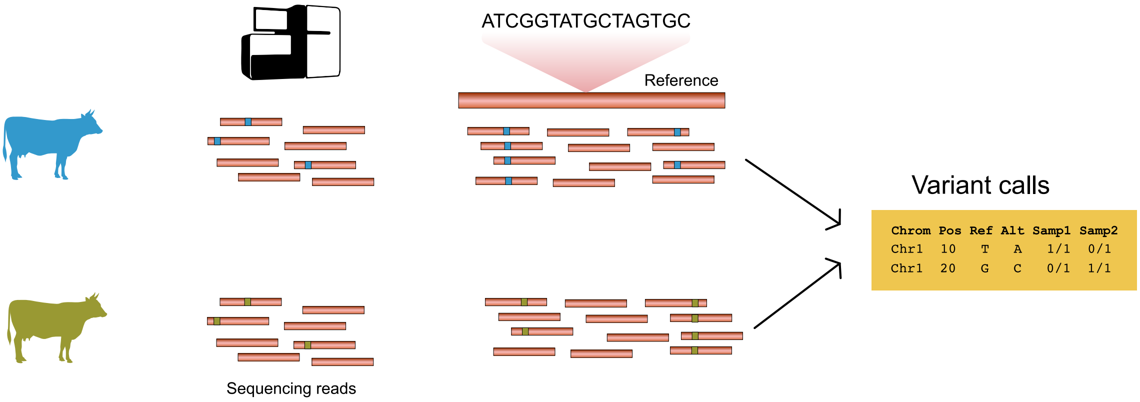
0.1 Genomic technologies to assess genetic variations in livestock

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 (Zhang et al., 2020). 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 (Georges et al., 2019).

Genomic selection had been proposed to further accelerate genetic gain (Meuwissen et al., 2001). To this end, the genetic value of an individual is estimated based on genome-wide molecular marker information. Genotyping arrays were first developed to assess variation at thousands of polymorphic sites in the genome. The genotype information is then linked to phenotype to determine markers associated with agriculturally-important traits (Goddard and Hayes, 2009) or to derive the prediction equation for genomic selection (Meuwissen et al., 2001). More than 3 million cattle in the USA have already been genotyped (Wiggans et al., 2017). However, variations covered by chip-based genotyping are not comprehensive enough to pinpoint causal mutations underlying the traits (Pausch et al., 2017).

This limitation prompted the utilization of whole-genome sequencing based on the short-read sequencing technology. 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 (DePristo et al., 2011). 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 (Regalado, 2020; Wetterstrand, 2020). 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 (8th) run, the consortium has already catalogued more than 150 million variants from more than 3500 cattle acrossf 200 breeds (Hayes and Daetwyler, 2019). 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 (Daetwyler et al., 2014). 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 (Snelling et al., 2020).



Chrom	Pos	Ref	Alt	Samp1	Samp2
Chr1	10	T	A	1/1	0/1
Chr1	20	G	C	0/1	1/1

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 (Bickhart et al., 2020).

The first cattle reference genome (Btau 3.1 and Btau 4.0) was assembled in 2009 from sequencing reads from bacterial artificial chromosome (BAC) and whole-genome shotgun (WGS) sequencing (Elsik et al., 2009). 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 (Zimin et al., 2009). The most recent cattle reference genome (ARS-UCD 1.2) was assembled from single-molecule real-time (SMRT) long-read sequencing data and scaffolded with optical mapping. 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 (Rosen et al., 2020). 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 (Bickhart et al., 2020). The recently generated assemblies exceed in quality the current bovine reference genome with contig N50 exceeding 70 Mb and could resolve complex genomic regions, e.g. major histocompatibility regions (Rice et al., 2020). 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) (Koren et al., 2018), but now has been applied to broad range cattle breeds, including undomesticated and/or cattle relatives (Yak, Gaur, Bison) (Oppenheimer et al., 2021). Recently, the Bovine Pangenome Consortium (Heaton et al., 2021) 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 are still a poor representation of the full genomic diversity of a species. A linear reference genome represents only one 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 (Rosen et al., 2020). Reference assemblies from other livestock species were generated using a similar approch, e.g., Duroc breed used for Sscrofa11.1 pig reference (Warr et al., 2020), San Clemente breed for domestic goat reference (Bickhart et al., 2017), and boxer breed for CanFam 3.1 dog reference (Lindblad-Toh et al., 2005). 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 (Ballouz et al., 2019). 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. (2019) 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. (Shukla et al., 2019; Ballouz et al., 2019) estimated that 2 million bases in the human reference genome are minor alleles.

