DISS. ETH NO.

Bovine Pangenome Graphs Facilitate Unbiased Genomic Analysis

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

Danang Crysnanto

M.Sc., The University of Edinburgh

Master in Quantitative Genetics and Genome Analysis

born on 08.01.1992

citizen of Indonesia

accepted on the recommendation of

Prof XXX

Prof YYY, ZZZ 2021

Table of Contents

List of Figures

List of Tables

Abstract

English abstract

Zusammenfassung

Deutsch abstract

Chapter 1

Analysis of the cohort-specific graphs

Preface: Bridging text between Chapter 1 and Chapter 2

In this chapter, I assessed the feasibility of the genome graphs in cattle genome. I assessed *Graphtyper* software for variant genotyping in cattle. *Graphtyper* performed two round of genotyping. The first is to discover variants from linear genome. And the second round used the variants discovered in the first round to construct a local genome graph and used it to refine the genotypes. I discovered that *graph genotyping* using *Graphtyper* is highly accurate in cattle and outperform current approaches e.g., *SAMtools,* GATK that are based on linear reference. My work is the first to apply *graph genome* for sequence variant genotyping in the livestock genome. I implemented the graph genotyping pipeline and it is now publicly available at https://github.com/danangcrysnanto/Graph-genotyping-paper-pipelines.

Contribution: I and Hubert Pausch conceived the study, I wrote the genotyping pipelines and performed all analyses. I wrote the initial draft of the manuscript with input from Hubert Pausch.

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

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

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

[2] Chair of Animal Breeding, TU München, Freising, Germany.

Published in *Genetic Selection Evolution (2019) 51:21.*

Abstract

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

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

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

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

1.1 Introduction

The sequencing of important ancestors of many cattle breeds revealed millions of sequence variants that are polymorphic in dairy and beef populations (Hoff et al., 2017; Stothard et al., 2015; Boussaha et al., 2016; Jansen et al., 2013). In order to compile an exhaustive catalog of polymorphic sites that segregate in Bos taurus, the 1000 Bull Genomes consortium was established (Daetwyler et al., 2014; Hayes and Daetwyler, 2019). The 1000 Bull Genomes Project imputation reference panel facilitates to infer sequence variant genotypes for large cohorts of genotyped animals thus enabling genomic investigations at nucleotide resolution (Daetwyler et al., 2014; Pausch et al., 2017a; Bouwman et al., 2018; Raymond et al., 2018).

Sequence variant discovery and genotyping typically involves two steps that are carried out successively (Nielsen et al., 2011; Guo et al., 2014; Goodwin et al., 2016; Pfeifer, 2017): first, raw sequencing data are generated, trimmed and filtered to remove adapter sequences and bases with low sequencing quality, respectively, and aligned towards a linear reference genome using, e.g., *Bowtie* (Langmead and Salzberg, 2012) or the Burrows-Wheeler Alignment (*BWA*) software (Li and Durbin, 2009). The aligned reads are subsequently compared to the nucleotide sequence of a reference genome in order to discover and genotype polymorphic sites using, e.g., *SAMtools* (Li et al., 2009) or the Genome Analysis Toolkit (*GATK*) (McKenna et al., 2010; Van der Auwera et al., 2013; Poplin et al., 2018). Variant discovery may be performed either in single- or multi-sample mode. The accuracy (i.e., ability to correctly genotype sequence variants) and sensitivity (i.e., ability to detect true sequence variants) of sequence variant discovery is higher using multi-sample than single-sample approaches particularly when the sequencing depth is low (Liu et al., 2013; Cheng et al., 2014; Baes et al., 2014; Kumar et al., 2014; DePristo et al., 2011). However, the genotyping of sequence variants from multiple samples simultaneously is a computationally intensive task, particularly when the sequenced cohort

is large and diverse and had been sequenced at high coverage (Poplin et al., 2018). The multi-sample sequence variant genotyping approach that has been implemented in the *SAMtools* software has to be restarted for the entire cohort once new samples are added. *GATK* implements two different approaches to multi-sample variant discovery, i.e., the *UnifiedGenotyper* and *HaplotypeCaller* modules, with the latter relying on intermediate files in *gVCF* format that include probabilistic data on variant and non-variant sites for each sequenced sample. Applying the *HaplotypeCaller* module allows for separating variant discovery within samples from the estimation of genotype likelihoods across samples. Once new samples are added to an existing cohort, only the latter needs to be performed for the entire cohort, thus enabling computationally efficient parallelization of sequence variant genotyping in a large number of samples.

Genome graph-based methods consider non-linear reference sequences for variant discovery (Rakocevic et al., 2019; Eggertsson et al., 2017; Novak et al., 2017b; Garrison et al., 2018; Sibbesen et al., 2018). A variation-aware genome graph may incorporate distinct (population-specific) reference sequences and known sequence variants. Recently, the *Graphtyper* software has been developed in order to facilitate sequence variant discovery from a genome graph that has been constructed and iteratively augmented using variation of the sequenced cohort (Eggertsson et al., 2017). So far, sequence variant genotyping using variation-aware genome graphs has not been evaluated in cattle.

An unbiased evaluation of the accuracy and sensitivity of sequence variant genotyping is possible when high confidence sequence variants and genotypes are accessible that were detected using genotyping technologies and algorithms different from the ones to be evaluated (Li et al., 2018). For species where such a resource is not available, the accuracy of sequence variant genotyping may be evaluated by comparing sequence variant to microarray-derived genotypes (e.g., (Jansen et al., 2013; DePristo et al., 2011)). Due to the ascertainment bias in SNP chip data, this com-

parison may overestimate the accuracy of sequence variant discovery particularly at variants that are either rare or located in less-accessible genomic regions (Li, 2014; Malomane et al., 2018).

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

1.2 Methods

Selection of animals We selected 49 Original Braunvieh (OB) bulls that were either frequently used in artificial insemination or explained a large fraction of the genetic diversity of the active breeding population. Semen straws of the bulls were purchased from an artificial insemination center and DNA was prepared following standard DNA extraction protocols.

Sequencing data pre-processing All samples were sequenced on either an Illumina HiSeq 2500 (30 animals) or an Illumina HiSeq 4000 (19 animals) sequencer using 150 bp paired-end sequencing libraries with insert sizes ranging from 400 to 450 bp. Quality control (removal of adapter sequences and bases with low quality) of the raw sequencing data was caried out using the *fastp* software (version 0.19.4) with default parameters (Chen et al., 2018). The filtered reads were mapped to the UMD3.1 version of the bovine reference genome (Zimin et al., 2009) using *BWA mem* (version 0.7.12) (Li and Durbin, 2009) with option-M to mark shorter split hits as secondary alignments, default parameters were applied in all other steps. Optical and PCR duplicates were marked using *Samblaster* (version 0.1.24) (Faust

and Hall, 2014). The output of *Samblaster* was converted into *BAM* format using *SAMtools view* (version 1.3) (Li et al., 2009), and subsequently coordinate-sorted using *Sambamba* (version 0.6.6) (Tarasov et al., 2015). We used the *GATK* (version 3.8) *RealignerTargetCreator* and *IndelRealigner* modules to realign reads around indels. The realigned BAM files served as input for *GATK* base quality score recalibration using 102,092,638 unique positions from the Illumina BovineHD SNP chip and Bovine dbSNP version 150, as known variants. The *mosdepth* software (version 0.2.2) (Pedersen and Quinlan, 2018) was used to extract the number of reads that covered a genomic position.

Sequence variant discovery We followed the best practice guidelines recommended for variant discovery and genotyping using *GATK* (version 4.0.6) with default parameters for all commands (McKenna et al., 2010; Vander Jagt et al., 2018; DePristo et al., 2011). First, genotype likelihoods were calculated separately for each sequenced animal using *GATK HaplotypeCaller* (Vander Jagt et al., 2018), which resulted in files in *gVCF* (genomic Variant Call Format) format for each sample (Danecek et al., 2011). The gVCF files from the 49 samples were consolidated using *GATK* GenomicsDBImport. Subsequently, *GATK GenotypeGVCFs* was applied to genotype polymorphic sequence variants for all samples simultaneously.

Graphtyper (version 1.3) was run in a multi-sample mode as recommended in Eggertsson et al. (Eggertsson et al., 2017). Because the original implementation of *Graphtyper* is limited to the analysis of the human chromosome complement, we cloned the *Graphtyper GitHub* repository (`https://github.com/DecodeGenetics/graphtyper`), modified the source code to allow analysis of the cattle chromosome complement, and compiled the program from the modified source code (see Additional file 2.1). The *Graphtyper* workflow consisted of four steps that were executed successively. First, sequence variants were identified from the read alignments that were produced using *BWA mem* (see above). Second, these cohort-specific variants

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

Our implementation of *SAMtools mpileup* (version 1.8) (Li, 2011) was run in a multi-sample mode to calculate genotype likelihoods from the aligned reads for all samples simultaneously. The parameters -E and -t were used to recalculate (and apply) base alignment quality and produce per-sample genotype annotations, respectively. Next, the estimated genotype likelihoods were converted into genotypes using *BCFtools call* using the -v and -m flags to output variable sites only, and permit sites to have more than two alternative alleles, respectively.

We implemented all pipelines using Snakemake (version 5.2.0) (Köster and Rahmann, 2012). The scripts for the pipelines are available via *Github* repository
https://github.com/danangcrysnanto/Graph-genotyping-paper-pipelines

Sequence variant filtering and genotype refinement The *GATK VariantFiltration* module was used to parse and filter the raw VCF files. Quality control on the raw sequencing variants and genotypes was applied according to guidelines that were recommended for each variant caller. Variants that were identified using *GATK* were retained if they met the following criteria: QualByDepth (QD) > 2.0, FisherStrand > 60.0, RMSMappingQuality (MQ) > 40.0, MappingQualityRankSumTest (MQRankSum) > 12.5, ReadPosRankSumTest (ReadPosRankSum) > -8.0, SOR < 3.0 (SNPs) and QD > 2.0, FS < 200.0, ReadPosRankSum > 20.0, SOR < 10.0 (indels). For the variants identified using *SAMtools*, the thresholds that have been

applied in the 1000 Bull Genomes project (Daetwyler et al., 2014) were considered to remove variants with indication of low quality. Variants were retained if they met the following criteria: QUAL $>$ 20, MQ $>$ 30, ReadDepth (DP) $>$ 10, DP $<$ median(DP) $+ 3 *$ mean(DP). Moreover, SNPs were removed from the data if they had the same positions as the starting position of an indel. The output of *Graphtyper* was filtered so that it included only variants that met criteria recommeded by Eggertsson et al. (Eggertsson et al., 2017): ABHet < 0.0 | ABHet > 0.33, ABHom $<$ 0.0 | ABHom > 0.97, MaxAASR > 0.4, and MQ > 30.

We used *Beagle* (version 4.1) (Browning and Browning, 2016) to improve the raw sequence variant genotype quality and impute missing genotypes. The genotype likelihood (*gl*) mode of *Beagle* was applied to infer missing and modify existing genotypes based on the phred-scaled likelihoods (*PL*) of all other non-missing genotypes of the 49 Original Braunvieh animals in our study.

Evaluation of sequence variant genotyping To ensure consistent variant representation across the different sequence variant genotyping methods evaluated, we applied the *vt normalize* software (version 0.5) (Tan et al., 2015). Normalized variants are parsimonious (i.e., represented by as few nucleotides as possible) and left aligned (Tan et al., 2015). The number of variants detected and transition to transversion (Ti/Tv) ratios were calculated using *vt peek* (Tan et al., 2015) and *BCFtools stats* (Li, 2011). The intersection of variants that were common to the evaluated tools was calculated and visualized using *BCFtools isec* (Li, 2011) and the UpSet R package (Conway et al., 2017), respectively.

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

All 49 sequenced cattle were also genotyped using either the Illumina BovineHD

(N = 29) or the BovineSNP50 (N = 20) Bead chip that comprise 777,962 and 54,001 SNPs, respectively. The average genotyping rate at autosomal SNPs was 98.91%. In order to assess the quality of sequence variant genotyping, the genotypes detected by the different variant calling methods were compared to the array-called genotypes in terms of genotype concordance, non-reference sensitivity and non-reference discrepancy (DePristo et al., 2011; Linderman et al., 2014), and for more details on the metrics (see Additional file 2.2). Non-parametric Kruskal–Wallis tests followed by pairwise Wilcoxon signed-rank tests were applied to determine if any of the three metrics differed significantly between the three tools evaluated.

Computing environment and statistical analysis All computations were performed on the ETH Zurich Leonhard Open Cluster with access to multiple nodes equipped with 18 cores Intel Xeon E5-2697v4 processors (base frequency rated at 2.3 GHz) and 128 GB of random-access memory. Unless otherwise stated, the R (version 3.3.3) software environment (R Core, 2013) was used for statistical analyses and ggplot2 (version 3.0.0) (Wickham, 2016) was used for data visualisation.

1.3 Results

Following quality control (removal of adapter sequences and low-quality bases), we aligned more than 13 billion paired-end reads (2 × 125 and 2 × 150 bp) from 49 Original Braunvieh cattle to the UMD3.1 assembly of the bovine genome. On average, 98.44% (91.06–99.59%) of the reads mapped to the reference genome and 4.26% (2.0–10.91%) of these were flagged as duplicates and not considered for further analyses. Sequencing depth ranged from 6.00 to 37.78 with an average depth per animal of 12.75 and was above 12-fold for 31 samples. Although the realignment of sequencing reads around indels is no longer required when sequence variants are genotyped using the latest version of *GATK* (v 4), it is still recommended to im-

prove the genotyping of indels by using *SAMtools*. To ensure a fair comparison of the three tools evaluated, we realigned the reads around indels on all BAM files and used the re-aligned files as a starting point for our comparisons (Fig. 1.1). The sequencing read data of 49 cattle were deposited at European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena) under primary accession PRJEB28191.

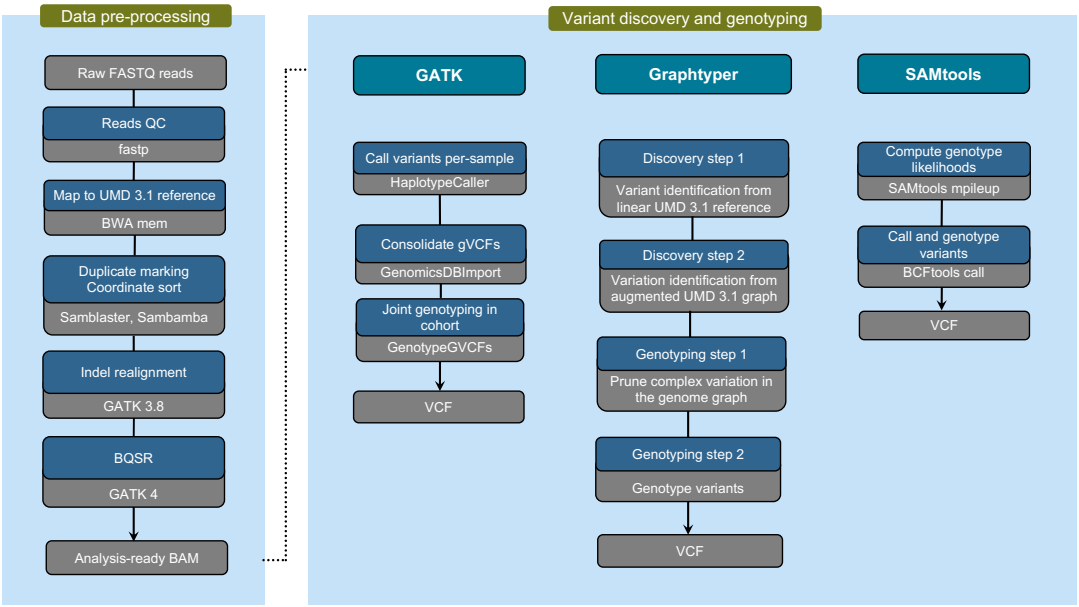


Figure 1.1: **Schematic representation of the three sequence variant discovery and genotyping methods evaluated.**
According to the best practice recommendations for sequence variant discovery using *GATK*, the VQSR module should be applied to distinguish between true and false positive variants. Because this approach requires a truth set of variants, which is not (publicly) available for cattle, the VQSR module was not considered in our evaluation

Sequence variant discovery and genotyping

Polymorphic sites (SNPs, indels) were discovered and genotyped in the 49 animals using either *GATK* (version 4), *Graphtyper* (version 1.3) or *SAMtools* (version 1.8). All software programs were run using default parameters and workflow descriptions for variant discovery (Fig. 1.1 and also see Methods). Only autosomal sequence

variants were considered to evaluate the accuracy and sen-sitivity of sequence variant genotyping. Because variant filtering has a strong impact on the accuracy and sensitivity of sequence variant genotyping (Carson et al., 2014; Jun et al., 2015), we evaluated both the raw variants that were detected using default parameters for variant discovery (Fig. 1.1) and variants that remained after applying filtering criteria that are commonly used but may differ slightly between different software tools. Note that *GATK* was run by using the suggested filtering parameters, when application of Variant Quality Score Recalibration (VQSR) is not possible.

Using default parameters for variant discovery for each of the software programs evaluated, 21,140,196, 20,262,913, and 20,668,459 polymorphic sites were discovered using *GATK*, *Graphtyper* and *SAMtools*, respectively (Table 1.1). The vast majority (86.79, 89.42 and 85.11%) of the detected variants were biallelic SNPs. Of the 18,594,182, 18,120,724 and 17,592,038 SNPs detected using *GATK*, *Graphtyper* and *SAMtools*, respectively, 7.46, 8.31 and 5.02% were novel, i.e., they were not among the 102,091,847 polymorphic sites of the most recent version (150) of the Bovine dbSNP database. The Ti/Tv ratio of the detected SNPs was equal to 2.09, 2.07 and 2.05 using *GATK*, *Graphtyper* and *SAMtools*, respectively. Using *GATK* revealed four times more multiallelic SNPs (246,220) than either *SAMtools* or *Graphtyper*.

Table 1.1: **Number of different types of autosomal sequence variants** detected in 49 Original Braunvieh cattle using three sequence variant genotyping methods (Full) and subsequent variant filtration based on commonly used criteria (Filtered)

	Full			Filtered		
	GATK	Graphtyper	SAMtools	GATK	Graphtyper	SAMtools
Variants	21,140,196	20,262,913	20,668,459	19,761,679	17,679,155	18,871,549
SNPs	18,594,182	18,120,724	17,592,038	17,248,593	15,777,446	16,272,917
Not in dbSNP	1,387,781	1,505,586	882,575	867,838	564,326	570,901
Biallelic	18,347,962	18,053,396	17,528,249	17,111,806	15,730,153	16,218,714
Multi-allelic	246,220	67,328	63,789	136,787	47,293	54,203
Ti/Tv ratio	2.09	2.07	2.05	2.17	2.18	2.16
SNP array (%)						
BovineHD	99.46	99.61	99.32	99.21	98.79	98.85
Bovine SNP50	99.14	99.26	99.12	98.91	98.87	98.9
Indels	2,478,489	2,044,585	3,076,421	2,445,766	1,826,808	2,598,632
Not in dbSNP	663,831	596,137	1,279,162	639,219	456,752	979,291
Biallelic	2,166,352	1,753,391	2,704,413	2,133,840	1,571,195	2,310,386
Multi-allelic	312,137	291,194	372,008	311,926	255,613	288,246
Insertion/Deletion	0.88	0.88	1	0.88	0.88	0.99
Complex variation	67,525	97,604	0	67,320	74,901	0

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

On average, each Original Braunvieh cattle carried between 7 and 8 million variants that differed from the UMD3.1 reference genome. Of these, between 2.4 and 2.6 million SNPs were homozygous for the alternate allele, between 3.8 and 4.7 million SNPs were heterozygous and between 0.7 and 1 million were indels (Table 1.2). An intersection of 15,901,526 biallelic SNPs was common to all sequence-variant discovery tools evaluated Fig 1.2a, i.e., between 85.51 and 90.39% of the detected SNPs of each tool, and 466,029 (2.93%, Ti/Tv: 1.81) of them were novel, i.e., they were not present in dbSNP 150. The Ti/Tv-ratio of the common SNPs was 2.22. *SAMtools* had the largest number of SNPs in common with the other two tools (90.39%). The number of private SNPs, i.e., SNPs that were detected by one but not the other tools was largest for *GATK* and smallest for *Graphtyper*.

In total, 1,299,467 biallelic indels Fig 1.2b were common to all evaluated tools and 98,931 (13.13%) of these were novel, i.e., they were not present in dbSNP 150. The intersection among the three tools was considerably smaller for indels than for SNPs. *Graphtyper* had the highest proportion of indels in common with the other tools (74.11%). *SAMtools* discovered the largest number (2,704,413) of biallelic indels and most of them (41.26%) were not detected using either *GATK* or *Graphtyper*. *GATK* (21.2%) and *Graphtyper* (12.38%) discovered fewer private indels than *SAMtools*.

