemblies [25] uncovered an extra copy of *FADS2P1* gene, which proposed to confer the heat resistance in *Bos indicus*. Analysis of the unaligned sequences between 10 goat assemblies [57] recovered 38.3 Mb non-reference insertions and identified 2 Mb assembly error in ARS-1 goat reference genome that includes prolactin gene region. Analysis of 12 Eurasian pig de-novo assemblies retrieved 72.5 Mb novel insertions absent in Duroc-based reference assemblies [55, 56]. Additionally, they also discovered a non-reference insertion segregates at high frequency in Chinese breeds (but not in European breeds) encompassing the *TIG3* gene region, which is important for fatty acid metabolism.

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/.

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

[10] 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.

[11] 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.

[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, 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.

[17] 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.

[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, Konstantinos Billis, William Chow, Lel Eory, Heather A Finlayson, Paul Flicek, 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, 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):720797, aug 2020. ISSN 2041-1723. doi: 10.1101/720797.

[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] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

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

[48] 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.

[49] 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.

[50] 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.

[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. ISSN 1476-4687. doi: 10.1038/s41586-021-03451-0.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] Sean Walkowiak, Liangliang Gao, Cecile Monat, Georg Haberer, Mulualem T Kassa, Jemima Brinton, Ricardo H Ramirez-Gonzalez, Markus C Kolodziej, Emily Delorean, Dinushika Thambugala, et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature*, pages 1–7, 2020.

[65] 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.

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

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

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

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

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

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

[84] 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.

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

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

[87] 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.

[88] 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.

[89] Diane P Genereux, Aitor Serres, Joel Armstrong, Jeremy Johnson, Marinescu, et al. A comparative genomics multitool for scientific discovery and conservation. *Nature*, 587(7833):240–245, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2876-6.

[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. doi: 10.1146/annurev-arplant-080720-105454.

[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*, page 782755, 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, 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.

[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, 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 dna and rna sequence reads from the bovine reference individual. *BMC genomics*, 16(1):1–7, 2015.