0.3.2 Insufficient representation of linear genomes cause reference bias

Because the alignment algorithms compare the reads towards the reference and try to minimize differences, the reference-guided variant discovery will always be 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) (Pritt et al., 2018). 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 (Garrison et al., 2018), allelic-specific expression (Salavati et al., 2019), or analysis in the highly polymorphic regions (Dilthey et al., 2015). Wu et al. (2018) 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) (Colquhoun et al., 2020). 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 (Low et al., 2019). 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. (2018); Audano et al. (2019) 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. (2019). 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).

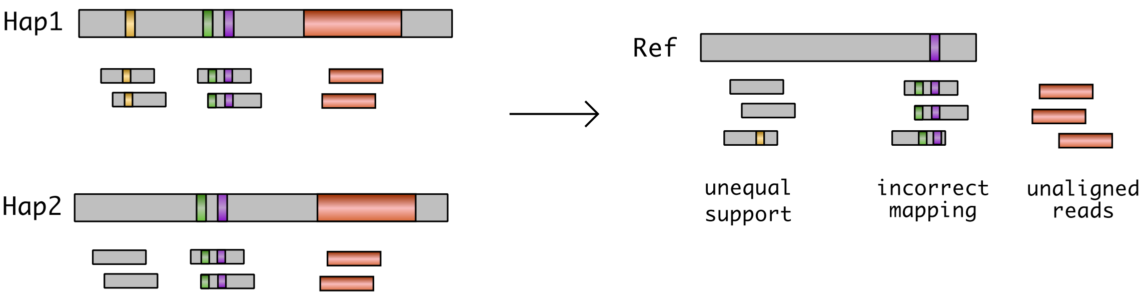


Figure 2: **Illustration of the reference allele bias.**
The origin of short sequencing reads of the sample (hap1 and hap2) are determined by alignments to the reference nucleotides. Thus, the comparison will always be biased towards nucleotides in the reference. Alignment of reads with alleles differing 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 magnified 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 (Zhang et al., 2020), 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 (Charlier et al., 2016). 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 (Pitt et al., 2019). 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 (Low et al., 2019). However, these characteristics are not strict as there are numerous local cattle breeds optimized for specialized breeding goals (Signer-Hasler et al., 2017; Upadhyay et al., 2019). Series of introgressions and hybridizations created specialized breeds with mosaic genomes, such as Brahman, composed of 10 % taurine and 90% indicine origin (Koufariotis et al., 2018). African cattle are generally admixture between *Bos taurus* x *Bos indicus*, where the introgressed regions are selected for African pastoralism (Kim et al., 2020). 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 (Daetwyler et al., 2014; Sudmant et al., 2015). The number of variants is higher in more diverged, indicine (Koufariotis et al., 2018) or under-studied African cattle (Kim et al., 2020, 2017). 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 (Mahmoud et al., 2019; Chaisson et al., 2019).

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. (2019) 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 (Kaminow et al., 2020). Chen et al. (2021) 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 (Sherman and Salzberg, 2020).