Table 1.2: **Average number of autosomal variants** identified per animal using three sequence variant genotyping methods

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

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

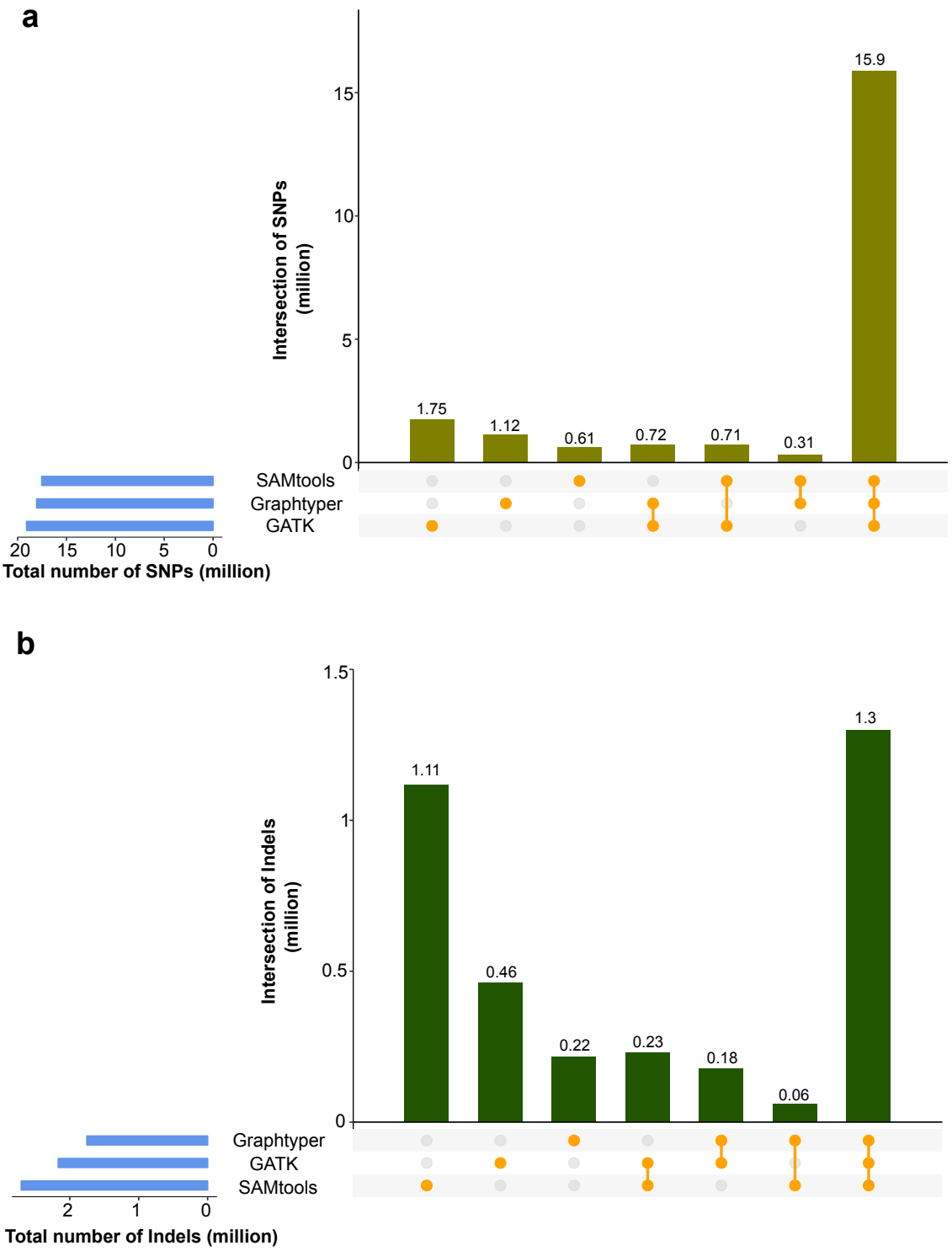


Figure 1.2: Number of biallelic SNPs (**a**) and indels (**b**) identified in 49 Original Braunvieh cattle using three sequence variant genotyping methods. Blue horizontal bars represent the total number of sites discovered for each method. Vertical bars indicate private and common variants detected by the methods evaluated

Sequence variant genotyping using *Graphtyper* is accurate

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

In the raw data, the proportion of missing non-reference sites was 1.90%, 0.56%, and 0.47% using *GATK*, *Graphtyper*, and *SAMtools*, respectively. The genotype concordance between the sequence- and microarray-derived genotypes was higher ($P < 0.005$) when *Graphtyper* (97.72%) was used than when either *SAMtools* (97.68%) or *GATK* (95.99%) was used (Table 1.3). For the three tools evaluated, the genotype concordance was higher at homozygous than heterozygous sites, particularly in animals that were sequenced at low depth (see Additional file 2.3). In order to take the variable proportions of missing genotypes in the sequence variants into account, we calculated non-reference sensitivity and non-reference discrepancy. Non-reference sensitivity was almost identical using *Graphtyper* (98.26%) and *SAMtools* (98.21%). However, non-reference sensitivity was clearly lower using *GATK* (93.81%, $P < 0.001$). Non-reference discrepancy was lower using *Graphtyper* (3.53%) than using either *SAMtools* (3.6%, $P = 0.003$) or *GATK* (6.35%, $P < 0.001$).

Table 1.3: **Comparisons between array-called and sequence variant genotypes.**

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

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

Next, we analysed the proportion of opposing homozygous genotypes for SNPs and indels in nine sire-son pairs that were included among the sequenced animals (Table 1.4). We observed that SNPs that were discovered using either *Graphtyper* or *SAMtools* had almost a similar proportion of genotypes with Mendelian inconsistencies in the full and filtered datasets, whereas the values were two times higher using *GATK*. The proportion of opposing homozygous genotypes was higher for indels than SNPs for all the tools evaluated. However, in the full and filtered datasets, it was lower when *Graphtyper* was used than when either *GATK* or *SAMtools* was used. Using filtering parameters that are commonly applied for the three evaluated tools (see Methods), we excluded 1,378,517 (6.52%, Ti/Tv 1.24), 2,583,758 (12.75%, Ti/Tv 1.47) and 1,796,910 (8.69%, Ti/Tv 1.36) variants due to low mapping or genotyping quality from the *GATK*, *Graphtyper*, and *SAMtools* datasets, respectively. The genotype concordance between sequence- and microarray-derived genotypes was slightly higher for the filtered than the raw genotypes, but the non-reference sensitivity was lower for the filtered than the raw genotypes, which indicates that the filtering step also removed some true variant sites from the raw data (Table 1.3). The filtering step had almost no effect on the proportion of Mendelian inconsistencies detected in the nine sire-son pairs (Table 1.4).

Beagle genotype refinement improved genotype quality

We used the *Beagle* software to refine the primary genotype calls and infer missing genotypes in the raw and filtered datasets. Following imputation, the quality of the sequence variant genotypes increased for all evaluated tools particularly for the individuals that had a sequencing coverage less than 12-fold (Fig. 1.3). The largest increase in the concordance metrics was observed for the sequence variants that were obtained using *GATK* (Tables 1.3 and 1.4). Following imputation, the variants identified using *Graphtyper* had a significantly higher quality ($P < 0.05$) for eight out of the ten metrics evaluated.

Table 1.4: **Proportions of opposing homozygous genotypes observed in nine sire-son pairs**

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

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

The quality of the sequence variant genotypes, particularly before Beagle genotype phasing and imputation, was influenced by the variable depth of coverage for the 49 sequenced samples of our study (Fig. 1.3). When we restricted the evaluations to 31 samples that had an average sequencing depth above 12-fold, the three tools performed almost identically (see Additional file 2.4). However, the performance of *Graphtyper* was significantly ($P < 0.05$) higher for 12 (out of the total 20) metrics than either that of *GATK* or *SAMtools*. When 18 samples with an average sequencing depth lower than 12-fold were considered, the differences observed in the three metrics were more pronounced between the three tools. In samples with a low sequencing coverage, *Graphtyper* performed significantly ($P < 0.05$) better than either *GATK* or *SAMtools* for all concordance metrics both before and after filtering and *Beagle* imputation, except for the non-reference sensitivity.

Computing requirements

The multi-sample sequence variant genotyping pipelines that were implemented using either *GATK* or *SAMtools* were run separately for each chromosome in a single-threading mode. The *SAMtools mpileup* module took between 3.07 and 11.4

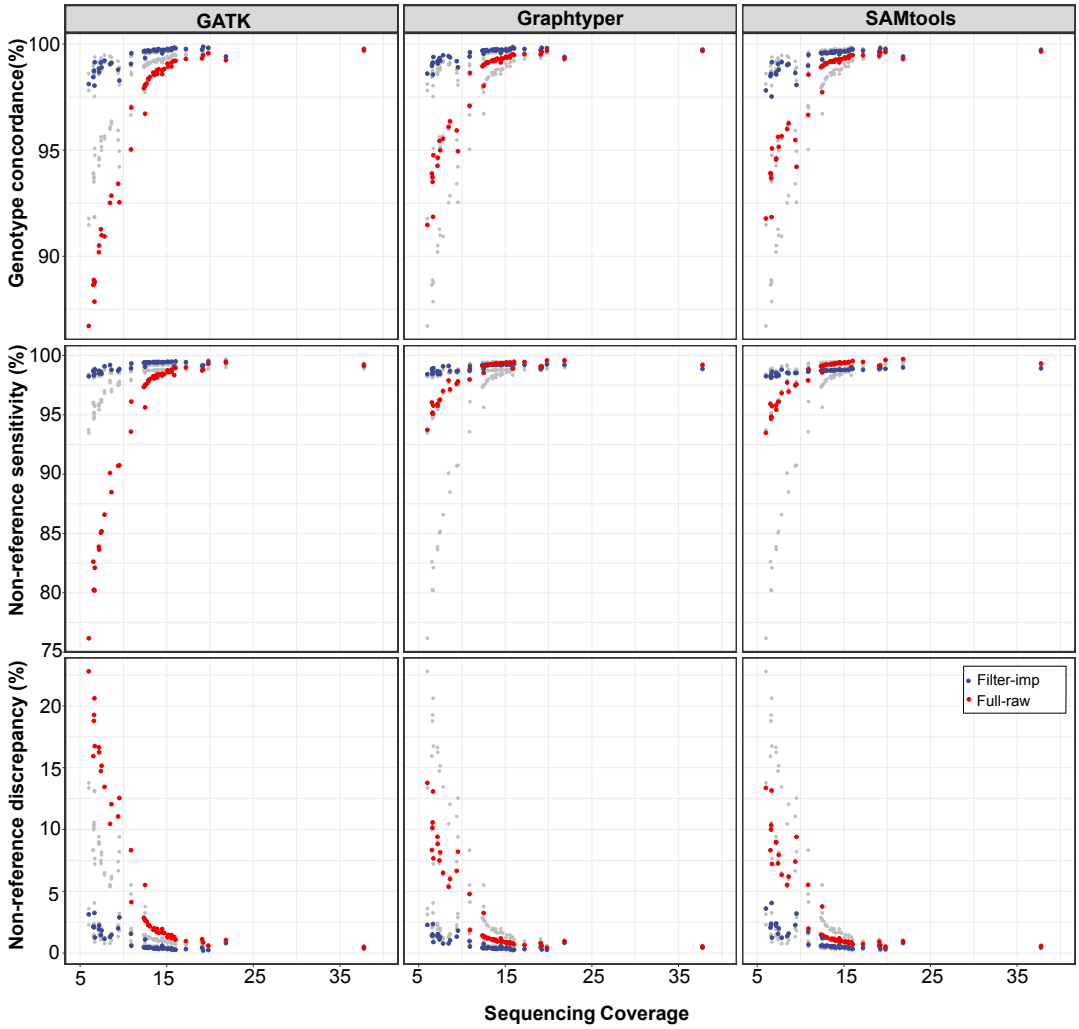


Figure 1.3: **Accuracy and sensitivity of sequence variant genotyping at different sequencing depths.** Genotype concordance, non-reference sensitivity and non-reference discrepancy were calculated for 49 Original Braunvieh cattle considering either raw (red) or filtered and imputed (blue) sequence variant genotypes. The grey points represent overlays of the results from the other methods

CPU hours and required between 0.12 and 0.25 gigabytes (GB) peak random-access memory (RAM) per chromosome. To genotype 20,668,459 sequence variants in 49 animals, *SAMtools* mpileup required 192 CPU hours (Fig. 1.4).

For *GATK*, we submitted 1421 parallel jobs of the *HaplotypeCaller* module (i.e., one job for each animal and chromosome) that required between 3.9 and 12.3 GB RAM and between 0.36 and 11 CPU hours to complete. To process 29 chromosomes

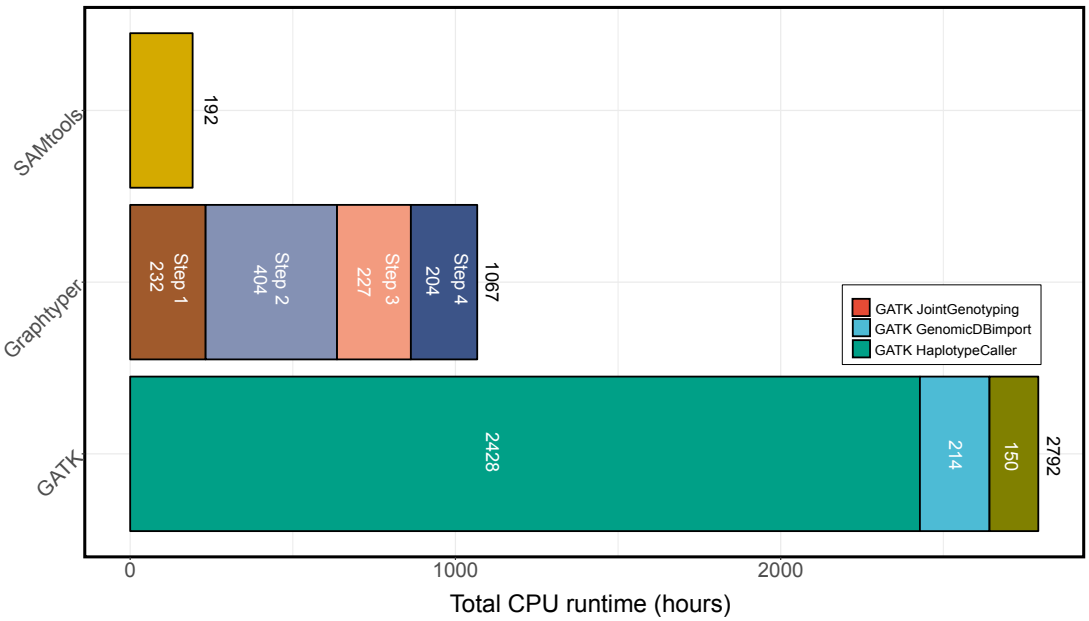


Figure 1.4: **Computing time required to genotype all autosomal sequence variants in 49 Original Braunvieh cattle.** The runtime of *GATK* and *Graphtyper* is shown for the different steps (see Fig. 1.1 for more details)

in 49 samples, the *HaploytpeCaller* module required 2428 CPU hours. Subsequently, we ran the *GATK GenomicsDBImport* module, which required between 7.98 and 20.88 GB RAM and between 2.81 and 19.31 CPU hours per chromosome. *GATK Joint Genotyping* required between 4.33 and 17.32 GB of RAM and between 1.81 and 14.01 h per chromosome. To genotype 21,140,196 polymorphic sequence variants in 49 animals, the *GATK* pipeline required 2792 CPU hours (Fig. 1.4).

The *Graphtyper* pipeline including construction of the variation graph and genotyping of sequence variants was run in parallel for 2538 non-overlapping segments of 1 million bp as recommended by (Eggertsson et al., 2017). The peak RAM required by *Graphtyper* was between 1 and 3 GB per segment. Twelve segments, for which *Graphtyper* either ran out of memory or did not finish within the allocated time, were subdivided into smaller segments of 10 kb and subsequently re-run (Additional file 2.5). The genotyping of 20,262,913 polymorphic sites in 49 animals using our implementation of the *Graphtyper* pipeline required 1066 CPU hours (Fig. 1.4).

The computing resources required by *SAMtools* and *GATK* increased linearly with chromosome length. The computing time required to genotype sequence variants was highly heterogeneous along the genome using *Graphtyper*. The CPU time for a 1-Mb segment ranged from 0.196 to 10.11 h, with an average CPU time of 0.42 h. We suspected that flaws in the reference genome might increase the complexity of the variation-aware graph and that the construction of the graph might benefit from an improved assembly. To test this hypothesis, we re-mapped the sequencing reads to the recently released new bovine reference genome (ARS-UCD1.2, https://www.ncbi.nlm.nih.gov/assembly/GCF_002263795.1) and repeated the graph-based sequence variant discovery. Indeed, we did observe a decrease in the computing time required to genotype polymorphic sites (particularly at chromosomes 12, 27 and 29) and a more uniform runtime along the genome, which possibly indicates that graph-based variant discovery in cattle will be faster and more accurate with highly contiguous reference sequences (Fig. 1.5).

1.4 Discussion

We used either *GATK*, *Graphtyper*, or *SAMtools* to discover and genotype polymorphic sequence variants in whole-genome sequencing data of 49 Original Braunvieh cattle that were sequenced at between 6 and 38-fold genome coverage. Whereas *SAMtools* and *GATK* discover variants from a linear reference genome, *Graphtyper* locally realigns reads to a variation-aware reference graph that incorporates cohort-specific sequence variants (Eggertsson et al., 2017). Our graph-based variant discovery pipeline that is implemented by using the *Graphtyper* software used the existing bovine reference sequence to construct the genome graph. Subsequently, the graph was augmented with variants that were detected from linear alignments of the 49 Original Braunvieh cattle. The use of more sophisticated genome graph-based approaches that have been developed very recently facilitates the mapping of raw se-

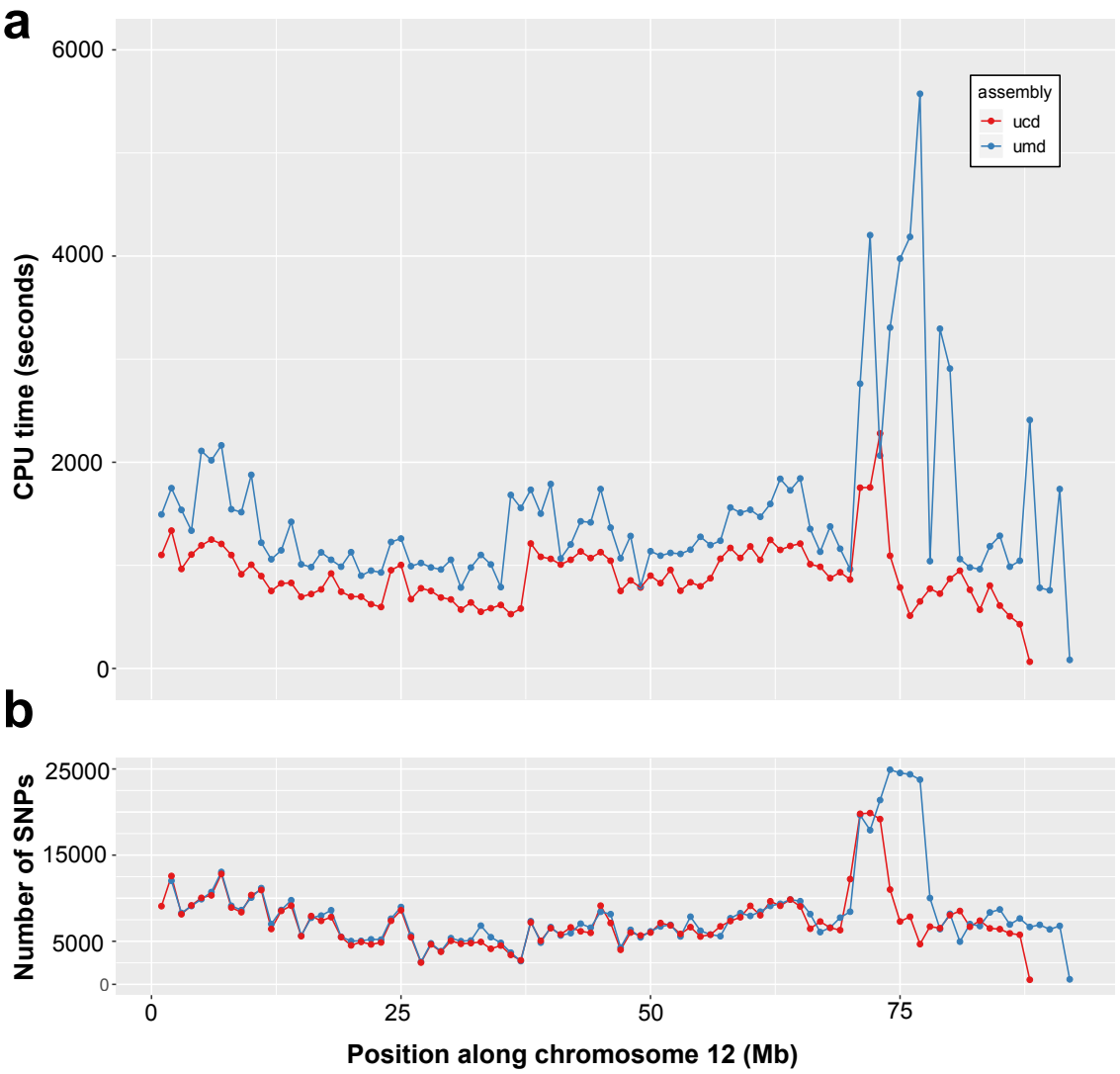


Figure 1.5: **Sequence variant genotyping on chromosome 12 using *Graphtyper*.** Computing time required (**a**) and number of variants discovered (**b**) for bovine chromosome 12 using *Graphtyper*. Each dot represents an interval of 1 million bp. Blue and red colours represent values for the UMD3.1 and ARS-UCD1.2 versions of the bovine assembly, respectively

quencing reads directly against a genome graph without the need to first align reads towards a linear reference genome (Garrison et al., 2018). Whereas genome graph-based variant discovery has been explored recently in mammalian-sized genomes (Dilthey et al., 2015; Rakocevic et al., 2019; Garrison et al., 2018; Sibbesen et al., 2018), our work is the first to apply graph-based sequence variant genotyping in cattle.