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 (Miga et al., 2020; Logsdon et al., 2021). 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 (Lewin et al., 2018) 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 further scaffolding with optical mapping results in the chromosome-scale assemblies and fulfilling the strict high-quality criteria (Rhie et al., 2020). 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 (Heaton et al., 2021) that aim to generate reference-quality assemblies across global cattle breeds. There are already dozens of genomes from livestock species publicly available in NCBI assembly. As in other species, the application of third generation sequencing technology results in an outstanding improvement in assembly quality. This was pioneered with goat genome that improved the contiguity over 400 times compared to the previous short-read based assembly (Bickhart et al., 2017). 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

Accumulating evidence suggests that a single genome cannot represent the full diversity of a species, motivating the development of pangenomes. 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 (Bayer et al., 2020; **?**; Sherman and Salzberg, 2020; Della Coletta et al., 2021). The term pan-genome (pan – whole, Greek) was first introduced by Tettelin et al. (2005) 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 (Li et al., 2017; Tian et al., 2019), goat (Li et al., 2019), and human (Duan et al., 2019; Sherman et al., 2019). There has been rapid growth in the number of pangenome publications across years (Bayer et al., 2020), 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 (Golicz et al., 2020). 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) (Duan et al., 2019). 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 (Gerdol et al., 2020). Pangenomes have been extensively characterized in plants, among them are in rice (Zhao et al., 2018), tomato (Gao et al., 2019), wheat (Walkowiak et al., 2020). They reported larger proportion of accessory genomes (>20%), particularly in polypoid, outcrossing, or species history of whole-genome duplications (Tao et al., 2019). Higher ratio of flexible to core genome indicates a species with higher adaptability (Tranchant-Dubreuil et al., 2018).

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 (Golicz et al., 2020). Many plant and animal pangenomes are a closed type in terms in the number of genes but open in terms of total sequence content (Duan et al., 2019; Golicz et al., 2020), which also suggests that the non-coding segments primarily drive the sequence variability across samples. Bacterial pangenomes are generally open due to prevalence of horizontal gene transfer (Soucy et al., 2015). Sampling bias of underrepresented diversity (such as genetically related samples) could lead to the falsely concludingthe pangenome is complete (Tranchant-Dubreuil et al., 2018). 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.

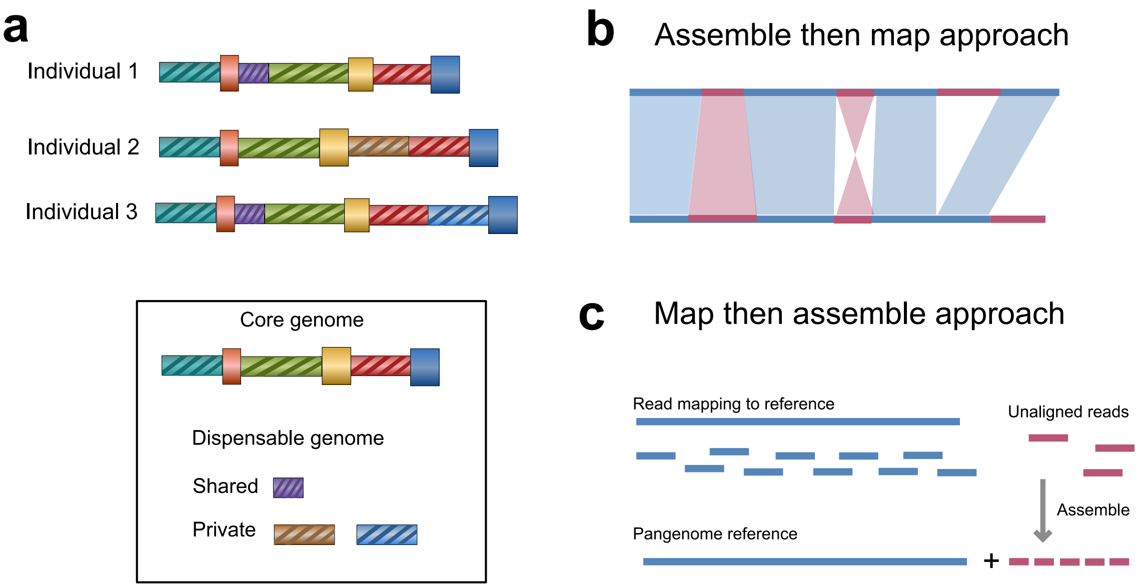


Figure 3: **The concept of pangenomes.**
(a) Pangenomes refers to a collection of individual genomes in the populations, which is further divided into core (shared by all members of populations) and flexible parts that the presence varies across individuals. Different strategies to build the pangenome **(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 (Sherman and Salzberg, 2020) and (Bayer et al., 2020).

0.5.3 Approaches building the 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) (Golicz et al., 2020). 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 (Duan et al., 2019; Li et al., 2019; Eisfeldt et al., 2020). 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 (Holden et al., 2018; Laine et al., 2019; Sherman et al., 2019). 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 (Sherman et al., 2019).

0.6 Graph-based pangenomics

0.6.1 Graphs as a richer reference structure 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 (Paten et al., 2017; com, 2018; Eizenga et al., 2020).

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 (Paten et al., 2018). Traversing (or walk through the graphs) recovers the initial input sequences as well as all possible recombinations.

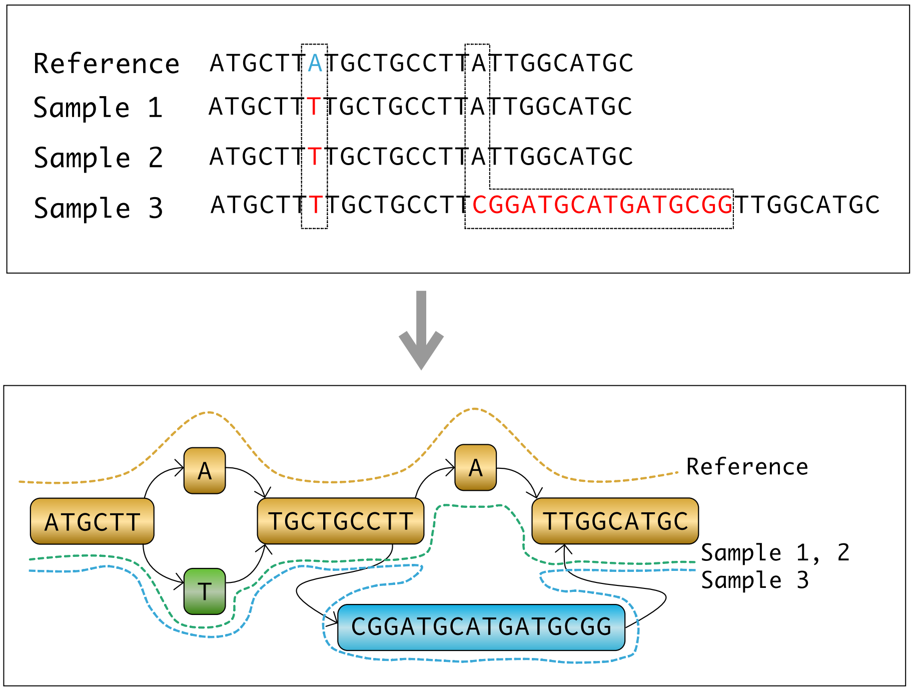

Reference ATGCTTATGCTGCCTTATTGGCATGC
Sample 1 ATGCTTTTGCTGCCTTATTGGCATGC
Sample 2 ATGCTTTTGCTGCCTTATTGGCATGC
Sample 3 ATGCTTTTGCTGCCTTCGGATGCATGATGCGGTTGGCATGC
A
A
ATGCTT
TGCTGCCTT
TTGGCATGC
T
CGGATGCATGATGCGG
Reference
Sample 1, 2
Sample 3

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).

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. (2012) developed *Cortex*, a coloured DBG-based pangenome tool. They used it to construct a population graph from 164 human samples and identified 3.2 Mb novel sequences that are absent in the human reference genome. Because the genomic coordinates are discarded by fragmenting the reads, DBG-based approaches are not suitable for resequencing study, although a recent study attempts to embed a long-range path information into the graph (Turner et al., 2018).