In order to evaluate graph-based variant discovery in cattle, we compared accuracy and sensitivity of *Graphtyper* to *GATK*, and *SAMtools* , i.e., two state-of-the-art methods on linear reference genomes that have been evaluated thoroughly in many species including cattle (Jansen et al., 2013; Baes et al., 2014). We ran each tool with default parameters for variant discovery and applied commonly used or recommended filtration criteria. However, our evaluation of the software tools may suffer from ascertainment bias because we relied on SNPs that are included in bovine SNP arrays, i.e., they are located predominantly at genomic regions where variants are easy to identify (Li, 2014; Malomane et al., 2018; Linderman et al., 2014). Thus, the global accuracy and sensitivity of sequence variant discovery might be overestimated in our study. However, this ascertainment bias is unlikely to affect the relative performance of the methods evaluated.

In 49 Original Braunvieh cattle, sequence variant genotyping was more accurate using *Graphtyper* than either *GATK* or *SAMtools*. Differences in accuracy are small between the three tools, particularly when samples are sequenced at an average coverage higher than 12-fold (see Additional file 4). Yet, *Graphtyper* performed significantly better than *GATK* and *SAMtools* for samples sequenced at medium ($>$ 12-fold) or low($<$ 12-fold) coverage indicating that genome graph-based variant discovery in cattle is accurate across a wide range of sequencing depths. *GATK* might perform better than observed in our study, when the VQSR module is applied to train the variant filtration algorithm on true and false variants (Pirooznia et al., 2014). However, to the best of our knowledge, the required sets of true and false variants are not available in cattle. An intersection of variants detected by different sequence variant genotyping software may be considered as a truth set (e.g., (Alberto et al., 2018)) and compiling such a set is possible using the 49 samples from our study. However, a truth set that has been constructed from the data that are used for evaluation is likely to be depleted for variants that are difficult to discover in the target data set, thus preventing an unbiased evaluation of variant calling (Li

et al., 2018). Variants from the 1000 Bull Genomes project (Daetwyler et al., 2014; Hayes and Daetwyler, 2019) could potentially serve as a truth/training set. However, variants from the 1000 Bull Genomes project were detected from short read sequencing data using either *GATK* or *SAMtools*, i.e., technologies and software that are part of our evaluation, thus precluding an unbiased comparison of variant discovery between*GATK*, *Graphtyper*, and *SAMtools* (Li et al., 2018). (Vander Jagt et al., 2018) showed in a subset of samples from the 1000 Bull Genomes project that *GATK VQSR* does not notably improve the concordance between sequence-derived and microarray-called genotypes compared to *GATK* hard filtering. Interestingly, the proportion of opposing homozygous genotypes in sire/offspring pairs was slightly higher in their study using *GATK VQSR* than *GATK* hard-filtering as used by the 1000 Bull Genomes project (Vander Jagt et al., 2018). Applying *GATK* VQSR to the variants of our dataset corroborates the findings of (Vander Jagt et al., 2018) (see Additional file 6). Considering that the quality of the truth/training sets has a strong impact on the capabilities of VQSR (Additional file 2.6) and that high-confidence variants are currently not publicly available for cattle, we report *GATK* results using the recommended filtering parameters when VQSR is not possible.

Regardless of the method evaluated, we observed heterozygous under-calling in animals that are sequenced at low coverage, i.e., heterozygous variants were erroneously genotyped as homozygous due to an insufficient number of sequencing reads supporting the heterozygous genotype (Nielsen et al., 2011; Sims et al., 2014; Fragoso et al., 2016; Bilton et al., 2018). In agreement with previous studies (Jansen et al., 2013; Daetwyler et al., 2014), *Beagle* imputation improved genotype concordance and reduced heterozygous under-calling particularly in individuals that are sequenced at low coverage. After the imputation step, the genotype concordance, non-reference sensitivity, and non-reference discrepancy of the three tools were almost identical, which indicates that genotyping sequence variants from samples with a medium genome coverage is possible at high accuracy (at least for common vari-

ants in more accessible regions of the genome) using any of the three tools evaluated and subsequent Beagle error correction. While such conclusions have been drawn previously for *SAMtools* and *GATK* (Jansen et al., 2013; Baes et al., 2014), our findings demonstrate that the genotype likelihoods estimated from the *Graphtyper* software are also compatible with and benefit from the imputation algorithm implemented in the Beagle software. Considering that sequence data are enriched for rare variants that are more difficult to impute than common variants from SNP microarrays (Pausch et al., 2017b), the benefits from Beagle error correction might be overestimated in our study. An integration of phasing and imputation of missing genotypes directly in a graph-based variant genotyping approach would simplify sequence variant genotyping from variation-aware graphs (Rakocevic et al., 2019; Sirén et al., 2020; Novak et al., 2017a). Using *Graphtyper* for variant genotyping and Beagle for genotype refinement enabled us to genotype sequence variants in 49 Original Braunvieh cattle at a genotypic concordance of 99.52%, i.e., higher than previously achieved using either *GATK* or *SAMtools* for variant calling in cattle that are sequenced at a similar genome coverage (Jansen et al., 2013; Stothard et al., 2015; Boussaha et al., 2016; Daetwyler et al., 2014; Baes et al., 2014; Stafuzza et al., 2017); this indicates that graph-based variant discovery might improve sequence variant genotyping. However, applying the filtering criteria that are recommended for *Graphtyper* (Eggertsson et al., 2017) removed more variants from the *Graphtyper* (12.75%) than from either *GATK* (6.52%) or *SAMtools* (8.69%) datasets. It should be mentioned that *GATK* VQSR would remove considerably more variants from the *GATK* dataset than *GATK* hard filtering as applied in our study (see Additional file 6). Fine-tuning of the variant filtering parameters is necessary to further increase the accuracy and sensitivity of sequencing variant genotyping, particularly for *Graphtyper* (Carson et al., 2014; Jun et al., 2015). Moreover, the accuracy and sensitivity of graph-based variant discovery may be higher when known variants are considered for the initial construction of the variation graph (Eggertsson et al., 2017).

Indeed, we observed a slight increase in genotype concordance (see Additional file 2.7) when we used *Graphtyper* to genotype sequence variants from a variation-aware genome-graph that incorporated bovine variants listed in dbSNP 150. However, additional research is required to prioritize a set of variants to augment bovine genome graphs for different cattle breeds (Pritt et al., 2018).

Using microarray-derived genotypes as a truth set may overestimate the accuracy of sequence variant discovery particularly at variants that are rare or located in less accessible regions of the genome. Moreover, it does not allow assessment of the accuracy and sensitivity of indel discovery because variants other than SNPs are currently not routinely genotyped with commercially available microarrays. Estimating the proportion of opposing homozygous genotypes between parent–offspring pairs may be a useful diagnostic metric to detect sequencing artefacts or flawed genotypes at indels (Patel et al., 2014). Our results show that genotypes at indels are more accurate using *Graphtyper* than either *SAMtools* or *GATK* because *Graphtyper* produced less opposing homozygous genotypes at indels in nine sire-son pairs than the other methods both in the raw and filtered datasets. These findings are in line with those reported by (Eggertsson et al., 2017), who showed that the mapping of the sequencing reads to a variation-aware graph could improve read alignment nearby indels, thus enabling highly accurate sequence variant genotyping also for variants other than SNPs. Recently, (Garrison et al., 2018) showed that graph-based variant discovery may also mitigate reference allele bias. An assessment of reference allele bias was, however, not possible in our study because the sequencing depth was too low for most samples.

In our study, *Graphtyper* required less computing time than *GATK* to genotype sequence variants for 49 individuals.*SAMtools* required the least computing resources, probably because the implemented mpileup algorithm produces genotypes from the aligned reads without performing the computationally intensive local realignment of the reads. However, with an increasing number of samples, the multi-

sample variant genotyping implementation of the *GATK HaplotypeCaller* module seems to be more efficient than *SAMtools mpileup* because variant discovery within samples can be separated from the joint genotyping across samples (Poplin et al., 2018; Vander Jagt et al., 2018). A highly parallelized graph-based variant discovery pipeline also offers a computationally feasible and scalable framework for variant discovery in thousands of samples (Eggertsson et al., 2017). However, the computing time necessary for graph-based variant genotyping might be high in genomic regions where the nucleotide diversity is high or the assembly is flawed (Sibbesen et al., 2018; Koren et al., 2013). In our study, the algorithm implemented in the *Graphtyper* software failed to finish within the allocated time for 12 1-Mb segments including a segment on chromosome 12 that contains a large segmental duplication (Pausch et al., 2017b; Liu et al., 2009; Bickhart et al., 2012) possibly because many mis-mapped reads increased graph complexity. The region on chromosome 12 contains an unusually large number of sequence variants and has been shown to suffer from low accuracy of imputation (Pausch et al., 2017b). *Graphtyper* also failed to finish within the allocated time for a region on chromosome 23 that encompasses the bovine major histocompatibility complex, which is known to have a high level of diversity. Our results show that *Graphtyper* may also produce genotypes for problematic segments when they are split and processed in smaller parts. Moreover, most of these problems disappeared when we considered the latest assembly of the bovine genome, which possibly corroborates that more complete and contiguous genome assemblies may facilitate more reliable genotyping from variation-aware graphs (Li, 2014; Guo et al., 2017).

1.5 Conclusions

Genome graphs facilitate sequence variant discovery from non-linear reference genomes. Sequence variant genotyping from a variation-aware graph is possible in cattle using

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

References

F. J. Alberto, F. Boyer, P. Orozco-terWengel, I. Streeter, B. Servin, P. De Villemereuil, B. Benjelloun, P. Librado, F. Biscarini, L. Colli, et al. Convergent genomic signatures of domestication in sheep and goats. *Nature Communications*, 9(1):1–9, 2018.

C. F. Baes, M. A. Dolezal, J. E. Koltes, B. Bapst, E. Fritz-Waters, S. Jansen, C. Flury, H. Signer-Hasler, C. Stricker, R. Fernando, et al. Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. *BMC genomics*, 15(1):948, 2014.

D. M. Bickhart, Y. Hou, S. G. Schroeder, C. Alkan, M. F. Cardone, L. K. Matukumalli, J. Song, R. D. Schnabel, M. Ventura, J. F. Taylor, et al. Copy number variation of individual cattle genomes using next-generation sequencing. *Genome research*, 22(4):778–790, 2012.

T. P. Bilton, J. C. McEwan, S. M. Clarke, R. Brauning, T. C. van Stijn, S. J. Rowe, and K. G. Dodds. Linkage disequilibrium estimation in low coverage high-throughput sequencing data. *Genetics*, 209(2):389–400, 2018.

M. Boussaha, P. Michot, R. Letaief, C. Hozé, S. Fritz, C. Grohs, D. Esquerré, A. Duchesne, R. Philippe, V. Blanquet, F. Phocas, S. Floriot, D. Rocha, C. Klopp, A. Capitan, and D. Boichard. Construction of a large collection of small genome variations in French dairy and beef breeds using whole-genome sequences. *Genetics Selection Evolution*, 48(1):87, dec 2016. ISSN 1297-9686. doi: 10.1186/s12711-016-0268-z.

A. C. Bouwman, H. D. Daetwyler, A. J. Chamberlain, C. H. Ponce, M. Sargolzaei, F. S. Schenkel, G. Sahana, A. Govignon-Gion, S. Boitard, M. Dolezal, H. Pausch, R. F. Brøndum, P. J. Bowman, B. Thomsen, B. Guldbrandtsen, M. S. Lund, B. Servin, D. J. Garrick, J. Reecy, J. Vilkki, A. Bagnato, M. Wang, J. L. Hoff, R. D. Schnabel, J. F. Taylor, A. A. Vinkhuyzen, F. Panitz, C. Bendixen, L. E. Holm, B. Gredler, C. Hozé, M. Boussaha, M. P. Sanchez, D. Rocha, A. Capitan, T. Tribout, A. Barbat, P. Croiseau, C. Drögemüller, V. Jagannathan, C. Vander Jagt, J. J. Crowley, A. Bieber, D. C. Purfield, D. P. Berry, R. Emmerling, K. U. Götz, M. Frischknecht, I. Russ, J. Sölkner, C. P. Van Tassell, R. Fries, P. Stothard, R. F. Veerkamp, D. Boichard, M. E. Goddard, and B. J. Hayes. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nature Genetics*, 50(3):362–367, feb 2018. ISSN 15461718. doi: 10.1038/s41588-018-0056-5.

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

A. R. Carson, E. N. Smith, H. Matsui, S. K. Brækkan, K. Jepsen, J.-B. Hansen, and K. A. Frazer. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC bioinformatics*, 15(1):125, 2014.

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

A. Y. Cheng, Y.-Y. Teo, and R. T.-H. Ong. Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics*, 30(12):1707–1713, 2014.

J. R. Conway, A. Lex, and N. Gehlenborg. Upsetr: an r package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18):2938–2940, 2017.

H. D. Daetwyler, A. Capitan, H. Pausch, P. Stothard, R. Van Binsbergen, R. F. Brøndum, X. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerré, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. Vantassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genetics*, 46(8):858–865, aug 2014. ISSN 15461718. doi: 10.1038/ng.3034.

P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

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.

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

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

C. A. Fragoso, C. Heffelfinger, H. Zhao, and S. L. Dellaporta. Imputing genotypes in biallelic populations from low-coverage sequence data. *Genetics*, 202(2):487–495, 2016.

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.

S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333, 2016.

Y. Guo, F. Ye, Q. Sheng, T. Clark, and D. C. Samuels. Three-stage quality control strategies for dna re-sequencing data. *Briefings in bioinformatics*, 15(6):879–889, 2014.

Y. Guo, Y. Dai, H. Yu, S. Zhao, D. C. Samuels, and Y. Shyr. Improvements and impacts of grch38 human reference on high throughput sequencing data analysis. *Genomics*, 109(2):83–90, 2017.

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(1):annurev–animal–020518–115024, feb 2019. ISSN 2165-8102. doi: 10.1146/annurev-animal-020518-115024.

J. L. Hoff, J. E. Decker, R. D. Schnabel, and J. F. Taylor. Candidate lethal haplotypes and causal mutations in Angus cattle. *BMC Genomics*, 18(1), 2017. ISSN 14712164. doi: 10.1186/s12864-017-4196-2.

S. Jansen, B. Aigner, H. Pausch, M. Wysocki, S. Eck, A. Benet-Pagès, E. Graf, T. Wieland, T. M. Strom, T. Meitinger, and R. Fries. Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. *BMC Genomics*, 14(1):446, jul 2013. ISSN 14712164. doi: 10.1186/1471-2164-14-446.

G. Jun, M. K. Wing, G. R. Abecasis, and H. M. Kang. An efficient and scalable analysis framework for variant extraction and refinement from population-scale dna sequence data. *Genome Research*, 25(6):918–925, 2015.

S. Koren, G. P. Harhay, T. P. Smith, J. L. Bono, D. M. Harhay, S. D. Mcvey, D. Radune, N. H. Bergman, and A. M. Phillippy. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome biology*, 14(9):R101, 2013.

J. Köster and S. Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

P. Kumar, M. Al-Shafai, W. A. Al Muftah, N. Chalhoub, M. F. Elsaid, A. A. Aleem, and K. Suhre. Evaluation of snp calling using single and multiple-sample calling algorithms by validation against array base genotyping and mendelian inheritance. *BMC research notes*, 7(1):747, 2014.

B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9 (4):357, 2012.

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

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

H. Li and R. Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.

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

H. Li, J. M. Bloom, Y. Farjoun, M. Fleharty, L. Gauthier, B. Neale, and D. MacArthur. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nature methods*, 15(8): 595–597, 2018.

M. D. Linderman, T. Brandt, L. Edelmann, O. Jabado, Y. Kasai, R. Kornreich, M. Mahajan, H. Shah, A. Kasarskis, and E. E. Schadt. Analytical validation of whole exome and whole genome sequencing for clinical applications. *BMC medical genomics*, 7(1):20, 2014.

G. E. Liu, M. Ventura, A. Cellamare, L. Chen, Z. Cheng, B. Zhu, C. Li, J. Song, and E. E. Eichler. Analysis of recent segmental duplications in the bovine genome. *BMC genomics*, 10(1):571, 2009.

X. Liu, S. Han, Z. Wang, J. Gelernter, and B.-Z. Yang. Variant callers for next-generation sequencing data: a comparison study. *PloS one*, 8(9), 2013.

D. K. Malomane, C. Reimer, S. Weigend, A. Weigend, A. R. Sharifi, and H. Simianer. Efficiency of different strategies to mitigate ascertainment bias when using snp panels in diversity studies. *BMC genomics*, 19(1):22, 2018.

A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song. Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, 2011.

A. M. Novak, E. Garrison, and B. Paten. A graph extension of the positional burrows–wheeler transform and its applications. *Algorithms for Molecular Biology*, 12(1):18, 2017a.

A. M. Novak, G. Hickey, E. Garrison, S. Blum, A. Connelly, A. Dilthey, J. Eizenga, M. S. Elmohamed, S. Guthrie, A. Kahles, et al. Genome graphs. *bioRxiv*, page 101378, 2017b.

Z. H. Patel, L. C. Kottyan, S. Lazaro, M. S. Williams, D. H. Ledbetter, G. Tromp, A. Rupert, M. Kohram, M. Wagner, A. Husami, et al. The struggle to find reliable results in exome sequencing data: filtering out mendelian errors. *Frontiers in genetics*, 5:16, 2014.

H. Pausch, R. Emmerling, B. Gredler-Grandl, R. Fries, H. D. Daetwyler, and M. E. Goddard. Meta-analysis of sequence-based association studies across three cattle breeds reveals 25 QTL for fat and protein percentages in milk at nucleotide resolution. *BMC Genomics*, 18(1):853, dec 2017a. ISSN 1471-2164. doi: 10.1186/s12864-017-4263-8.

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):24, 2017b.

B. S. Pedersen and A. R. Quinlan. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*, 34(5):867–868, 2018.

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

M. Pirooznia, M. Kramer, J. Parla, F. S. Goes, J. B. Potash, W. R. McCombie, and P. P. Zandi. Validation and assessment of variant calling pipelines for next-generation sequencing. *Human genomics*, 8(1):14, 2014.

R. Poplin, V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, G. A. Van der Auwera, D. E. Kling, L. D. Gauthier, A. Levy-Moonshine, D. Roazen, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, page 201178, 2018.

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

T. R Core. R: A language and environment for statistical computing. 2013.

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.

B. Raymond, A. C. Bouwman, C. Schrooten, J. Houwing-Duistermaat, and R. F. Veerkamp. Utility of whole-genome sequence data for across-breed genomic prediction. *Genetics Selection Evolution*, 50(1):27, dec 2018. ISSN 1297-9686. doi: 10.1186/s12711-018-0396-8.

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.

D. Sims, I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121, 2014.

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

N. B. Stafuzza, A. Zerlotini, F. P. Lobo, M. E. B. Yamagishi, T. C. S. Chud, A. R. Caetano, D. P. Munari, D. J. Garrick, M. A. Machado, M. F. Martins, et al. Single nucleotide variants and indels identified from whole-genome re-sequencing of guzerat, gyr, girolando and holstein cattle breeds. *PLoS One*, 12(3), 2017.

P. Stothard, X. Liao, A. S. Arantes, M. De Pauw, C. Coros, G. S. Plastow, M. Sargolzaei, J. J. Crowley, J. A. Basarab, F. Schenkel, S. Moore, and S. P. Miller. A large and diverse collection of bovine genome sequences from the Canadian Cattle Genome Project. *GigaScience*, 2015. doi: 10.1186/s13742-015-0090-5.

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

A. Tarasov, A. J. Vilella, E. Cuppen, I. J. Nijman, and P. Prins. Sambamba: fast processing of ngs alignment formats. *Bioinformatics*, 31(12):2032–2034, 2015.

G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, et al. From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, 43(1):11–10, 2013.