Current well-established graph genome implementations establish a variation graph as an extension of the linear reference genome (Eggertsson et al., 2017; Garrison et al., 2018;

Sibbesen et al., 2018; Rakocevic et al., 2019; Kim et al., 2019). 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 (Garrison et al., 2018).

Graphtyper is the first open-source variation graph-based software designed for genotyping from a local (region-specific) graph (Eggertsson et al., 2017, 2019). 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 (Eggertsson et al., 2019). 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 (Eggertsson et al., 2017).

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 (Garrison et al., 2018). 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 (Paten et al., 2017). 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 (Sirén et al., 2020a). Graph

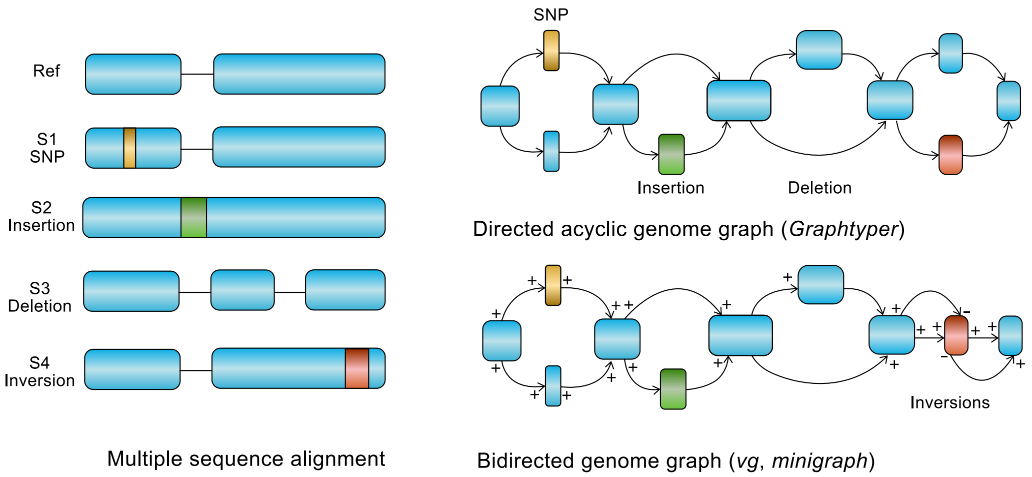

Ref
S1
SNP
S2
Insertion
S3
Deletion
S4
Inversion
Multiple sequence alignment
SNP
Insertion
Deletion
Directed acyclic genome graph (*Graphtyper*)
Inversions
Bidirected genome graph (*vg*, *minigraph*)

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 (Eizenga et al., 2020).

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 (Sirén, 2017). 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 (Sirén, 2017). 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 structural variant genotyping (Sirén et al., 2020b).

To be continued with more comments

References

Computational pan-genomics: status, promises and challenges. *Briefings in bioinformatics*, 19(1):118–135, 2018.

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

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

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

P. E. Bayer, A. A. Golicz, A. Scheben, J. Batley, and D. Edwards. Plant pan-genomes are the new reference. *Nat. Plants*, 6:914–920, 2020.

D. Bickhart, J. McClure, R. Schnabel, B. Rosen, J. Medrano, and T. Smith. Symposium review: advances in sequencing technology herald a new frontier in cattle genomics and genome-enabled selection. *Journal of dairy science*, 2020.

D. M. Bickhart, B. D. Rosen, S. Koren, B. L. Sayre, A. R. Hastie, S. Chan, J. Lee, E. T. Lam, I. Liachko, S. 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.

M. J. Chaisson, A. D. Sanders, X. Zhao, A. Malhotra, D. Porubsky, T. Rausch, E. J. Gardner, O. L. Rodriguez, L. Guo, R. L. Collins, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature communications*, 10(1):1–16, 2019.

C. Charlier, W. Li, C. Harland, M. Littlejohn, W. Coppieters, F. Creagh, S. Davis, T. Druet, P. Faux, F. Guillaume, et al. Ngs-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. *Genome research*, 26(10):1333–1341, 2016.