C. Vander Jagt, A. Chamberlain, R. Schnabel, B. Hayes, and H. Daetwyler. Which is the best variant caller for large whole-genome sequencing datasets. In *Proceedings of the 11th world congress on genetics applied to livestock production*, pages 11–16, 2018.

H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer, 2016.

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):R42, 2009.

Chapter 2

Analysis of the cattle whole-genome graphs

Preface: Bridging text between Chapter 2 and Chapter 3

In this chapter, I constructed the first cattle whole genome graph and performed the first assessment of the gigabase genome graph on the species other than human. I showed using both real and simulated datasets that the graph facilitate accurate read mapping and unbiased sequence variant genotyping. I developed the graph pipeline further from previous implementation based on *vg toolkit* allowing graph analysis performed in a full genome scale. Additionally, I included catalogues of previously discovered variants to the graph, which showed that *breed-specific* graph perform similarly as the *multi-breed pangenome graph*.

Contribution: I and Hubert Pausch conceived the study, I wrote the full genome graph pipelines and performed all analyses. I wrote the initial draft of the manuscript with input from Hubert Pausch.

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

Danang Crysnanto*, Hubert Pausch

Animal Genomics, ETH Zurich, Zurich, Switzerland.

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

Abstract

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

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

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

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

2.1 Introduction

A reference sequence is an assembly of digital nucleotides that are representative for a species' genetic constitution. Discovery and genotyping of polymorphic sites from whole-genome sequencing data typically involve reference-guided alignment and genotyping steps that are carried out successively (DePristo et al., 2011). Variants are discovered at positions where aligned sequencing reads differ from corresponding reference nucleotides. Long-read sequencing and sophisticated genome assembly methods enabled spectacular improvements in the quality of linear reference sequences particularly for species with gigabase-sized genomes (Koren et al., 2018). Recently generated de novo assemblies exceed in quality and continuity all current reference sequences (Miga et al., 2020; Rice et al., 2020). However, modifications and amendments to existing linear reference sequences causes shifts in their coordinates that require large efforts from the genomics community to make data compatible with updated reference sequences (Ballouz et al., 2019).

Domestication and selection for beef and milk production under various environmental conditions have led to the formation of more than thousand breeds of cattle (*Bos taurus*) with distinct genetic characteristics and high allelic variation within and between breeds (Scherf et al., 2015). The 1000 Bull Genomes Project discovered almost 100 million sequence variants that are polymorphic in 2700 cattle from worldwide cattle breeds (Daetwyler et al., 2014; Hayes and Daetwyler, 2019). Nucleotide diversity is higher in cattle than human populations (Daetwyler et al., 2014; Charlier et al., 2016). Yet, all bovine DNA sequences are aligned to the linear consensus reference sequence of a highly inbred Hereford cow to facilitate reference-guided variant discovery and genotyping (Worley and Gibbs, 2012; Elsik et al., 2009). A genome-wide alignment of DNA fragments from a *B. taurus* individual differs from the Hereford-based reference sequence at between 7 and 8 million single-nucleotide polymorphisms (SNPs) and small (< 50 bp) insertions and deletions (Indels) (Crys-

nanto et al., 2019; Jansen et al., 2013). The number of differences is higher for DNA samples with greater divergence from the reference (Kim et al., 2017; Koufariotis et al., 2018).

The bovine linear reference sequence lacks allelic variation and nucleotides that might segregate at high frequency in animals from breeds other than Hereford. Lack of allelic diversity is an inherent drawback of linear reference sequences because it causes reference allele bias. DNA sequencing reads that contain only alleles that match corresponding reference nucleotides are more likely to align correctly than DNA fragments that also contain non-reference alleles (Van De Geijn et al., 2015; Paten et al., 2017). Reads originating from DNA fragments that are highly diverged from corresponding reference nucleotides will either obtain low alignment scores, or align at incorrect locations, or remain un-mapped (Pritt et al., 2018). Reference bias compromises analyses that are sensitive to accurately mapped reads and prevents the precise estimation of allele frequencies (Van De Geijn et al., 2015; Günther and Nettelblad, 2019; Salavati et al., 2019; Degner et al., 2009).

Graph-based (Garrison et al., 2018; Paten et al., 2017) and personalized reference genomes (Ballouz et al., 2019; Groza et al., 2020) mitigate reference allele bias. Existing linear reference coordinates can serve as backbones for variation-aware genome graphs. Nodes in the graph represent alleles at sites of variation and edges connect adjacent alleles. Once a variation-aware genome graph contains all alleles at known polymorphic sites, every haplotype can be represented as a walk through the graph (Sirén et al., 2020). However, an optimal balance between graph density and computational complexity is key to efficient whole-genome graph-based variant analysis because adding sites of variation to the graph incurs computational costs. Recently, Pritt et al. (Pritt et al., 2018) developed the FORGe software tool to prioritize variants for graph genomes. Their results provide a framework to build genome graphs that enable read mapping accuracy improvements over linear references at tractable computational complexity. A genome graph-based sequence analysis workflow is

implemented in the variation graph toolkit (*vg*, https://github.com/vgteam/vg) (Garrison et al., 2018). The *vg toolkit* enables the mapping of sequence reads to variation-aware graphs that incorporate linear reference coordinates as a backbone. It also facilitates to augment genome graphs with genetic variants that have more complex topology (e.g., duplications, inversions, and translocations) (Hickey et al., 2020). Graph-based references have been investigated primarily in humans and species with small genome sizes (Paten et al., 2017). High nucleotide diversity and the separation of individuals by breeds make cattle an ideally suited species to investigate the optimal composition of reference graphs for gigabase-sized genomes.

Here, we investigate sequence read mapping and variant genotyping accuracy using variation-aware reference structures in cattle. Using sequence variant genotypes of 288 cattle from four dairy and dual-purpose breeds, we construct breed-specific augmented and pan-genome reference graphs using the *vg toolkit* (Garrison et al., 2018). We prioritize sequence variants to be added to the graphs and assess accuracy of read mapping for variation-aware and linear references (Fig. 2.1). We show that breed-specific augmented and pan-genome graphs allow for significant read mapping accuracy improvements over linear reference sequences. We also construct a bovine whole-genome reference graph and show that unbiased and accurate sequence variant genotyping is possible from this novel reference structure. Together, we hope that our study can serve as a first step towards the transition from linear to variation-aware references in species with high genetic diversity and many sub-populations.

2.2 Results

Construction of bovine breed-specific augmented genome graphs

Breed-specific augmented reference graphs were constructed for four genetically distinct dairy (Brown Swiss (BSW), Holstein (HOL)) and dual-purpose (Fleckvieh

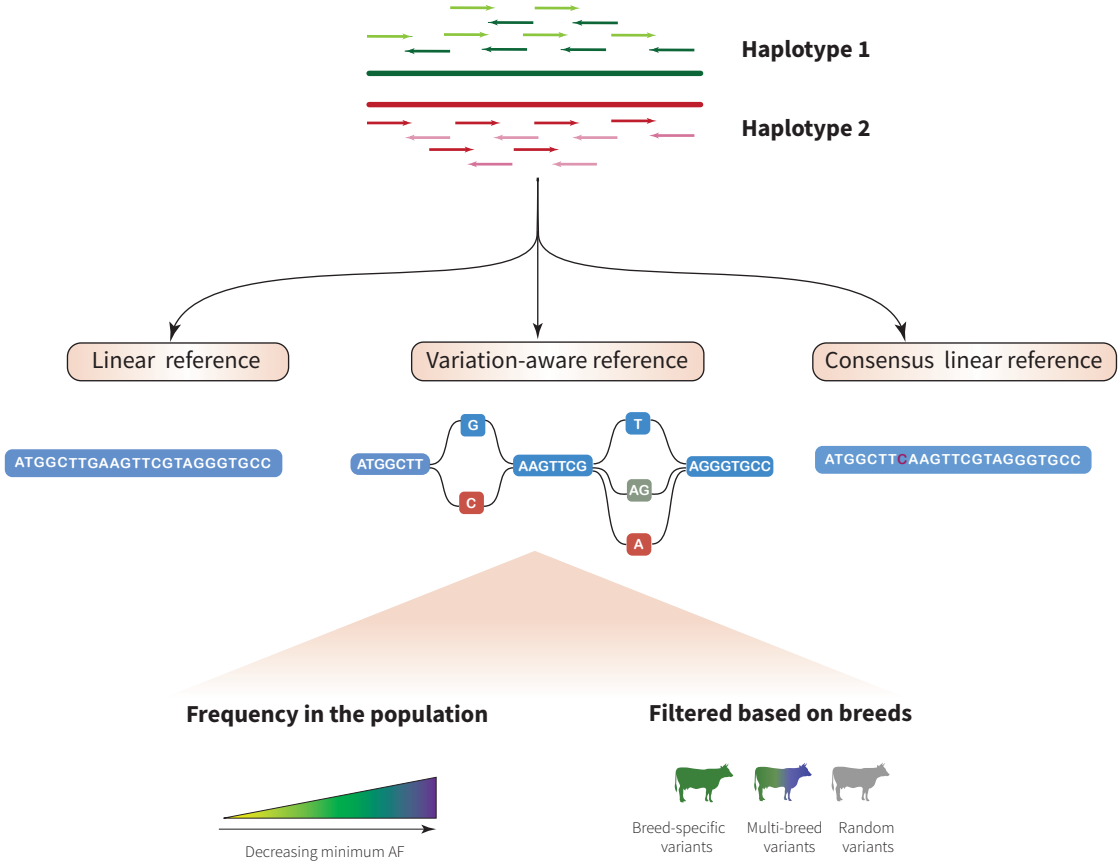


Figure 2.1: **Schematic overview of the construction of breed-specific augmented genome graphs.** We used the *vg toolkit* to augment the bovine linear reference sequence (ARS-UCD1.2) with alleles at SNPs and Indels that were discovered in 288 cattle from four breeds. Alleles that were added to the linear reference were prioritized based on their alternate allele frequency (AF). Reads simulated from true haplotypes were aligned to variation-aware, linear and consensus reference sequences to assess read mapping accuracy on cattle chromosome 25. Short-read sequencing data of Brown Swiss cattle were used to investigate sequence variant genotyping accuracy and reference allele bias using a bovine whole-genome graph as a novel reference.

(FV), Original Braunvieh (OBV)) cattle breeds using the Hereford-based linear reference sequence (ARS-UCD1.2) of chromosome 25 as a backbone (Fig. 2.2a). Average nucleotide diversity (π) estimated using 295,801 (HOL), 336,390 (FV), 347,402 (BSW), and 387,855 (OBV) biallelic variants of chromosome 25 ranged from 0.00177 (BSW) to 0.0019 (OBV) for the four breeds (Fig. 2.2b). To determine the optimal composition of bovine variation-aware references, we augmented the linear reference of chromosome 25 with an increasing number of variants (SNPs and Indels) that

were filtered for alternate allele frequency in 82 BSW, 49 FV, 49 HOL, and 108 OBV cattle. In total, we constructed 20 variation-aware graphs for each breed that contained between 2046 (variants had alternate allele frequency > 0.9) and 293,804 (no alternate allele frequency threshold) alleles.

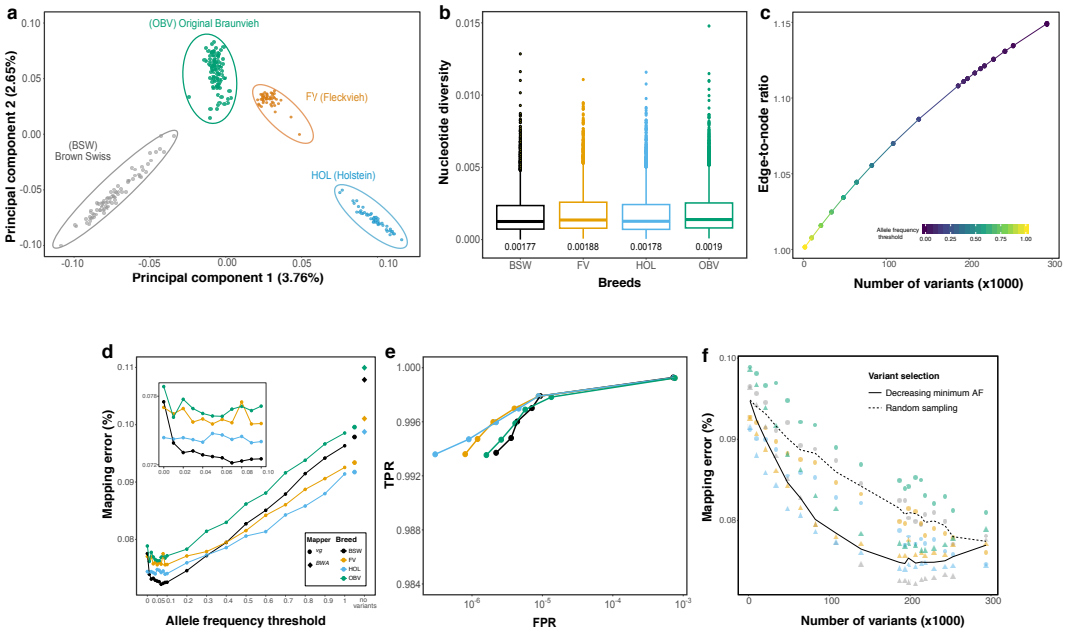


Figure 2.2: **Accuracy of mapping simulated paired-end reads to genome graphs that contained variants filtered for allele frequency at chromosome 25. a** The top principal components of a genomic relationship matrix constructed from whole-genome sequence variants reflect the genetic diversity of the four cattle breeds considered. **b** Nucleotide diversity of the four breeds calculated in non-overlapping 10-kb windows for variants of chromosome 25. The values below each boxplot indicate the nucleotide diversity for the four breeds averaged across all sliding-windows. **c** Edge-to-node ratio of graphs that contained between 2046 and 293,804 variants filtered for allele frequency. **d** Proportion of incorrectly mapped reads for four breed-specific augmented genome graphs. Diamonds and large dots represent values from linear mapping using *BWA mem* and *vg*, respectively. The inset represents a larger resolution of the mapping accuracy for alternate allele frequency thresholds less than 0.1. **e** True-positive (sensitivity) and false-positive mapping rate (specificity) parameterized on mapping quality of the best performing graph from each breed. **f** Read mapping accuracy for breed-specific augmented graphs that contained variants that were either filtered for alternate allele frequency (triangles) or sampled randomly (circles) from all variants detected within a breed. The dashed and solid line represents the average proportion of mapping errors across four breeds using random sampling and variant prioritization, respectively. Colors indicate values obtained for different breeds. Results for single-end mapping are presented in Fig. S3.2

The graph-based representation of bovine chromosome 25 (42,350,435 nucleotides)

had 1,323,451 nodes and 1,323,450 edges. The number of nodes increased proportionally with the number of variants added to the reference. When we added a maximum number of 293,804 variants to the linear reference sequence of chromosome 25, the variation-aware graph contained 2.02 million nodes. The number of edges increased faster than the number of nodes, ranging from 1.32 (empty) to 2.33 (293,804 variants included) million. Consequently, the edge-to-node ratio increased when variants were added to the graph (Fig. 2.2c). The number of paths through a graph grows rapidly with the number of variants being added to the graph. The index for the chromosome 25 reference graph contained 84.69 and 118.82 million k-mers (k = 256) when 2046 and 293,804 variants, respectively, were added to the graphs (Fig. S3.1).

Variant prioritization based on allele frequency

We simulated 10 million paired-end reads (2 x 150 bp) corresponding to approximately 35-fold coverage of bovine chromosome 25 from haplotypes of BSW, FV, HOL, and OBV cattle. Using either *BWA mem* or *vg*, we mapped the simulated reads to the respective breed-specific augmented reference graphs and the linear reference sequence. Variants that were only detected in animals used for read simulation were not added to the breed-specific augmented genome graphs. We observed fewer mapping errors using *vg* than *BWA mem* when simulated reads were aligned to a linear reference sequence. This finding was consistent for the four breeds investigated (Fig. 2.2d). Variation-aware references that contained variants filtered for allele frequency in the respective breed reduced the mapping errors for all breeds. The proportion of reads with mapping errors decreased significantly with the number of variants added to the genome graph (Fig. 2.2d, Pearson R = 0.94, P $< 10^{16}$).

Read mapping accuracy increased almost linearly between alternate allele frequency threshold 1 and 0.1, i.e., until 186,680 variants with allele frequency greater

than 0.1 were added to the graph (Pearson R = 0.94, P < 10^{16}). Adding additional alleles that had alternate allele frequency between 0.1 and 0.01 to the graphs did not further improve read mapping accuracy over the scenario with an alternate allele frequency threshold of 0.1 (P = 0.13, Fig. 2.2d inset). Read mapping accuracy declined (particularly in BSW) when the graphs contained rare alleles (alternate allele frequency < 0.01) likely because such alleles are not observed in most animals of a population. Maximum read mapping accuracy was achieved at allele frequency thresholds between 0.2 and 0.01, when the graphs contained between 139,322 and 293,628 variants filtered for allele frequency. The number of erroneously mapped reads was clearly higher for graphs that contained randomly sampled than prioritized variants (Fig. 2.2f). This finding corroborates that variant prioritization based on alternate allele frequency is important to achieve high mapping accuracy with graph-based reference structures.

We also applied the methods implemented in the FORGe software (Pritt et al., 2018) to prioritize variants for the breed-specific augmented graphs (Note S3.1). It turned out that genome graphs that were constructed with variants selected by the *Pop Cov* strategy, which relies solely on variant frequency information, enabled the highest mapping accuracy improvements over the linear reference. For example, we achieved the highest paired-end read mapping accuracy for the Brown Swiss reference graph (0.0722% erroneously mapped reads) using the Pop Cov method when 208,288 variants were added to the chromosome 25 reference (i.e., the top 60% of the ranked variants). The prioritized variants correspond to an alternate allele frequency threshold of 0.06. Variant prioritization approaches that also take into account factors other than allele frequency, e.g., the proximity of a variant to an already added variant in the graph or the repetitiveness of the resulting genome graph, did not lead to additional accuracy improvements.

Read mapping accuracy was highly correlated (*Pearson* R = 0.94, P < 10^{16}) for single- and paired-end reads (Fig. S3.2). However, the accuracy improvement of

variation-aware over linear mapping was higher for single- than paired-end reads, possibly because distance and sequence information from paired reads facilitate linear read alignment.

Read mapping accuracy differed significantly among the four breeds analyzed ($P = 10^{15}$, linear model with allele frequency as covariate) although all breed-specific augmented graphs contained the same number of variants at each allele frequency threshold (Fig. 2.2d). Linear mapping accuracy also differed among the breeds. We observed the highest error rate for reads aligned to the OBV-specific augmented reference graph. In 500 randomly sampled subsets of 35 sequenced cattle per breed, we discovered more sequence variants on chromosome 25 in OBV ($N = 305 \pm 5K$) than either FV ($N = 291 \pm 3K$), BSW ($N = 276 \pm 6K$) or HOL ($N = 259 \pm 2K$), reflecting that nucleotide diversity is higher in OBV than the other three breeds, which agrees with a recent study (Bhati et al., 2020). Across all alternate allele frequency thresholds considered, read mapping was more accurate for HOL than FV and OBV cattle, possibly because both genetic diversity and effective population size is less in HOL than the other breeds considered (Signer-Hasler et al., 2017). At allele frequency thresholds between 0.02 and 0.3, read mapping was more accurate for BSW than the other breeds. The proportion of variants with alternate allele frequency larger than 0.02 was lower for BSW(84.1%) than other breeds (86.3−89.2%). We detected more rare variants (allele frequency less than 0.05) in BSW and OBV than FV and HOL, likely reflecting differences in sample size (Fig. S3.3). An excess of singletons and rare variants in BSW and OBV cattle may have contributed to the decline in mapping accuracy at low alternate allele frequency thresholds (Fig. 2.2d inset, Table S3.2). Our findings indicate that differences in nucleotide diversity and allele frequency distributions across populations may affect read mapping accuracy to both linear and breed-specific augmented reference structures.

Comparison between bovine and human genome graphs

We used publicly available whole-genome sequence variant data from phase 3 of the 1000 Genomes Project (Consortium et al., 2015) to construct genome graphs for four genetically distinct human populations (Fig. 2.3a, GBR (British, European), YRI (Yoruba Nigeria, African), STU (Sri Lankan Tamil, South Asia), and JPT (Japanese, East Asia)). The effective population size is more than 20-fold higher for the human than cattle populations (e.g., $\sim$ 3100 for JPT and $\sim$ 7500 for YRI (Tenesa et al., 2007) vs. $\sim$ 80 for OBV and $\sim$ 160 for FV (Pausch et al., 2013; Hagger, 2005)). While the average number of sequence variants detected per sample was lower for the human than cattle populations (4,248,082 vs. 6,973,036), the proportion of singletons is higher in the human than cattle samples (23.00% in human vs. 14.01% in cattle) (S3.1). The proportion of sequence variants that had minor allele frequency less than 0.05 was between 44.88 and 55.45% in the four human and between 23.65 and 38.70% in the four cattle populations (Fig. S3.4). Nucleotide diversity ranged from 0.00098 (JPT) to 0.00141 (YRI) (Fig. 2.3b).

We considered the linear reference sequence of human chromosome 19 (g1k_v37 ref) as a backbone for the human genome graphs because its length (59,128,893 bp) and the number of variants detected per sample was similar to the values for bovine chromosome 25. Genetic diversity and allele frequency distributions were similar using either chromosome 19 or whole-genome variants indicating that the results obtained using chromosome 19 are representative for the human genome (Figs. S3.4,S3.5, Table S3.2). To construct population-specific augmented graphs, we used phased genotypes at 291,303, 306,304, 355,107, and 521,021 variants of chromosome 19 that were available for 104 JPT, 91 GBR, 102 STU, and 108 YRI individuals. Once the variants that were only detected in individuals used for simulating reads were removed from the graphs, the population-specific augmented graphs for the GBR, YRI, STU, and JPT populations contained between 3153 (alternate allele

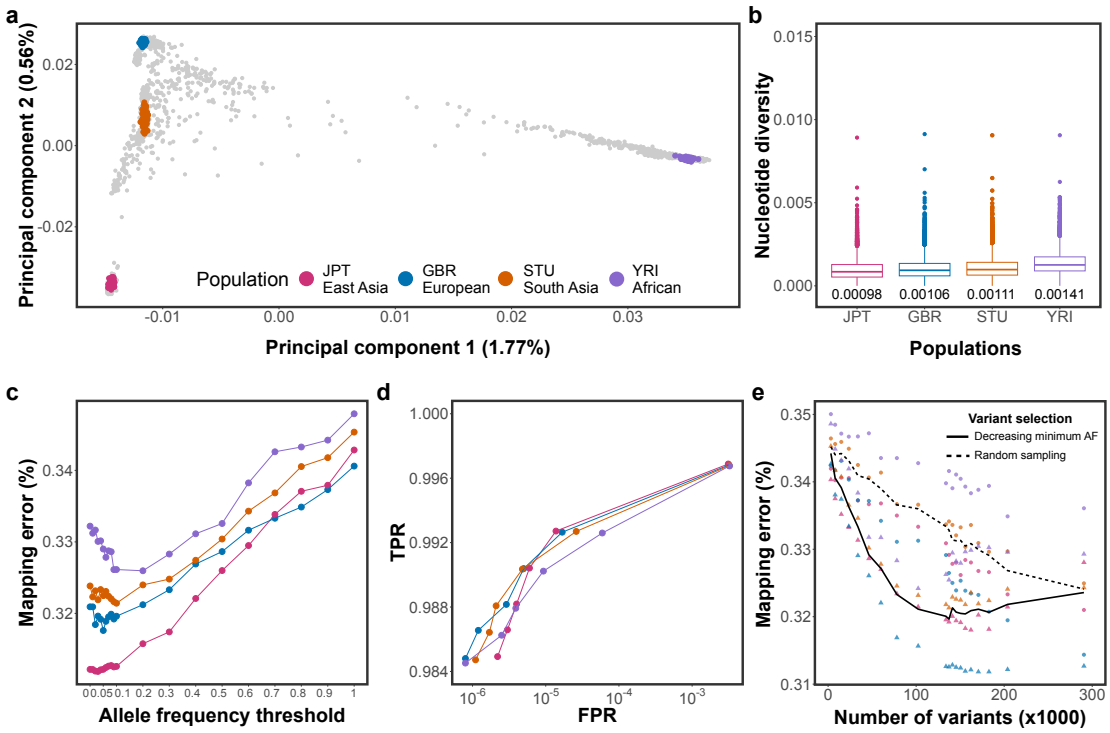


Figure 2.3: **Accuracy of mapping simulated paired-end reads to human population-specific augmented genome graphs. a** The top principal components of a genomic relationship matrix constructed from autosomal variants detected in 2504 individuals that were included in phase 3 of the 1000 Genomes Project. The colored points indicate 405 samples from the GBR (European), YRI (African), STU (South Asia), and JPT (East Asia) populations. **b** Nucleotide diversity of the four populations calculated in non−overlapping 10 kb windows for variants of chromosome 19. The values below each boxplot indicate the nucleotide diversity for the four populations averaged across all sliding-windows. **c** Proportion of incorrectly mapped reads for four population-specific augmented genome graphs. **d** Truepositive (sensitivity) and falsepositive mapping rate (specificity) parameterized on mapping quality of the best performing graph from each population. **e** Read mapping accuracy for population-specific augmented graphs that contained variants that were either filtered for alternate allele frequency (triangles) or sampled randomly (circles) from all variants detected within a population. The dashed and solid line represents the average proportion of mapping errors across four populations using variant prioritization and random sampling, respectively. Results for single-end mapping are presented in Fig S3.6

frequency $>$ 0.9) and 290,593 (no alternate allele frequency threshold) variants. We subsequently simulated 10 million reads from haplotypes of one individual per population and mapped the reads to the respective population-specific augmented genome graphs.

As observed for the bovine breed-specific augmented genome graphs, read map-

ping accuracy increased almost linearly between alternate allele frequency threshold 1 (no variants included) and 0.1 (133,891 variants added to the graph) (Fig. 2.3c). Adding low-frequency variants (alternate allele frequency between 0.01 and 0.1) did not further improve the mapping accuracy. Mapping accuracy decreased for all graphs when we added very rare variants and singletons to the graphs. This pattern was most apparent for YRI which had the highest proportion of rare variants and nucleotide diversity among the four populations considered. Read mapping accuracy differed among the four populations analyzed. We observed the lowest number of mismapped reads when reads simulated from a JPT individual were aligned to a JPT-specific augmented genome graph. The highest number of mis-mapped reads was observed when reads simulated from a YRI individual were aligned to a YRI-specific augmented genome graph. Mapping accuracy was higher for GBR than STU. These findings indicate that the mapping accuracy is negatively correlated with nucleotide diversity. Mapping accuracy improvements over the linear reference sequence were less when randomly sampled variants were added to the graphs (Fig. 2.3e).

While the overall pattern of the mapping accuracy improvements over the linear reference was similar for human and bovine genome graphs across all allele frequency thresholds considered, the proportion of mis-mapped paired-end reads was approximately four-fold higher in the human than bovine alignments (two-fold for single-end reads; S3.6). This finding was also apparent when the population-specific augmented graphs were parameterized on mapping quality to obtain sensitivity and specificity (Fig. 2.2e and Fig. 2.3d).

Mapping to breed-specific augmented genome graphs

Next, we compared read mapping accuracy between bovine breed-specific augmented and pan-genome graphs (i.e., graphs that contained variants filtered for allele fre-

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

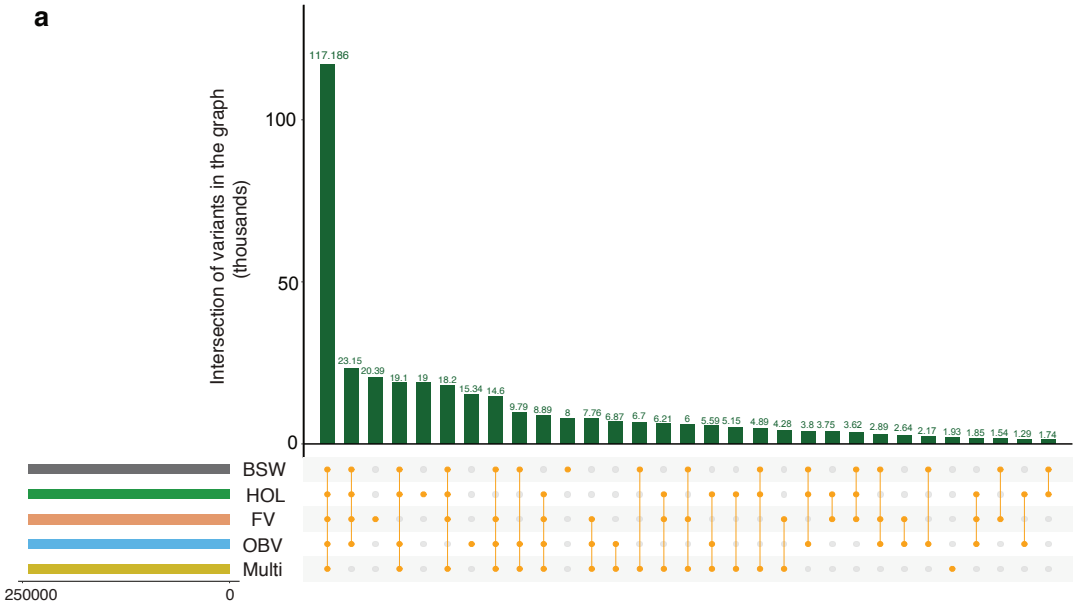
The average length, number of nodes, number of edges, and edge-to-node ratio of the variation-aware graphs were 42.60 Mb, 1,907,248, 2,155,799, and 1.13, respectively. Most variants of the random graph (87.81%) were not detected at alternate allele frequency > 0.03 in BSW, FV, OBV, and HOL indicating that they were either very rare or did not segregate in the four breeds considered in our study. Of 243,145 variants, an intersection of 48.13% had alternate allele frequency greater than 0.03 in the four breeds considered (Fig. 2.4a). The average number of variants that were

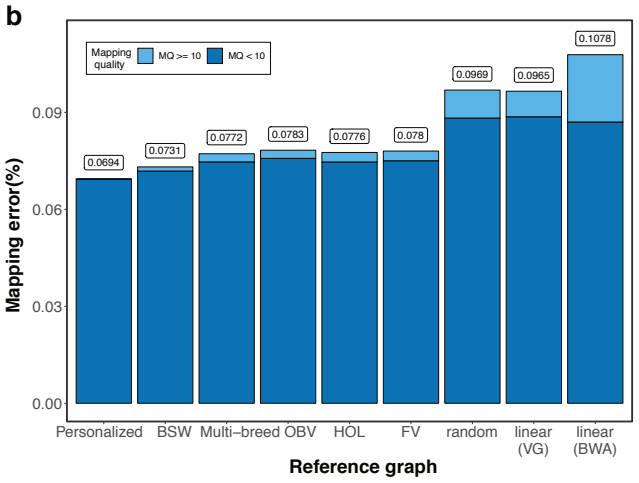
specific to the breed-specific augmented graphs ranged from 8010 in BSW to 20,392 in FV.

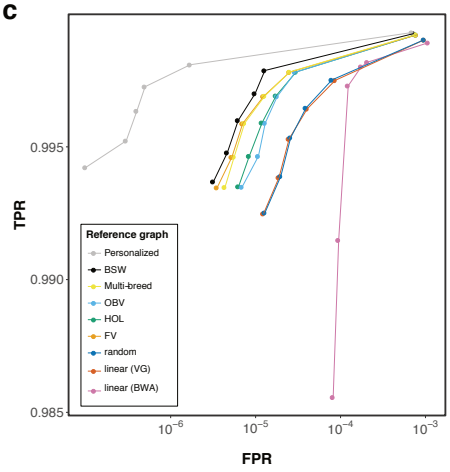
Personalized genome graphs, i.e., graphs that are tailored to a specific individual, enable the largest read mapping accuracy improvements over linear references. The proportion of mis-mapped reads was 0.0694% when a personalized BSW graph was used as a reference. Apart from the personalized graph, the highest mapping accuracy, sensitivity, and specificity was achieved when the simulated BSW reads were aligned to a BSW-specific augmented graph (Fig. 2.4b–d). The proportion of erroneously mapped paired-end reads was 0.073% for the BSW-specific augmented graph. Sensitivity and specificity were slightly lower and the number of reads with mapping errors was slightly higher when the same reads were aligned to a pan-genome graph. The read mapping accuracy differed only slightly between the breed-specific augmented and pan-genome graph because the overlap of variants that were included in both variation-aware references was high (Fig. **??**). The number of mapping errors was higher (adjusted $P < 10{-}16$, *pairwise t test*, S3.9) when BSW reads were aligned to genome graphs that contained variants filtered for allele frequency in either the FV, HOL, or OBV populations.

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

Mapping reads to a linear reference sequence using *BWA mem* with default parameter settings was the least sensitive and least specific mapping approach tested. Linear mapping using *vg* was also less accurate than variation-aware mapping. This finding indicates that accuracy improvements of variation-aware over linear mapping are attributable to differences in the reference structure rather than mapping algorithms. All graphs that contained pre-selected variants that had alternate allele frequency greater than 0.03 enabled significantly ($P = 10^{-16}$, two-sided t test)


a
Intersection of variants in the graph (thousands)
117.186
23.15
20.39
19.1
19
18.2
15.34
14.6
9.79
8.89
8
7.76
6.87
6.7
6.21
6
5.59
5.15
4.89
4.28
3.8
3.75
3.62
2.89
2.64
2.17
1.93
1.85
1.54
1.29
1.74
BSW
HOL
FV
OBV
Multi
250000
0


b
Mapping quality
MQ >= 10
MQ < 10
Mapping error(%)
0.0694
0.0731
0.0772
0.0783
0.0776
0.078
0.0969
0.0965
0.1078
Personalized
BSW
Multi−breed
OBV
HOL
FV
random
linear (VG)
linear (BWA)
Reference graph


c
TPR
FPR
Reference graph
Personalized
BSW
Multi-breed
OBV
HOL
FV
random
linear (VG)
linear (BWA)

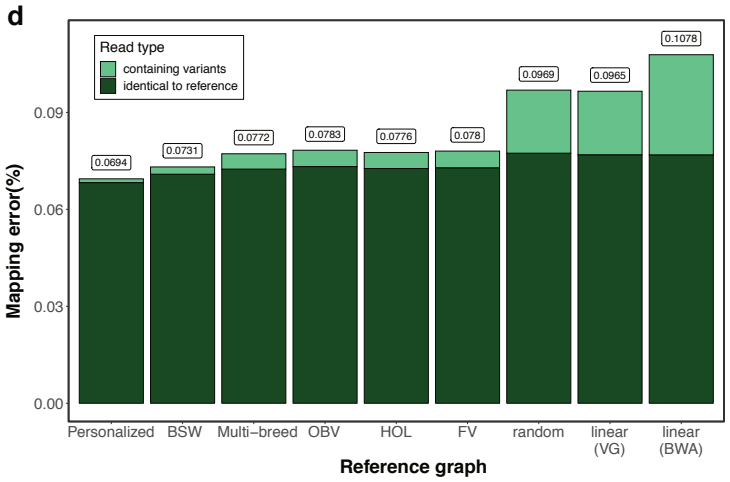

d
Read type
containing variants
identical to reference
Mapping error(%)
0.0694
0.0731
0.0772
0.0783
0.0776
0.078
0.0969
0.0965
0.1078
Personalized
BSW
Multi−breed
OBV
HOL
FV
random
linear (VG)
linear (BWA)
Reference graph

Figure 2.4: **The accuracy of mapping simulated BSW paired-end reads to variation-aware and linear reference structures.a** We added 243,145 chromosome 25 variants to the Hereford-based reference sequence that were filtered for alternate allele frequency > 0.03 in either the BSW, FV, HOL, or OBV populations. The pan-genome graph (Multi) contained 243,145 variants that had alternate allele frequency threshold > 0.03 across 288 cattle from the four breeds considered. The bars indicate the overlap of variants (averaged across ten replications) that were added to different graphs. **b** Proportion of simulated BSW reads that mapped erroneously against personalized graphs, breed-specific augmented graphs, pan-genome graphs (Multi-breed), random graphs, or linear reference sequences. We used *vg* and *BWA mem* for linear mapping. Dark and light blue colors represent the proportion of incorrectly mapped reads that had phred-scaled mapping quality (MQ) < 10 and MQ > 10, respectively. **c** True-positive (sensitivity) and false-positive mapping rate (specificity) parameterized on mapping quality. **d** Proportion of BSW reads that mapped incorrectly against breed-specific augmented graphs, pan-genome graphs (Multi-breed), random graphs, or linear reference sequences. Dark and light green colors represent the proportion of incorrectly mapped reads that matched corresponding reference nucleotides and contained non-reference alleles, respectively. Results for single-end mapping are presented in Fig. S3.7

higher mapping accuracy than linear references. This was also true when reads were mapped to graphs that contained variants that were filtered for allele frequency in a different breed, likely because many common variants segregated across the four breeds considered (Fig. 2.4a).

Recently, (Grytten et al., 2020) showed that an adjusted linear alignment approach that relies on a combination of *BWA mem* and *Minimap2* (Li, 2018) may improve linear mapping accuracy because the default setting of *BWA mem* might miss sub-optimal alignments and overestimate mapping quality for multi-mapping reads (Grytten et al., 2020; Li, 2013). We found that this approach enables to reduce the proportion of mis-mapped from 0.1078 to 0.0983 in cattle (Note S3.2). Improved mapping accuracy from the combination of *BWA mem* and *Minimap2* primarily results from less incorrectly mapped reads that had mapping quality > 10, indicating a better mapping quality assignment. The mapping accuracy from the adjusted linear alignment approach is similar to the linear mapping accuracy obtained using *vg* but considerably lower than using breed-specific augmented graphs (Note S3.2). The number of paired-end reads with mapping errors is 26% higher using the adjusted

linear alignment approach than breed-specific augmented reference graphs.

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

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

Using the BSW-specific augmented graph as a reference, only 1.76% of the incorrectly mapped reads had mapping quality (MQ) greater than 10. The MQ of the vast majority (98.24%) of incorrectly mapped reads was less than 10, i.e., they would not qualify for sequence variant discovery and genotyping using *GATK* with default parameter settings. The proportion of incorrectly mapped reads with MQ > 10 was twice as high using either the pan-genome or an across-breed augmented reference graph (3.21–3.85%). The proportion of incorrectly mapped reads with MQ > 10 was higher using either the random graph (8.92%) or linear reference sequence (*vg*: 8.19%, *BWA mem*: 19.3%).

Of 10 million simulated reads, 19.16% contained at least one nucleotide that differed from corresponding Hereford-based reference alleles. Using *BWA mem*, 47.44% and 28.72% of the erroneously mapped single- (SE) and paired-end (PE) reads, respectively, contained alleles that differed from corresponding reference nucleotides indicating that incorrectly mapped reads were enriched for reads that contained non-reference alleles (Fig. 2.4d, Figs. S3.7, S3.11). The proportion of erroneously mapped reads that contained non-reference alleles was similar for reads that were aligned to either random (47.62% and 20.13%) or empty graphs (48.20% and 20.35%)

using *vg*. However, the proportion of incorrectly mapped reads that contained non-reference alleles was clearly lower for the breed-specific augmented (SE: 1.37%, PE: 3.08%) and pan-genome graphs (SE: 2.12%, PE: 6.14%). The proportion of incorrectly mapped reads that matched corresponding reference nucleotides was almost identical across all mapping scenarios tested (Figs. 2.4d, S3.7, S3.11).

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

We further augmented the BSW-specific genome graph with 157 insertion and deletion polymorphisms of bovine chromosome 25 that were detected from short paired-end reads (2×150 bp) of 82 BSW animals using *Delly*. Adding these variants to the graph either alone or in addition to 243,145 variants that were detected using *GATK* did not improve the mapping accuracy over the corresponding scenarios that did not include these variants (Note S3.3).

Linear mapping accuracy using a consensus reference sequence

Previous studies reported that linear mapping may be more accurate using population-specific than universal linear reference sequences (Ballouz et al., 2019; Shukla et al., 2019; Dewey et al., 2011). In order to construct bovine linear consensus reference

sequences, we replaced the alleles of the chromosome 25 ARS-UCD1.2 reference sequence with corresponding major alleles at 67,142 and 73,011 variants that were detected in 82 BSW and 288 cattle from four breeds, respectively. Subsequently, we aligned 10 million simulated BSW reads to the linear adjusted sequences using either *vg* or *BWA mem*. Read mapping was more accurate to the consensus than original linear reference sequence (Figs. 2.5, S3.13). The accuracy of mapping was higher when reference nucleotides were replaced by corresponding major alleles that were detected in the target than multi-breed population. However, the mapping of reads was less accurate, sensitive, and specific using either of the consensus linear reference sequences than the breed-specific augmented graphs (Fig. 2.5b).