N.-C. Chen, B. Solomon, T. Mun, S. Iyer, and B. Langmead. Reference flow: reducing reference bias using multiple population genomes. *Genome biology*, 22(1):1–17, 2021.

R. M. Colquhoun, M. B. Hall, L. Lima, L. W. Roberts, K. M. Malone, M. Hunt, B. Letcher, J. Hawkey, S. George, L. Pankhurst, et al. Nucleotide-resolution bacterial pan-genomics with reference graphs. *bioRxiv*, 2020.

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

R. Della Coletta, Y. Qiu, S. Ou, M. B. Hufford, and C. N. Hirsch. How the pan-genome is changing crop genomics and improvement. *Genome biology*, 22(1):1–19, 2021.

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

A. Dilthey, C. Cox, Z. Iqbal, M. R. Nelson, and G. McVean. Improved genome inference in the mhc using a population reference graph. *Nature genetics*, 47(6):682–688, 2015.

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

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

H. P. Eggertsson, S. Kristmundsdottir, D. Beyter, H. Jonsson, A. Skuladottir, M. T. Hardarson, D. F. Gudbjartsson, K. Stefansson, B. V. Halldorsson, and P. Melsted. Graphtyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature communications*, 10(1):1–8, 2019.

J. Eisfeldt, G. Mårtensson, A. Ameur, D. Nilsson, and A. Lindstrand. Discovery of novel sequences in 1,000 swedish genomes. *Molecular biology and evolution*, 37(1):18–30, 2020.

J. M. Eizenga, A. M. Novak, J. A. Sibbesen, S. Heumos, A. Ghaffaari, G. Hickey, X. Chang, J. D. Seaman, R. Rounthwaite, J. Ebler, et al. Pangenome graphs. *Annual Review of Genomics and Human Genetics*, 21:139–162, 2020.

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

L. Gao, I. Gonda, H. Sun, Q. Ma, K. Bao, D. M. Tieman, E. A. Burzynski-Chang, T. L. Fish, K. A. Stromberg, G. 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.

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

M. Georges, C. Charlier, and B. Hayes. Harnessing genomic information for livestock improvement. *Nature Reviews Genetics*, 20(3):135–156, 2019.

M. Gerdol, R. Moreira, F. Cruz, J. Gómez-Garrido, A. Vlasova, U. Rosani, P. Venier, M. A. Naranjo-Ortiz, M. Murgarella, S. Greco, et al. Massive gene presence-absence variation shapes an open pan-genome in the mediterranean mussel. *Genome biology*, 21(1):1–21, 2020.

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

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

B. J. Hayes and H. 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.

M. P. Heaton, T. P. Smith, D. M. Bickhart, B. L. Vander Ley, L. A. Kuehn, J. Oppenheimer, W. R. Shafer, F. T. Schuetze, B. Stroud, J. C. McClure, et al. A reference genome assembly of simmental cattle, bos taurus taurus. *Journal of Heredity*, 2021.

L. A. Holden, M. Arumilli, M. K. Hytönen, S. Hundi, J. Salojärvi, K. H. Brown, and H. Lohi. Assembly and analysis of unmapped genome sequence reads reveal novel sequence and variation in dogs. *Scientific reports*, 8(1): 1–11, 2018.

Z. Iqbal, M. Caccamo, I. Turner, P. Flicek, and G. McVean. De novo assembly and genotyping of variants using colored de bruijn graphs. *Nature genetics*, 44(2):226–232, 2012.

B. Kaminow, S. Ballouz, J. Gillis, and A. Dobin. Virtue as the mean: Pan-human consensus genome significantly improves the accuracy of rna-seq analyses. *bioRxiv*, 2020.

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

J. Kim, O. Hanotte, O. A. Mwai, T. Dessie, S. Bashir, B. Diallo, M. Agaba, K. Kim, W. Kwak, S. Sung, et al. The genome landscape of indigenous african cattle. *Genome biology*, 18(1):1–14, 2017.