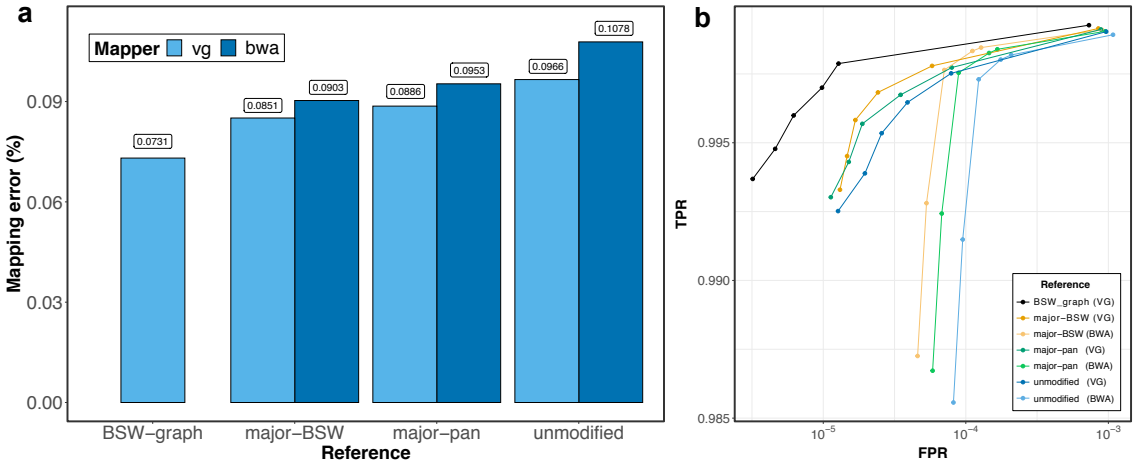


Figure 2.5: **Paired-end read mapping accuracy using breed-specific augmented genome graphs and consensus linear reference sequences. a** Dark and light blue represent the proportion of reads that mapped incorrectly using *BWA mem* and *vg*, respectively, to the BSW-specific augmented reference graph (BSW-graph), the BSW-specific (major-BSW) and the multi-breed linear consensus sequence (major-pan) and the bovine linear reference sequence (unmodified). **b** True-positive (sensitivity) and false-positive mapping rate (specificity) parameterized based on the mapping quality. The results of an analysis where reference nucleotides were only replaced at SNPs is available in Fig. S3.13

Read mapping and variant genotyping using whole genome graphs

In order to develop a breed-specific augmented reference structure for whole-genome applications, we constructed a BSW-specific augmented whole-genome variation-aware reference graph using 14,163,824 autosomal biallelic variants (12,765,895 SNPs and 1,397,929 Indels) that had alternate allele frequency greater than 0.03 in 82 BSW cattle. The resulting graph contained 111,511,367 nodes and 126,058,052 edges (an edge-to-node ratio of 1.13) and 6.32×10^9 256-mer paths. We also constructed a linear (empty) whole-genome graph that did not contain allelic variation. Subsequently, we mapped paired-end (2×150 bp) sequencing reads of 10 BSW cattle that had been sequenced at between 6- and 40-fold coverage (Table S3.4) to the variation-aware and linear reference sequence using either *vg map* or *BWA mem*. The 10 BSW cattle used for sequence read mapping were different to the 82 animals used for variant discovery, graph construction, and haplotype indexing.

62.19, 51.35 and 49.16% of the reads aligned perfectly (i.e., reads that aligned with full length (no clipping) and without any mismatches or Indels) to the BSW-specific augmented graph, the empty graph, and the linear reference sequence, respectively (Fig. 2.6a). We observed slightly less uniquely mapped reads using either the whole-genome (82.46%) or empty graph (82.18%) than the linear reference sequence (83.18%) indicating that variation-aware references can increase mapping ambiguity due to providing alternative paths for read alignment.

We converted (surjected) the graph-based read alignments of 10 BSW cattle to corresponding linear reference coordinates and genotyped polymorphic sites using *SAMtools mpileup*. In order to assess genotyping accuracy, we compared the sequence variant genotypes with array-called genotypes at corresponding positions. Sequence variant genotyping accuracy was correlated with sequencing coverage (Fig. 2.6b). Genotype concordance, non-reference sensitivity, non-reference discrepancy,

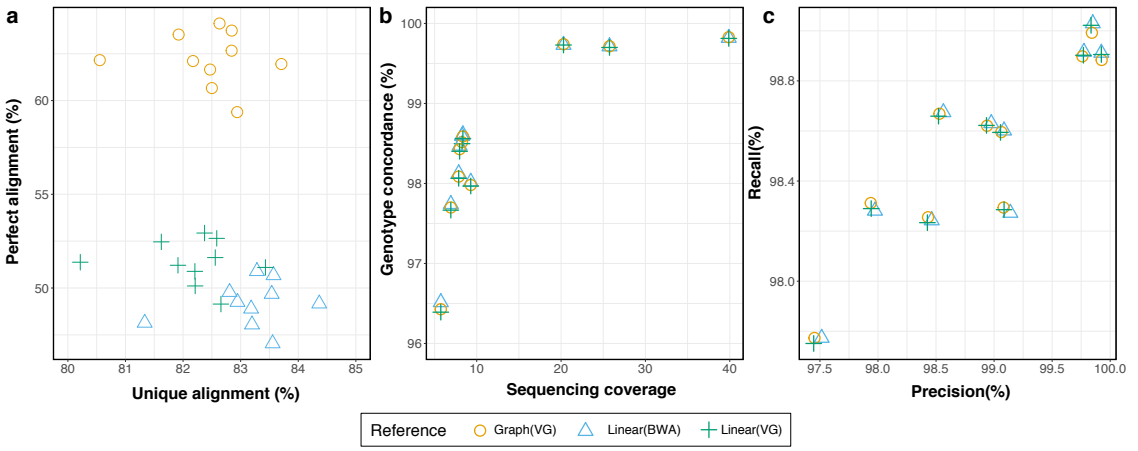


Figure 2.6: **Sequence read mapping and variant genotyping using a breed-specific augmented whole-genome graph. a** Proportion of sequencing reads that mapped perfectly and uniquely to the BSW-specific augmented (circle) and Hereford-based linear (triangle, cross) reference. **b** Concordance between sequence variant and corresponding microarray-derived genotypes as a function of sequencing depth. Sequence variant genotypes were obtained using the multi-sample variant calling approach implemented in *SAMtools*. **c** Corresponding precision-recall statistic. Each symbol represents one BSW animal

and precision did not differ between the graph-based and linear alignments for both raw and hard-filtered genotypes (Fig. 2.6b, c, Table S3.3). The average concordance, precision and recall from the graph-based alignments was 99.76, 99.84, and 98.93, respectively, for three samples (SAMEA6163185, SAMEA6163188, SAMEA6163187) with sequencing coverage greater than 20-fold. We observed similar values for genotypes called using either *GATK* or *Graphtyper* (Table S3.3). In agreement with our previous findings (Crysnanto et al., 2019), genotype concordance was slightly higher using *Graphtyper*, than either *SAMtools* or *GATK*.

Variation-aware alignment mitigates reference allele bias

To investigate reference allele bias in genotypes called from linear and graph-based alignments, we aligned sequencing reads of a BSW animal that was sequenced at

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

Using *SAMtools* to genotype sequence variants from variation-aware and linear alignments, the support for reference and alternate alleles was almost equal at heterozygous SNPs (Fig. 2.7a), indicating that SNPs are not notably affected by reference allele bias regardless of the reference structure. Alternate allele support decreased with variant length for the linear alignments. As expected, bias towards the reference allele was more pronounced at insertion than deletion polymorphisms. For instance, for 456 insertions that were longer than 30 bp, only 26% of the mapped reads supported the alternate alleles. The allelic ratio of Indel genotypes was closer to 0.5 using graph-based than linear alignments indicating that variation-aware alignment mitigates reference allele bias. However, slight bias towards the reference allele was evident at insertions with length > 12 bp, particularly if the alternate alleles were not included in the graph (Fig. 7a). Inspection of the read alignments using the Sequence Tube Map graph visualization tool (Beyer et al., 2019) corroborated that the support for alternate alleles is better using graph-based than linear references (Fig. S3.14).

Both the number of reads mapped and the number of mapped reads supporting alternate alleles was higher at Indels using graph-based than linear alignments (Fig. S3.15). The difference in the number of mapped reads between graph-based and linear alignments increased with variant length. However, the number of mapped reads supporting the reference alleles did not differ between the graph-based and linear alignments. This finding indicates that reduced reference allele bias at Indel

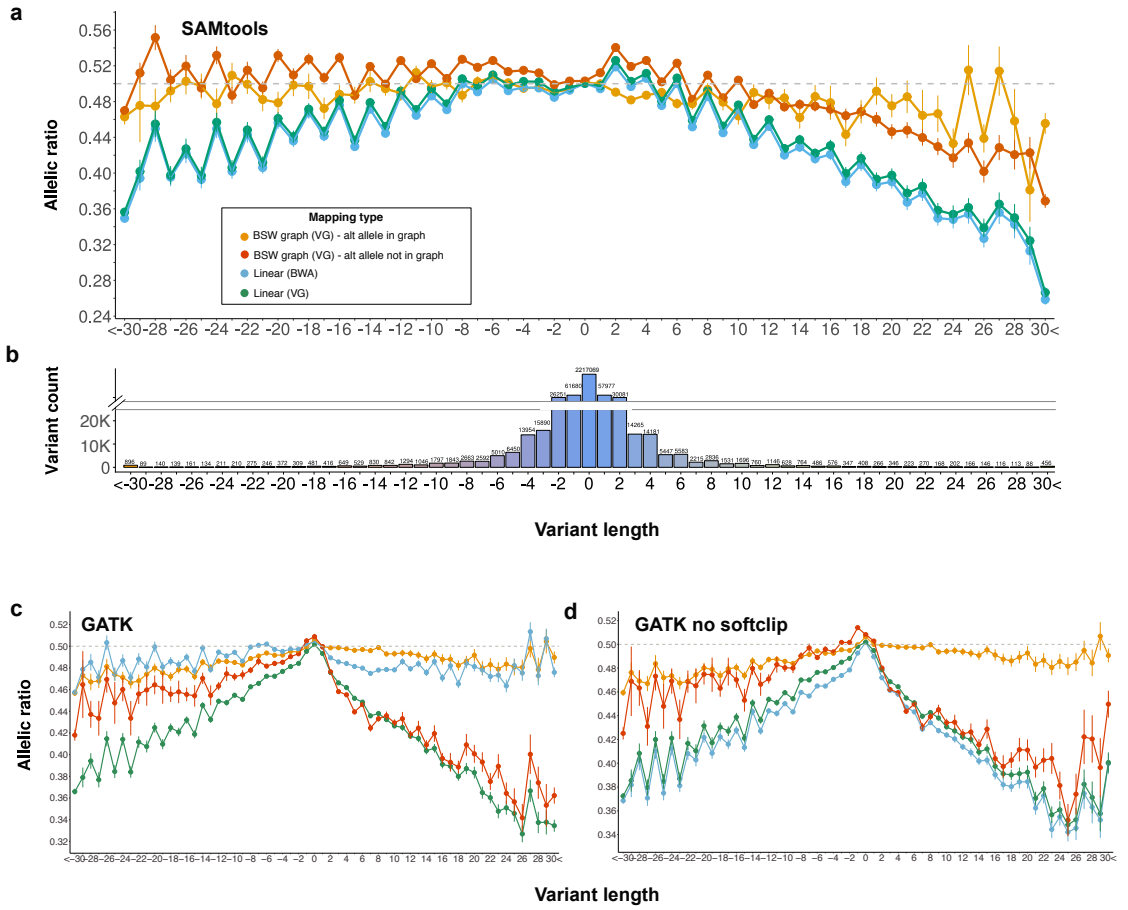


Figure 2.7: **Reference allele bias from graph-based and linear alignments.** Reference allele bias from graph-based and linear alignments using **a** *SAMtools*, **c** *GATK*, or **d** *GATK* without soft-clip for variant genotyping and either *BWA mem* or *vg* for alignment. Allelic ratio reflects the proportion of mapped reads supporting the alternate allele. The gray dashed line indicates equal support (0.5) for both alleles. Negative values, zero, and positive values along the x-axis represent deletions, SNPs, and insertions respectively. Each dot represents the mean ($\pm$ s.e.m.) allelic ratio for a given variant length. b Number of variants with a given length. To improve the readability, the values above the breakpoint of the y-axis do not scale proportionately with the height of the bars

genotypes called from graph-based alignments is due to the improved mapping of reads that contain non-reference alleles.

We next investigated if these conclusions also hold for genotypes called by *GATK*. While *SAMtools mpileup* detects variants directly from the aligned reads (Li et al., 2009), *GATK HaplotypeCaller* locally realigns the reads and calls variants from the refined alignments (Poplin et al., 2017). Using *GATK*, the allelic ratio was close to 0.5 for genotypes called from graph-based alignments across different lengths of

variants that were included in the reference graph (Fig. 2.7c). However, reference allele bias was evident at insertions that were not included in the reference graph. We also observed an almost equal number of reference and alternate alleles at variants genotyped from linear alignments using *GATK*. These findings confirm that the local realignment and haplotype-based genotyping approach of *GATK* might also mitigate reference alleles from linear alignments.

The percentage of soft-clipped reads increased with Indel length in the linear alignments (Fig. S3.16). However, the graph-based alignments contained almost no soft-clipped reads across all Indel lengths. In order to investigate the impact of soft-clipping on variant genotyping, we repeated *GATK* variant discovery and genotyping for the graph-based and linear alignments after all soft-clipped reads were removed (Fig. 2.7d). As expected, the allelic ratio of genotypes called from the graph-based alignments was not affected by the removal of (very few) soft-clipped reads. However, bias towards the reference allele became evident in genotypes called from linear alignments. This finding confirms that the local realignment of *GATK* rescues Indels that are initially soft-clipped, thus mitigating reference allele bias. This finding also implies that the original pileup information from graph-based alignments facilitates to confidently detect known Indels while avoiding local realignment as implemented in the *GATK HaplotypeCaller*.

2.3 Discussion

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

reference structures.

Using sequencing reads simulated from haplotypes of BSW, FV, OBV, and HOL cattle, our findings confirm that a breed-specific consensus sequence improves linear mapping (Ballouz et al., 2019; Shukla et al., 2019). However, read mapping is less accurate using linear consensus than variation-aware references that contain pre-selected variants. (Grytten et al., 2020) reported that an adjusted parameter setting of *BWA mem* and subsequent application of *Minimap2* may further improve the linear mapping accuracy. However, the adjusted linear mapping approach still performs worse than graph-based mapping on reads that contain variants. The accuracy improvements of the adjusted linear mapping approach were small in our study, because the number of sequence variants detected per sample and thus the proportion of reads with variants is almost twice as high in cattle than humans (Table S3.1).

Using a bovine variation-aware reference reduced the proportion of erroneously mapped reads by more than 30% compared to the most widely used linear mapping approach. A similar improvement in mapping accuracy over the linear reference was achieved for a human variation-aware reference genome (Pritt et al., 2018). The graph-based alignments using the most accurate breed-specific augmented reference graph contained 0.073% erroneously mapped reads. Incorrectly mapped reads that had high mapping quality (MQ >10) were less frequent in the graph-based than linear alignments. Thus, a variation-aware reference may reduce the number of flawed genotypes arising from mapping errors that would remain unnoticed due to high mapping quality. Similar to findings in human genome graphs (Pritt et al., 2018; Hickey et al., 2020), bovine variation-aware references did not improve the mapping of short reads that originate from low-complexity regions.

Our findings demonstrate that variant prioritization is key to accurate variation-aware read mapping. Based on investigations in four genetically distinct cattle breeds and human populations, we make three important observations: first, variation-aware references that contain random variants for which the allele frequency and

haplotype phase in the target populations is unknown do not improve read mapping accuracy over linear references. Our previous study also showed that adding many random variants does barely affect sequence variant genotyping from reference graphs (Crysnanto et al., 2019). Adding random unphased variants increases the number of alternative alignment paths that are not necessarily biologically plausible haplotypes, thus increasing mapping ambiguity. Second, read mapping accuracy increases approximately linearly with the number of randomly sampled breed-specific variants being added to the genome graph. Similar findings in the four human population-specific augmented graphs confirm that this observation also holds for populations that are strongly enriched for rare alleles and singletons. Third, the highest mapping accuracy at tractable graph complexity can be achieved when variants filtered for allele frequency are added to the graph. Using variant prioritization approaches that are based on allele frequency, we observed the highest mapping accuracy at allele frequency thresholds between 0.01 and 0.10 in four cattle breeds and four human populations. In order to reduce the computational complexity of variation-aware read mapping, previous studies used arbitrarily chosen allele frequency thresholds to prioritize variants to be included in the graphs (e.g., 1% (Garrison et al., 2018; Eggertsson et al., 2017), 5% (Maciuca et al., 2016), 10% (Kim et al., 2019)). Using fine-grained allele frequency inclusion thresholds, we find that the read mapping accuracy does not notably differ between the 0.01 and 0.1% thresholds in most populations. Yet, mapping accuracy declined rapidly for the YRI-specific augmented graph when variants with frequency less than 10% were added indicating that the optimal inclusion threshold may vary across populations. Variant prioritization approaches that also take into account factors other than allele frequency (Pritt et al., 2018) did not lead to further accuracy improvements in our study. Considering that most cattle breeds have an effective population size between 50 and 200 (Hall, 2016; Leroy et al., 2013), the vast majority of variants with allele frequency greater than 0.1 can be detected from a few sequenced key ancestor

animals (Jansen et al., 2013). As a matter of fact, key ancestor animals have been sequenced for many cattle breeds (Daetwyler et al., 2014; Bouwman et al., 2018). Thus, the construction of variation-aware reference structures that are informative for many cattle breeds is readily possible using, e.g., the sequence variant catalog of the 1000 Bull Genomes Project (Daetwyler et al., 2014; Hayes and Daetwyler, 2019).

A pan-genome graph that contained variants filtered for allele frequency across the four cattle breeds enabled almost similar accuracy improvements over the linear reference than breed-specific augmented graphs (Fig. 2.4b). Although the principal component analysis confirmed that the breeds considered in our study are genetically distinct populations, they share many common alleles. Moreover, compared to human populations, the proportion of rare alleles and singletons is low in cattle. The bovine pan-genome graph constructed in our study contained between 75.28 and 80.82% of the variants that were also added to the breed-specific augmented graphs. Instead of building many breed-specific graphs, the construction of a universal pan-genome graph is likely possible without notably compromising the accuracy of read mapping. This conclusion may hold for many species with genetically distinct sub-populations that share common alleles. Compared to the linear reference, the mapping accuracy was also significantly higher when reads from one breed were mapped to a genome graph that contained variants filtered for allele frequency in another somewhat related breed. Thus, the BSW-specific augmented whole-genome graph constructed in our study will likely improve read mapping accuracy over the linear reference and mitigate reference allele bias also for breeds other than BSW, FV, HOL, and OBV. Our BSW-specific augmented whole-genome graph is available at https://doi.org/10.5281/zenodo.3759712 (Crysnanto and Pausch, 2020a). In order to facilitate the construction of variation-aware reference structures, the entire workflow to establish whole-genome graphs is also available at https://github.com/danangcrysnanto/bovine-graphs-mapping.

The number of sequencing reads that aligned to the BSW-specific whole-genome graph with full identity increased considerably (+13%) over the linear reference sequence at the cost of a slightly reduced (−0.72%) number of unique alignments. A two-step graph alignment approach that exploits a refined search space might reduce the number of multiple mappings in dense variation-aware graphs (Grytten et al., 2020). Compared to a human whole-genome graph, the improvement in perfect mapping over the linear reference was slightly larger in our bovine whole genome graph (9.2%) (Garrison et al., 2018). However, the proportion of reads with perfect alignments (62.19%) was lower in our BSW-specific whole-genome graph, likely because it contained only sequences that were assembled to the 29 autosomes. The graph did not contain 269.77 Mb of the sex chromosomes, mitochondrial DNA, and 2180 unplaced contigs. A more sophisticated assembly of the bovine genome with increased continuity particularly at the sex chromosomes (Rice et al., 2020; Liu et al., 2019) might serve as a backbone for an improved variation-aware genome graph.

In order to detect SNPs and Indels from the variation-aware reference graph using widely used sequence variant genotyping methods, we had to make the graph-based alignments compatible with linear coordinates. Thus, our assessment of sequence variant genotyping from the bovine whole-genome graph is based on surjected graph-based alignments. It is possible that converting graph-based to linear alignments compromises variant discovery. However, the accuracy and sensitivity of genotyping did not differ between graph-based and linear alignments indicating that our whole-genome graph facilitates accurate sequence variant (SNPs and small Indels) genotyping. It is worth noting that our analysis considered only SNPs that are located in well-accessible regions of the genome, thus possibly overestimating genotyping accuracy (Li, 2014; Malomane et al., 2018). A benchmark dataset that enables unbiased evaluation of sequence variant genotyping (Li et al., 2018) is not available for the four cattle breeds considered in our study. Because approximately 90% of the

considered SNPs were already included in the BSW-specific whole-genome graph, they can be detected and genotyped easily from graph-based alignments (Paten et al., 2017). These variants can also be detected and genotyped accurately from linear alignments (Crysnanto et al., 2019; Zook et al., 2019).

As expected, bias towards the reference allele was less in graph-based than linear alignments particularly at variants that were included in the graph. Unbiased genotyping of heterozygous variants from graph-based alignments is possible because reads supporting alternate alleles align better to variation-aware than linear references. Thus, our bovine whole-genome graph offers an appealing novel reference for investigations that either rely on low-coverage sequencing or are sensitive to unbiased allele frequencies (Van De Geijn et al., 2015; Günther and Nettelblad, 2019; Rozowsky et al., 2011). Because a benchmark dataset for an unbiased evaluation of sequence variant genotyping performance (Li et al., 2018) is not available in cattle, our assessment was restricted to heterozygous variants that were identified from both linear and graph-based alignments. This set of variants is possibly enriched for variants that can be called confidently from linear alignments, thus underestimating the graph-based genotyping performance (e.g., (Garrison et al., 2018)).