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

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

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

V. N. Laine, T. I. Gossmann, K. van Oers, M. E. Visser, and M. A. Groenen. Exploring the unmapped dna and rna reads in a songbird genome. *BMC genomics*, 20(1):1–12, 2019.

H. A. Lewin, G. E. Robinson, W. J. Kress, W. J. Baker, J. Coddington, K. A. Crandall, R. Durbin, S. V. Edwards, F. Forest, M. T. 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.

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

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

K. Lindblad-Toh, C. M. Wade, T. S. Mikkelsen, E. K. Karlsson, D. B. Jaffe, M. Kamal, M. Clamp, J. L. Chang, E. J. Kulbokas, M. C. Zody, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069):803–819, 2005.

G. A. Logsdon, M. R. Vollger, P. Hsieh, Y. Mao, M. A. Liskovykh, S. Koren, S. Nurk, L. Mercuri, P. C. Dishuck, A. Rhie, et al. The structure, function and evolution of a complete human chromosome 8. *Nature*, pages 1–7, 2021.

W. Y. Low, R. Tearle, C. Liu, S. Koren, A. Rhie, D. M. Bickhart, B. D. Rosen, Z. N. Kroneberg, S. B. Kingan, E. Tseng, et al. Haplotype-resolved cattle genomes provide insights into structural variation and adaptation. *BioRxiv*, page 720797, 2019.

M. Mahmoud, N. Gobet, D. I. Cruz-Dávalos, N. Mounier, C. Dessimoz, and F. J. Sedlazeck. Structural variant calling: the long and the short of it. *Genome biology*, 20(1):1–14, 2019.

T. H. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, 2001.

K. H. Miga, S. Koren, A. Rhie, M. R. Vollger, A. Gershman, A. Bzikadze, S. Brooks, E. Howe, D. Porubsky, G. A. Logsdon, et al. Telomere-to-telomere assembly of a complete human x chromosome. *Nature*, 585(7823):79–84, 2020.

J. Oppenheimer, B. D. Rosen, M. P. Heaton, B. L. Vander Ley, W. R. Shafer, F. T. Schuetze, B. Stroud, L. A. Kuehn, J. C. McClure, J. P. Barfield, et al. A reference genome assembly of american bison, bison bison bison. *Journal of Heredity*, 112(2):174–183, 2021.

B. Paten, A. M. Novak, J. M. Eizenga, and E. Garrison. Genome graphs and the evolution of genome inference. *Genome research*, 27(5):665–676, 2017.

B. Paten, J. M. Eizenga, Y. M. Rosen, A. M. Novak, E. Garrison, and G. Hickey. Superbubbles, ultrabubbles, and cacti. *Journal of Computational Biology*, 25(7):649–663, 2018.

H. Pausch, I. M. MacLeod, R. Fries, R. Emmerling, P. J. Bowman, H. D. Daetwyler, and M. 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.

D. Pitt, N. Sevane, E. L. Nicolazzi, D. E. MacHugh, S. D. Park, L. Colli, R. Martinez, M. W. Bruford, and P. Orozco-terWengel. Domestication of cattle: Two or three events? *Evolutionary applications*, 12(1):123–136, 2019.

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

G. Rakocevic, V. Semenyuk, W.-P. Lee, J. Spencer, J. Browning, I. J. Johnson, V. Arsenijevic, J. Nadj, K. Ghose, M. C. Suciu, et al. Fast and accurate genomic analyses using genome graphs. *Nature genetics*, 51(2):354–362, 2019.

A. 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/.

A. Rhie, S. A. McCarthy, O. Fedrigo, J. Damas, G. Formenti, S. Koren, M. Uliano-Silva, W. Chow, A. Fungtammasan, G. L. Gedman, et al. Towards complete and error-free genome assemblies of all vertebrate species. *BioRxiv*, 2020.

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

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

M. Salavati, S. J. Bush, S. Palma-Vera, M. E. McCulloch, D. A. Hume, and E. L. Clark. Elimination of reference mapping bias reveals robust immune related allele-specific expression in crossbred sheep. *Frontiers in genetics*, 10:863, 2019.