Our study has three limitations. First, variants used to construct the breed-specific augmented genome graphs might be biased because they were detected from linear alignments of short sequencing reads. Variant discovery from an independent variation-aware reference structure might allow for a more complete assessment of genetic variation (Li et al., 2018). Second, we used the Hereford-based linear reference sequence as backbone to construct breed-specific augmented reference sequences. However, the Hereford-based reference sequence might lack millions of basepairs that segregate in the four breeds considered in our study (Sherman et al., 2019; Hehir-Kwa et al., 2016; Holden et al., 2018; Li et al., 2010). These nucleotides are likely missing in the breed-specific augmented reference graphs constructed in our study. Accurate and continuous genome assemblies from BSW, FV, HOL, and OBV

cattle are not available. All bovine genome assemblies that are available to date had been compiled from individuals that are distantly related to the breeds in our study (Koren et al., 2018; Rice et al., 2020; Rosen et al., 2020). Haplotype-resolved genome assemblies of cattle from different breeds will facilitate the construction of more informative genome graphs and make non-reference sequences and their sites of variation amenable to genetic investigations (Koren et al., 2018; Rice et al., 2020). Third, we did not investigate the impact of large sequence variation on sequence read mapping and variant genotyping performance because neither a high-quality benchmark set of large structural variants (cf. (Chaisson et al., 2019)) nor long-read sequencing data is available for the four cattle breeds considered. Adding insertion and deletion polymorphisms detected from short-read sequencing data did not lead to accuracy improvements in our study likely because structural variants detected from short reads are notoriously biased and incomplete (Alkan et al., 2011). Recent studies indicated that large structural variants can be identified accurately from genome graphs (Hickey et al., 2020; Eggertsson et al., 2019; Chen et al., 2019; Rakocevic et al., 2019). Eventually, a bovine genome graph that unifies multiple breed-specific haplotype-resolved genome assemblies and their sites of variation might provide access to sources of variation that are currently neglected when short sequencing reads are aligned to a linear reference sequence (Duan et al., 2019; Beyter et al., 2020; Li et al., 2020).

2.4 Conclusions

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

breeds.

2.5 Methods

Whole-genome sequencing data

We used short paired-end sequencing reads of 288 cattle from dairy (n = 82 Brown Swiss (BSW), n = 49 Holstein (HOL)) and dual-purpose (n = 49 Fleckvieh (FV), n = 108 Original Braunvieh (OBV)) breeds to detect variants that segregate in these populations. The average sequencing depth of the 288 cattle was 12.71-fold, and it ranged from 3.49 to 70.04. Most of the sequencing data were generated previously (Daetwyler et al., 2014; Crysnanto et al., 2019; Jansen et al., 2013; Baes et al., 2014; Hofstetter et al., 2019). Accession numbers for all animals are available in Table S3.4.

We trimmed adapter sequences from the raw data and discarded reads for which the phred-scaled quality was below 15 for more than 15% of the bases using fastp (Chen et al., 2018). Subsequently, the sequencing reads were aligned to the linear reference assembly of the bovine genome (ARS-UCD1.2, GCF_002263795.1) using *BWA mem* (Li, 2013). Duplicates were marked and the aligned reads were coordinate sorted using the Picard tools software suite (`http://broadinstitute.github.io/picard`) and Sambamba (Tarasov et al., 2015), respectively. We discovered and genotyped polymorphic sites from the linear read alignments using the Best Practices Workflow descriptions for multi-sample variant calling with *GATK* (version 4.1.0) (DePristo et al., 2011). Because a truth set of variants required for variant quality score recalibration (VQSR) is not available for Bos taurus, we followed the recommendations for sequence variant discovery and filtration when applying VQSR is not possible. Genotypes of the hard-filtered variants were subsequently refined, and sporadically missing genotypes were imputed with *BEAGLE* v4 (Browning and

Browning, 2016) using the genotype likelihoods from the *GATK HaplotypeCaller* model as input values. Additional information on the sequence variant genotyping workflow and the expected genotyping accuracy can be found in (Crysnanto et al., 2019). Nucleotide diversity was calculated in non-overlapping 10 kb windows separately for each breed using the π (nucleotide diversity) module implemented in the vcftools software (Danecek et al., 2011).

We discovered and genotyped large structural variants (> 50 bp) including insertions, deletions, inversions, duplications, and translocations in 82 sequenced BSW animals using *Delly* v0.7.8 (Rausch et al., 2012) with the default settings. We retained only insertion and deletion variants that had been refined using split-reads (PRECISE-flag in the vcf file).

The principal components of a genomic relationship matrix constructed from whole-genome sequence variant genotypes were calculated using PLINK v1.9 (Chang et al., 2015). The top principal components separated the animals by breeds, corroborating that the four breeds are genetically distinct (Fig. 2.2a). To take haplotype diversity and different linkage disequilibrium phases across breeds into account, the sequence variant genotypes were phased for each breed separately using *BEAGLE* v5 (Browning et al., 2018).

Unless stated otherwise, our analyses included 541,876 biallelic SNPs and Indels that were detected on bovine chromosome 25. The *vg toolkit* version 1.17.0 "Candida" (Garrison et al., 2018) was used for all graph-based analyses.

Haplotype-aware simulation of short sequencing reads

We simulated 10 million reads (150 bp) from reference haplotypes of one animal per breed that had sequencing coverage greater than 20-fold (see Table S3.4). Therefore, we added the phased sequence variants of each of the four animals to the linear reference to construct individualized reference graphs using *vg construct*. The

haplotype-aware indexes of the resulting graphs were built using *vg index xg* and *gbwt*. *vg paths* and *vg mod* were used to extract the haplotype paths from the individualized reference graphs. Subsequently, we simulated 2.5 million paired-end reads (2×150 nt) from each haplotype using *vg sim*, yielding 10 million 150 bp reads per breed corresponding to approximately 35-fold sequencing coverage of bovine chromosome 25. The simulation parameter setting for read and fragment length was 150 and 500 (± 50), respectively. The substitution and indel error rate was 0.01 and 0.002, respectively, according to the settings used in (Garrison et al., 2018).

Read mapping to graphs augmented with variants filtered for allele frequency

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

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

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

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

Read mapping to human population-specific augmented genome graphs

We downloaded phased whole-genome variants of 2504 individuals from phase 3 of the 1000 Genomes Project (Consortium et al., 2015) as well as the corresponding reference sequence (g1k_v37; https://www.internationalgenome.org/category/reference/). We selected four populations which we considered to be genetically distinct based on the results of a principal components analysis and for which the number of individuals was similar to the number of individuals for the four cattle breeds, i.e., GBR (British in England and Scotland, European), YRI (Yoruba in Ibadan Nigeria, African), JPT (Japanese in Tokyo, East Asia), and STU (Sri Lankan Tamil, South Asia). The principal components were calculated from a genomic relationship matrix constructed using 81.27 million autosomal variants using the *PLINK* (v1.9) software (Chang et al., 2015). Alternate allele frequency was calculated separately for the four populations for all variants of human chromosome 19. Nucleotide diversity was calculated with the vcftools software as detailed above. In order to construct population-specific augmented genome graphs, we used the refer-

ence sequence (g1k_v37) of human chromosome 19 as a backbone and added variants filtered for alternate allele frequency in the four populations (following the approach explained above). For each population, we constructed 20 graphs that contained between 3153 and 290,593 variants. We simulated 10 million paired-end reads for each population from reference haplotypes (as detailed above) of four selected samples (GBR: HG00096, YRI: NA18486, JPT: NA18939, STU: HG03642). The simulated reads were then mapped to the population-specific augmented genome graphs using the *vg toolkit*.

Read mapping to bovine breed-specific augmented graphs

We simulated 10 million reads from the haplotypes of a BSW animal (SAMEA6272105) and mapped them to variation-aware reference graphs that were constructed using variants (SNPs and Indels) filtered for alternate allele frequency greater than 0.03. Alleles that were only detected in SAMEA6272105 were excluded from the graphs. All graphs contained 243,145 variants. The number of variants was determined according to the HOL cattle breed because the lowest number of variants segregated at an alternate allele frequency greater than 0.03 in that breed. To investigate the utility of targeted genome graphs, we mapped the simulated BSW reads to a graph that contained variants filtered for allele frequency in BSW cattle. To investigate across-breed mapping, we mapped the simulated BSW reads to graphs that contained variants filtered for allele frequency in either FV, HOL, or OBV cattle. We also mapped the BSW reads to a bovine pan-genome graph that contained variants that were filtered for allele frequencies across the four cattle breeds. Additionally, we investigated the accuracy of mapping reads to a graph that was built from randomly selected variants. To construct the random graph, we randomly sampled from 2,294,416 variants that were detected on bovine chromosome 25 from animals of various breeds of cattle (http://www.1000bullgenomes.com/doco/ARS1.2PlusY_BQSR_v2.vcf.gz).

The allele frequencies and haplotype phases of the random variants were not known. We constructed personalized graphs that contained only variants and haplotypes that were detected in the animals used for read simulation. The variation-aware graphs were subsequently indexed using *vg index* (see above). The simulated BSW reads were mapped to the different graphs using *vg map* (see above). The construction and indexing of graphs as well as read simulation and mapping were repeated ten times. We report in the main part of the paper the average values of ten replicates. This entire procedure was repeated with reads that were simulated from the haplotypes of FV (SAMN02671626), HOL (SAMN02671584), and OBV animals (SAMEA5059743).

Read mapping to consensus reference sequences

We modified alleles of the ARS-UCD1.2 linear reference sequence using the vcf2diploid tool (Rozowsky et al., 2011). We created two adjusted linear reference sequences for bovine chromosome 25:

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

Ten million BSW reads were simulated (see above) and mapped to the original and modified linear reference sequences, as well as the corresponding variation aware reference structures using either *BWA mem* or *vg map* (see above) with default parameter settings. Since the replacement of reference alleles with Indels causes a shift in the reference coordinate system, we converted the coordinates of simulated reads between the original and modified reference using a local instance of the *UCSC liftOver* tool (Haeussler et al., 2019) that was guided using a chain file produced by

vcf2diploid. In order to prevent possible errors arising from coordinate shifts when reference nucleotides are either deleted or inserted at Indels, we repeated the analysis when only the alleles at SNPs were replaced.

Assessment of the read mapping accuracy

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

Using *vg annotate*, we annotated the simulated reads with respect to the linear reference coordinates and determined if they contained non-reference alleles. Comparing the true and mapped positions of the simulated reads enabled us to differentiate between correctly and incorrectly mapped reads. Following the approach of (Garrison et al., 2018) and taking into account the possibility that aligned reads may be clipped at Indels, we considered reads as incorrectly mapped if their starting positions were more than $k = 150$ ($k =$ read length) bases distant from true positions. The functional relevance genomic regions where the simulated reads originated from were determined based on the *Ensembl* annotation (version 99, (Yates et al., 2020)) of the bovine ARS-UCD 1.2 reference sequence. The coordinates of repetitive elements were determined based on RepeatMasker (Smith et al., 2013) annotation tables of the *UCSC* Genome Browser.

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

tic) curve (Garrison et al., 2018) using:

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

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

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

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

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

To evaluate sequence variant genotyping from the whole-genome graph, we used between 122,753,846 and 904,047,450 million paired-end (2×150 bp) sequencing reads from 10 BSW cattle (SAMEA6163185, SAMEA6163188, SAMEA6163187, SAMEA6163177, SAMEA6163178, SAMEA6163176, SAMEA6163179, SAMEA6163183, SAMEA6163181, SAMEA6163182, Table S3.4) that had been sequenced at between

5.74 and 39.88-fold genome coverage. These animals were not part of the 82 BSW animals that were used to detect the variants that were added to the graph. We trimmed adapter sequences and removed reads that had more than 20% bases with phred-scaled quality less than 20 using *fastp* (Chen et al., 2018). Subsequently, we mapped the pruned reads to either the BSW-specific augmented whole-genome graph or the linear reference sequence using either *vg map* while supplying both graph (*xg*, *gcsa*) and haplotype (*gbwt*) index to produce GAM files for each sample or *BWA mem*. To make the coordinates of the graph-based alignments compatible with linear reference coordinates, we converted the GAM- to BAM-files using *vg surject*. Variants were detected and genotyped from the surjected files using the multi-sample variant calling approach of either *GATK* (Poplin et al., 2017), *Graphtyper* (Eggertsson et al., 2017), or *SAMtools* (Li et al., 2009), as stated above and detailed in (Crysnanto et al., 2019).

In order to assess the read mapping accuracy from real sequencing data, we calculated the proportion of reads that aligned (i) perfectly and (ii) uniquely (Pritt et al., 2018; Shukla et al., 2019; Novak et al., 2017). A read was considered to map perfectly if the edit distance was zero along the entire read (NM:0 tag in *BWA mem*-aligned BAM files; identity 1 in *vg map*-aligned GAM-files), and without hard clipping (H tag) or soft clipping (S tag) in CIGAR string. A read was considered to map uniquely if either a single primary alignment was reported for the respective read or reads that had secondary alignments (XA tag in *BWA mem*-aligned BAM files; secondary_score >0 in *vg map*-aligned GAM-files) had one alignment with phred-scaled mapping quality score of 60.

The sequenced BSW animals also had Illumina SNP BeadChip-derived genotypes at between 24,512 and 683,752 positions. The sequence variant genotypes were compared to microarray-called genotypes at corresponding positions to calculate recall/non-reference sensitivity, genotype concordance, precision, and non-reference discrepancy (DePristo et al., 2011; Linderman et al., 2014). The concordance metrics

are explained in Fig. S3.17.

Snakemake workflows (Köster and Rahmann, 2012) for whole-genome graph construction, read mapping, and variant discovery are available in the *Github* repository (https://github.com/danangcrysnanto/bovine-graphs-mapping).

Assessment of reference allele bias

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

Availability of data and materials

The scripts and data used in this study are available via *GitHub* repository (https://github.com/danangcrysnanto/bovine-graphs-mapping) and archived in Zenodo (data: https://doi.org/10.5281/zenodo.3759712 (Crysnanto and Pausch, 2020a) and scripts: https://doi.org/10.5281/zenodo.3763286 (Crysnanto and Pausch, 2020b)). Raw sequencing read data of 298 cattle used for graph construction,

evaluation of variant genotyping accuracy, and assessment of reference allele bias are available at the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena) with study accession of PRJNA238491 (Daetwyler et al., 2014), PRJEB28191 (Crysnanto et al., 2019), and PRJEB18113 (Hofstetter et al., 2019). Detailed accession numbers for each sample are provided in Table S3.4.

References

C. Alkan, B. P. Coe, and E. E. Eichler. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5):363–376, 2011.

C. F. Baes, M. A. Dolezal, J. E. Koltes, B. Bapst, E. Fritz-Waters, S. Jansen, C. Flury, H. Signer-Hasler, C. Stricker, R. Fernando, et al. Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. *BMC genomics*, 15(1):1–18, 2014.

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

W. Beyer, A. M. Novak, G. Hickey, J. Chan, V. Tan, B. Paten, and D. R. Zerbino. Sequence tube maps: making graph genomes intuitive to commuters. *Bioinformatics*, 35(24):5318, 2019.

D. Beyter, H. Ingimundardottir, A. Oddsson, H. P. Eggertsson, E. Bjornsson, H. Jonsson, B. A. Atlason, S. Kristmundsdottir, S. Mehringer, M. 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.

M. Bhati, N. K. Kadri, D. Crysnanto, and H. Pausch. Assessing genomic diversity and signatures of selection in original braunvieh cattle using whole-genome sequencing data. *BMC genomics*, 21(1):1–14, 2020.

A. C. Bouwman, H. D. Daetwyler, A. J. Chamberlain, C. H. Ponce, M. Sargolzaei, F. S. Schenkel, G. Sahana, A. Govignon-Gion, S. Boitard, M. Dolezal, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nature genetics*, 50(3):362–367, 2018.

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

B. L. Browning, Y. Zhou, and S. R. Browning. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3):338–348, 2018.

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. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):s13742–015, 2015.

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.

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

S. Chen, P. Krusche, E. Dolzhenko, R. M. Sherman, R. Petrovski, F. Schlesinger, M. Kirsche, D. R. Bentley, M. C. Schatz, F. J. Sedlazeck, et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome biology*, 20(1):1–13, 2019.

. G. P. Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

D. Crysnanto and H. Pausch. Data for bovine graphs experiments (Version 1.1) [Data set], 2020a. URL https://doi.org/10.5281/zenodo.3759712.

D. Crysnanto and H. Pausch. Scripts for bovine graphs experiments (Version 1.1), 2020b. URL https://doi.org/10.5281/zenodo.3763286.

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

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

P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

J. F. Degner, J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori, Y. Gilad, and J. K. Pritchard. Effect of read-mapping biases on detecting allele-specific expression from rna-sequencing data. *Bioinformatics*, 25(24):3207–3212, 2009.

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.

F. E. Dewey, R. Chen, S. P. Cordero, K. E. Ormond, C. Caleshu, K. J. Karczewski, M. Whirl-Carrillo, M. T. Wheeler, J. T. Dudley, J. K. Byrnes, et al. Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet*, 7(9):e1002280, 2011.

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.

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.

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.

C. Groza, T. Kwan, N. Soranzo, T. Pastinen, and G. Bourque. Personalized and graph genomes reveal missing signal in epigenomic data. *Genome biology*, 21:1–22, 2020.

I. Grytten, K. D. Rand, A. J. Nederbragt, and G. K. Sandve. Assessing graph-based read mappers against a baseline approach highlights strengths and weaknesses of current methods. *BMC genomics*, 21:1–9, 2020.

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

M. Haeussler, A. S. Zweig, C. Tyner, M. L. Speir, K. R. Rosenbloom, B. J. Raney, C. M. Lee, B. T. Lee, A. S. Hinrichs, J. N. Gonzalez, et al. The ucsc genome browser database: 2019 update. *Nucleic acids research*, 47(D1):D853–D858, 2019.

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

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

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.

J. Y. Hehir-Kwa, T. Marschall, W. P. Kloosterman, L. C. Francioli, J. A. Baaijens, L. J. Dijkstra, A. Abdellaoui, V. Koval, D. T. Thung, R. Wardenaar, et al. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nature communications*, 7(1):1–10, 2016.

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

S. Hofstetter, F. Seefried, I. M. Häfliger, V. Jagannathan, T. Leeb, and C. Drögemüller. A non-coding regulatory variant in the 5′-region of the mitf gene is associated with white-spotted coat in brown swiss cattle. *Animal genetics*, 50(1):27–32, 2019.

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.

S. Jansen, B. Aigner, H. Pausch, M. Wysocki, S. Eck, A. Benet-Pagès, E. Graf, T. Wieland, T. M. Strom, T. Meitinger, et al. Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. *BMC genomics*, 14(1):1–9, 2013.

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.

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.

J. Köster and S. Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

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.

G. Leroy, T. Mary-Huard, E. Verrier, S. Danvy, E. Charvolin, and C. Danchin-Burge. Methods to estimate effective population size using pedigree data: Examples in dog, sheep, cattle and horse. *Genetics Selection Evolution*, 45(1):1–10, 2013.

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

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

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

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

H. Li, J. M. Bloom, Y. Farjoun, M. Fleharty, L. Gauthier, B. Neale, and D. MacArthur. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nature methods*, 15(8): 595–597, 2018.

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

R. Li, Y. Li, H. Zheng, R. Luo, H. Zhu, Q. Li, W. Qian, Y. Ren, G. Tian, J. Li, et al. Building the sequence map of the human pan-genome. *Nature biotechnology*, 28(1):57–63, 2010.

M. D. Linderman, T. Brandt, L. Edelmann, O. Jabado, Y. Kasai, R. Kornreich, M. Mahajan, H. Shah, A. Kasarskis, and E. E. Schadt. Analytical validation of whole exome and whole genome sequencing for clinical applications. *BMC medical genomics*, 7(1):1–11, 2014.

R. Liu, W. Y. Low, R. Tearle, S. Koren, J. Ghurye, A. Rhie, A. M. Phillippy, B. D. Rosen, D. M. Bickhart, T. P. Smith, et al. New insights into mammalian sex chromosome structure and evolution using high-quality sequences from bovine x and y chromosomes. *BMC genomics*, 20 (1):1–11, 2019.

S. Maciuca, C. del Ojo Elias, G. McVean, and Z. Iqbal. A natural encoding of genetic variation in a burrows-wheeler transform to enable mapping and genome inference. In *International Workshop on Algorithms in Bioinformatics*, pages 222–233. Springer, 2016.

D. K. Malomane, C. Reimer, S. Weigend, A. Weigend, A. R. Sharifi, and H. Simianer. Efficiency of different strategies to mitigate ascertainment bias when using snp panels in diversity studies. *BMC genomics*, 19(1):1–16, 2018.

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.

A. M. Novak, G. Hickey, E. Garrison, S. Blum, A. Connelly, A. Dilthey, J. Eizenga, M. S. Elmohamed, S. Guthrie, A. Kahles, et al. Genome graphs. *bioRxiv*, page 101378, 2017.

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.