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

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

H. G. Shukla, P. S. Bawa, and S. Srinivasan. hg19kindel: ethnicity normalized human reference genome. *BMC genomics*, 20(1):1–17, 2019.

J. A. Sibbesen, L. Maretty, and A. Krogh. Accurate genotyping across variant classes and lengths using variant graphs. *Nature genetics*, 50(7):1054–1059, 2018.

H. Signer-Hasler, A. Burren, M. Neuditschko, M. Frischknecht, D. Garrick, C. Stricker, B. Gredler, B. Bapst, and C. Flury. Population structure and genomic inbreeding in nine swiss dairy cattle populations. *Genetics Selection Evolution*, 49(1):1–13, 2017.

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

J. Sirén, E. Garrison, A. M. Novak, B. Paten, and R. Durbin. Haplotype-aware graph indexes. *Bioinformatics*, 36 (2):400–407, 2020a.

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

W. M. Snelling, J. L. Hoff, J. H. Li, L. A. Kuehn, B. N. Keel, A. K. Lindholm-Perry, and J. K. Pickrell. Assessment of imputation from low-pass sequencing to predict merit of beef steers. *Genes*, 11(11):1312, 2020.

S. M. Soucy, J. Huang, and J. P. Gogarten. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16(8):472–482, 2015.

P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H.-Y. Fritz, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571): 75–81, 2015.

Y. Tao, X. Zhao, E. Mace, R. Henry, and D. Jordan. Exploring and exploiting pan-genomics for crop improvement. *Molecular plant*, 12(2):156–169, 2019.

H. Tettelin, V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. 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.

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

C. Tranchant-Dubreuil, M. Rouard, and F. Sabot. Plant pangenome: impacts on phenotypes and evolution. *Annual Plant Reviews Online*, pages 453–478, 2018.

I. Turner, K. V. Garimella, Z. Iqbal, and G. McVean. Integrating long-range connectivity information into de bruijn graphs. *Bioinformatics*, 34(15):2556–2565, 2018.

M. Upadhyay, S. Eriksson, S. Mikko, E. Strandberg, H. Stålhammar, M. A. Groenen, R. P. Crooijmans, G. Andersson, and A. M. Johansson. Genomic relatedness and diversity of swedish native cattle breeds. *Genetics Selection Evolution*, 51(1):1–11, 2019.

S. Walkowiak, L. Gao, C. Monat, G. Haberer, M. T. Kassa, J. Brinton, R. H. Ramirez-Gonzalez, M. C. Kolodziej, E. Delorean, D. Thambugala, et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature*, pages 1–7, 2020.

A. Warr, N. Affara, B. Aken, H. Beiki, D. M. Bickhart, K. Billis, W. Chow, L. Eory, H. A. Finlayson, P. Flicek, et al. An improved pig reference genome sequence to enable pig genetics and genomics research. *GigaScience*, 9(6): giaa051, 2020.

K. A. Wetterstrand. Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP), 2020. URL `www.genome.gov/sequencingcostsdata`.

G. R. Wiggans, J. B. Cole, S. M. Hubbard, and T. S. Sonstegard. Genomic selection in dairy cattle: the usda experience. *Annual review of animal biosciences*, 5:309–327, 2017.

D.-D. Wu, X.-D. Ding, S. Wang, J. M. Wójcik, Y. Zhang, M. Tokarska, Y. Li, M.-S. Wang, O. Faruque, R. Nielsen, et al. Pervasive introgression facilitated domestication and adaptation in the bos species complex. *Nature ecology & evolution*, 2(7):1139–1145, 2018.

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

Q. Zhao, Q. Feng, H. Lu, Y. Li, A. Wang, Q. Tian, Q. Zhan, Y. Lu, L. Zhang, T. Huang, et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature genetics*, 50(2):278–284, 2018.

A. V. Zimin, A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard, et al. A whole-genome assembly of the domestic cow, bos taurus. *Genome biology*, 10(4): 1–10, 2009.