H. Pausch, B. Aigner, R. Emmerling, C. Edel, K.-U. Götz, and R. Fries. Imputation of high-density genotypes in the fleckvieh cattle population. *Genetics Selection Evolution*, 45(1):1–10, 2013.

R. Poplin, V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, G. A. Van der Auwera, D. E. Kling, L. D. Gauthier, A. Levy-Moonshine, D. Roazen, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, page 201178, 2017.

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.

T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel. Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, 2012.

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.

J. Rozowsky, A. Abyzov, J. Wang, P. Alves, D. Raha, A. Harmanci, J. Leng, R. Bjornson, Y. Kong, N. Kitabayashi, et al. Alleleseq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology*, 7(1):522, 2011.

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.

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

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.

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

A. Smith, R. Hubley, and P. Green. Repeatmasker open-4.0. *RepeatMasker Open-4.0*, 2013.

A. Tarasov, A. J. Vilella, E. Cuppen, I. J. Nijman, and P. Prins. Sambamba: fast processing of ngs alignment formats. *Bioinformatics*, 31(12):2032–2034, 2015.

A. Tenesa, P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke, M. E. Goddard, and P. M. Visscher. Recent human effective population size estimated from linkage disequilibrium. *Genome research*, 17(4):520–526, 2007.

B. Van De Geijn, G. McVicker, Y. Gilad, and J. K. Pritchard. Wasp: allele-specific software for robust molecular quantitative trait locus discovery. *Nature methods*, 12(11):1061–1063, 2015.

K. C. Worley and R. A. Gibbs. Sequencing the bovine genome. *Bovine Genomics*, page 109, 2012.

A. D. Yates, P. Achuthan, W. Akanni, J. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, A. G. Azov, R. Bennett, et al. Ensembl 2020. *Nucleic acids research*, 48(D1):D682–D688, 2020.

J. M. Zook, J. McDaniel, N. D. Olson, J. Wagner, H. Parikh, H. Heaton, S. A. Irvine, L. Trigg, R. Truty, C. Y. McLean, et al. An open resource for accurately benchmarking small variant and reference calls. *Nature biotechnology*, 37(5):561–566, 2019.

Chapter 3

Analysis of the multi-assembly graphs

Preface: Bridging text between Chapter 2 and Chapter 3

In this chapter, I constructed the first cattle whole genome graph and performed the first assessment of the gigabase genome graph on the species other than human. I showed using both real and simulated datasets that the graph facilitate accurate read mapping and unbiased sequence variant genotyping. I developed the graph pipeline further from previous implementation based on *vg toolkit* allowing graph analysis performed in a full genome scale. Additionally, I included catalogues of previously discovered variants to the graph, which showed that *breed-specific* graph perform similarly as the *multi-breed pangenome graph*.

Contribution: I and Hubert Pausch conceived the study, I wrote the full genome graph pipelines and performed all analyses. I wrote the initial draft of the manuscript with input from Hubert Pausch.

Novel functional sequences uncovered through a bovine multi-assembly graph

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

Animal Genomics, ETH Zürich, Zürich, 8315 Switzerland

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

Abstract

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

Keywords:Genetic diversity, Genome graphs, Pangenome

Significance

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

3.1 Introduction

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

well as for yak (*Bos grunniens*), a closely related species to domestic cattle (Low et al., 2020; Rice et al., 2020).

DNA sequences from taurine and indicine cattle are typically aligned to the Hereford-based reference genome to discover and genotype variable sites. Reference-guided read alignment and variant genotyping has revealed millions of polymorphic variants that segregate within and between taurine and indicine cattle breeds (Kim et al., 2020; Daetwyler et al., 2014; Koufariotis et al., 2018). However, using the linear reference in this alignment approach is susceptible to reference allele bias, particularly for DNA samples that are greatly diverged from the reference (Ballouz et al., 2019; Pritt et al., 2018). Moreover, reference-guided methods are blind to variations in sequences that are not present in the reference genome (Wong et al., 2020). Recent estimates suggest that millions of bases are missing in mammalian reference genomes (Sherman et al., 2019; Whitacre et al., 2015), indicating a high potential for bias.

Efforts to mitigate reference allele bias and increase the genetic diversity of reference genomes have led to graph-based references (Garrison et al., 2018,?). We have previously shown that a genome graph, which integrates linear reference coordinates and pre-selected variants, improves the mapping of reads and enables unbiased variant genotyping in different breeds of cattle (Crysnanto et al., 2019; Crysnanto and Pausch, 2020). However, previous attempts focused on augmenting the *Bos taurus* reference genome with small variations (¡50bp), not the larger class of structural variations. Despite being an important source of genotypic and phenotypic diversity (Song et al., 2020; Kehr et al., 2017), little is known about the prevalence and functional impact of structural variations in the cattle genome. The availability of reference-quality assemblies and long read sequencing data from different breeds of cattle now provides an opportunity to characterize sequence diversity beyond small variations (Hickey et al., 2020; Li et al., 2020). In this paper, we integrate reference-quality assemblies from multiple taurine breeds as well as two close relatives into a

multi-assembly graph with minigraph (21). We detect autosomal sequences that are missing in the *Bos taurus* reference genome and investigate their functional significance using transcriptome data. We show that the non-reference sequences contain novel transcripts that are differentially expressed as well as polymorphic sites that segregate within and between breeds of cattle.

3.2 Results

References

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

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

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

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.

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.

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

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

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.

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

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

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

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.

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

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

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

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

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

Supplementary Materials Chapter 1

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

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

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

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

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

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

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

1. Clone the *Graphtyper Github*

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

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

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

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

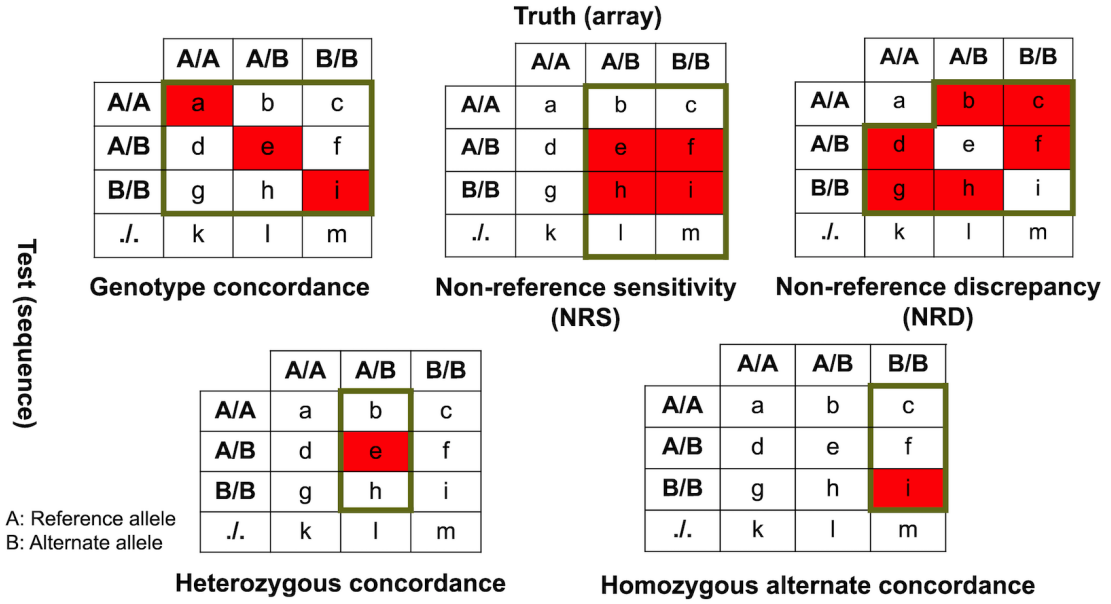
4. Make sure that these dependencies are installed:
    - C++ compiler with C++11 supported (we tested gcc 4.8.5 or gcc 6.3.0
    - Boost≥1.57.0
    - zlib≥1.2.8
    - libbz2
    - liblzma
    - Autotools, Automake, libtool, Make, and CMake¿=2.8.8
5. Follow installation procedures as below. This will put the software in *releasebuild/bin/graphtyper*

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

Additional file 2.2

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

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

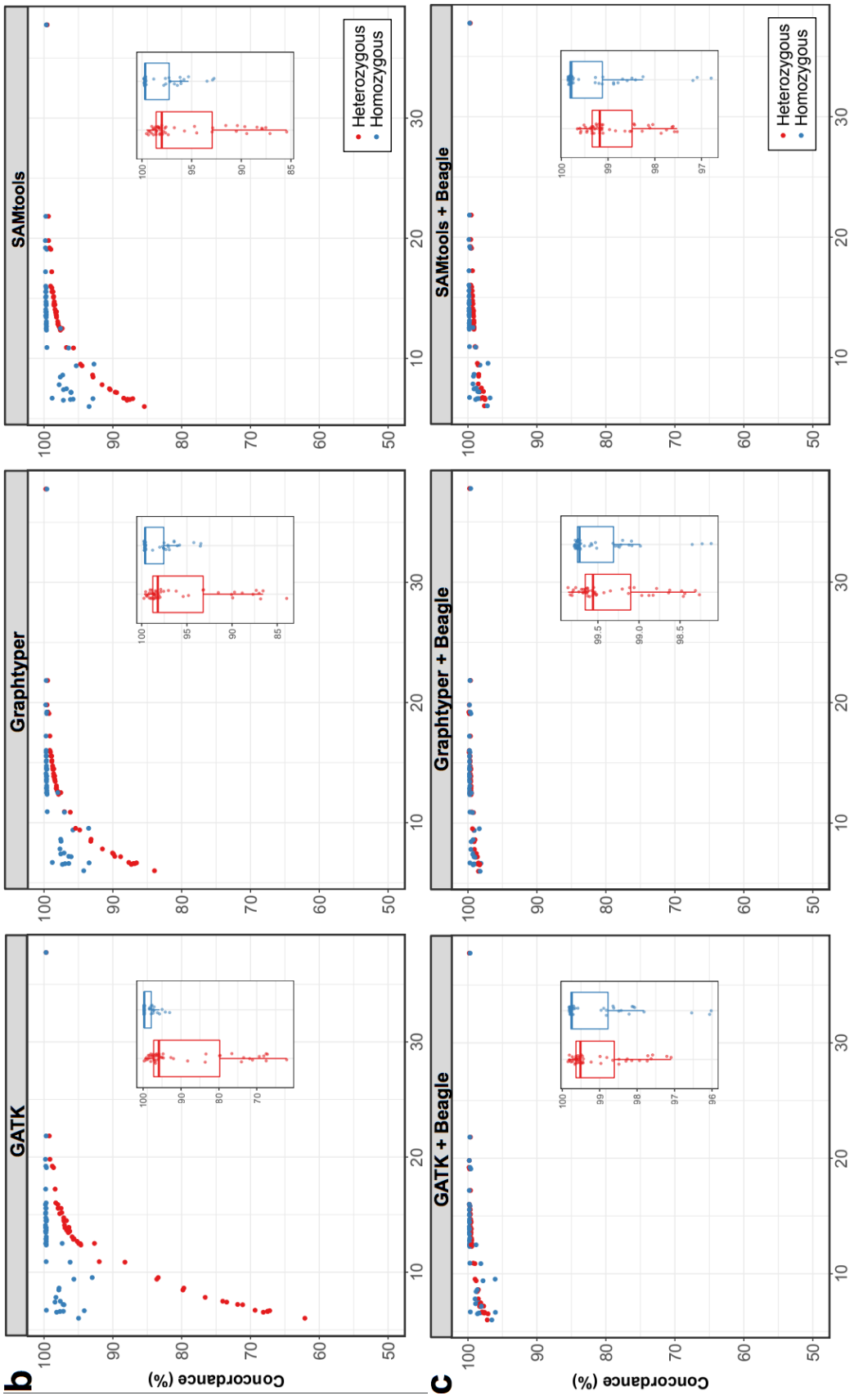


Additional file 2.3

Concordance statistics

The concordance of heterozygous and alternate homozygous genotypes in 49 Original Braunvieh cattle (**a**) and the concordance at the different sequencing depth for the (**b**) raw and (**c**) imputed datasets.

	Heterozygous concordance				Homozygous concordance			
	full		filtered		full		filtered	
	raw	imp	raw	imp	raw	imp	raw	imp
GATK	89.17	99.11	89.24	99.21	*98.74*	99.18	*98.75*	99.27
Graphtyper	*95.79*	*99.36*	*95.82*	*99.44*	98.55	*99.51*	98.59	*99.57*
SAMtools	95.73	98.91	95.77	98.99	98.46	99.37	98.48	99.41


b
c
GATK
Graphtyper
SAMtools
GATK + Beagle
Graphtyper + Beagle
SAMtools + Beagle
Heterozygous
Homozygous
Concordance (%)

Additional file 2.4

Sequence variant genotyping quality for 18 and 31 animals that were sequenced at a lower and higher than 12-fold sequencing coverage, respectively.

Asterisks denote significant differences with the best value (italic) for a respective parameter.

Coverage less than 12

	Genotype concordance				Non-reference sensitivity				Non-reference discrepancy			
	full		filtered		full		filtered		full		filtered	
	raw	imp	raw	imp	raw	imp	raw	imp	raw	imp	raw	imp
GATK	90.99***	98.7***	91.02***	98.82***	85.63***	98.91	85.51***	98.73	14.64***	2.09***	14.59***	1.91***
Graphtyper	94.89	99.07	94.91	99.17	96.44	99	96.13	98.71	8.04	1.49	8	1.31
SAMtools	94.87	98.61***	94.89	98.67***	96.24***	98.94	95.75***	98.45***	8.11	2.24***	8.09	2.11***

Coverage more than 12

	Genotype concordance				Non-reference sensitivity				Non-reference discrepancy			
	full		filtered		full		filtered		full		filtered	
	raw	imp	raw	imp	raw	imp	raw	imp	raw	imp	raw	imp
GATK	98.73***	99.66	98.76***	99.71	98.3***	99.61	98.14***	99.39	1.8***	0.48*	1.76***	0.42
Graphtyper	99.26	99.67	99.3	99.72	99.25	99.54***	98.88	99.16***	1.04	0.45	0.99	0.4
SAMtools	99.21***	99.59***	99.24***	99.62***	99.21**	99.58***	98.51***	98.79***	1.12***	0.58***	1.08***	0.54***

Additional file 2.5

Twelve 1-Mb regions for which *Graphtyper* initially failed to genotype sequence variants

The algorithm either ran out of memory or exceeded the allocated runtime (12 h). Graphtyper eventually produced genotypes for the sequence variants when these regions were re-run in 10-kb segments.

No	Chromosome	Region (Mb)
1	1	0-1
2	1	145-146
3	3	69-70
4	7	58-57
5	8	110-111
6	12	76-77
7	23	26-27
8	23	29-30
9	26	50-51
10	27	37-38
11	28	39-40
12	29	30-31

Additional file 2.6
Variant filtration using *GATK*

The best practice guidelines for variant discovery using *GATK* recommend sequence variants to be filtered using Variant Quality Score Recalibration (VQSR) because it implements advanced machine learning-based methods to differentiate between true and false-positive variants. However, VQSR relies on sets of high confidence truth/training variants, which are currently not (publicly) available in cattle. Thus, we ran *GATK* with best practice recommendations for variant filtering when applying VQSR is not possible, i.e., we used a generic baseline hard-filtering threshold for each variant annotation (see https://gatkforums.broadinstitute.org/GATK/discussion/2806/howto-apply-hard-filters-to-acall-set). This threshold-based filtering is commonly applied the cattle genomics community (Koufariotis et al., 2018; Chen et al., 2018)

To facilitate running the VQSR module in sheep and goat, i.e., species where sets of truth/training variants are not (publicly) available, (Alberto et al., 2018) used an intersection of high confidence variants that had been discovered from multiple variant callers as truth/training sets, i.e., they derived truth/training sets directly from the analyzed data. We implemented their approach to apply *GATK* VQSR to our variant dataset. Training and truth sets were constructed using the overlap of the filtered variants from the *GATK*, *Graphtyper* and *SAMtools* pipelines (truth=false, training=true, known=false, prior= 10) and markers from the BovineHD BeadChip (truth=true, training=true, known=false, prior= 15), respectively. Moreover, we used variants listed in dbSNP (version 150) as known variants (truth=false, training=false, known=true, prior=3.0). Following *GATK* VQSR, we retained variants in the 99.9% tranche sensitivity threshold (best practice).

Variant filtration using *GATK* VQSR removed more variants from the raw data than *GATK* hard filtering (Table 1). However, VQSR retained more HD SNPs than *GATK* hard filtering, possibly reflecting bias that results from the use of HD SNPs as training/truth sets. The values of the concordance statistics (genotype concordance, non-reference sensitivity, nonreference discrepancy) were almost identical between *GATK* VQSR and *GATK* hard filtration (Table 2) indicating that the choice of either filtration option does not notably affect the concordance between sequence-derived and BovineHD SNP array-derived genotypes. These findings are in line with (Vander Jagt et al., 2018) who showed that the concordance between microarray-called and sequence-derived genotypes is almost identical using either *GATK* VQSR or the *GATK* 1000 bull genomes project hard filters, even though they used stringently filtered truth/training sets based on a more comprehensive catalogue of variants than in our study. Interestingly, in agreement with (Vander Jagt et al., 2018), the proportion of opposing homozygous genotypes in sire/son-pairs (which does not suffer from ascertainment bias because it is calculated using sequence-derived SNPs) is less using *GATK* hard filter than *GATK* VQSR.

The performance of *GATK* VQSR may be assessed using the novel variant sensitivity tranche plot (Figure 2). In the lowest 90% tranches (highest specificity) the filtering model still retained many false positive variants (orange box and low Ti/Tv ratio). However, when the 99.9% tranche sensitivity is used as filtration criterion as recommended by the *GATK* best practice guidelines, a high proportion of true positive variants is removed from the data.
Overall, our findings suggest that

(i) *GATK* VQSR removes more variants from the data than *GATK* hard filtering,

(ii) *GATK* VQSR does not notably improve the concordance between sequencederived and microarray-called genotypes compared to *GATK* hard filtering,

(iii) the proportion of opposing homozygous genotypes in sire/son-pairs is higher using *GATK* VQSR than *GATK* hard filtering, and

(iv) improving VQSR may be possible by providing more sophisticated truth/training variant datasets produced by orthogonal sequencing technology other than the ones used for training, e.g. (Li et al., 2018)

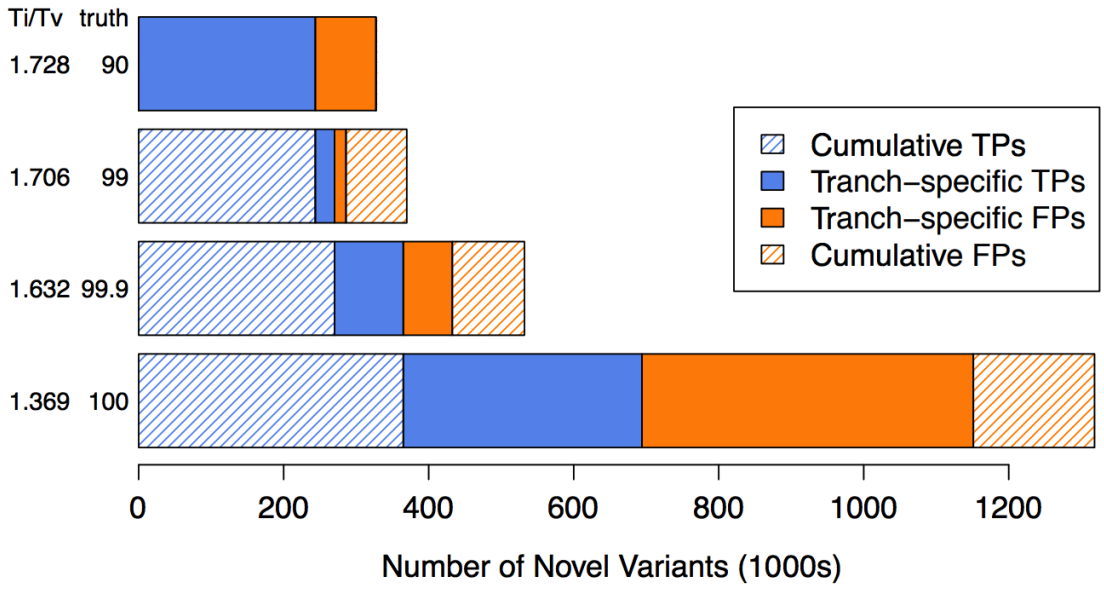
Table 1 Comparison of variants statistics between unfiltered and filtered datasets using either hard-filtering or VQSR.

	GATK full	*GATK* hard-filter	*GATK* VQSR
Total SNPs	18,594,182	17,248,593	16,537,577
Biallelic	18,347,962	17,111,806	16,430,734
Multi-allelic	246,220	136,787	106,843
Ti/Tv ratio	2.09	2.17	2.16
BovineHD	99.46	99.21	99.38
BovineSNP50	99.14	98.91	98.98

Table 2 The concordance statistics between hard-filtered and VQSR

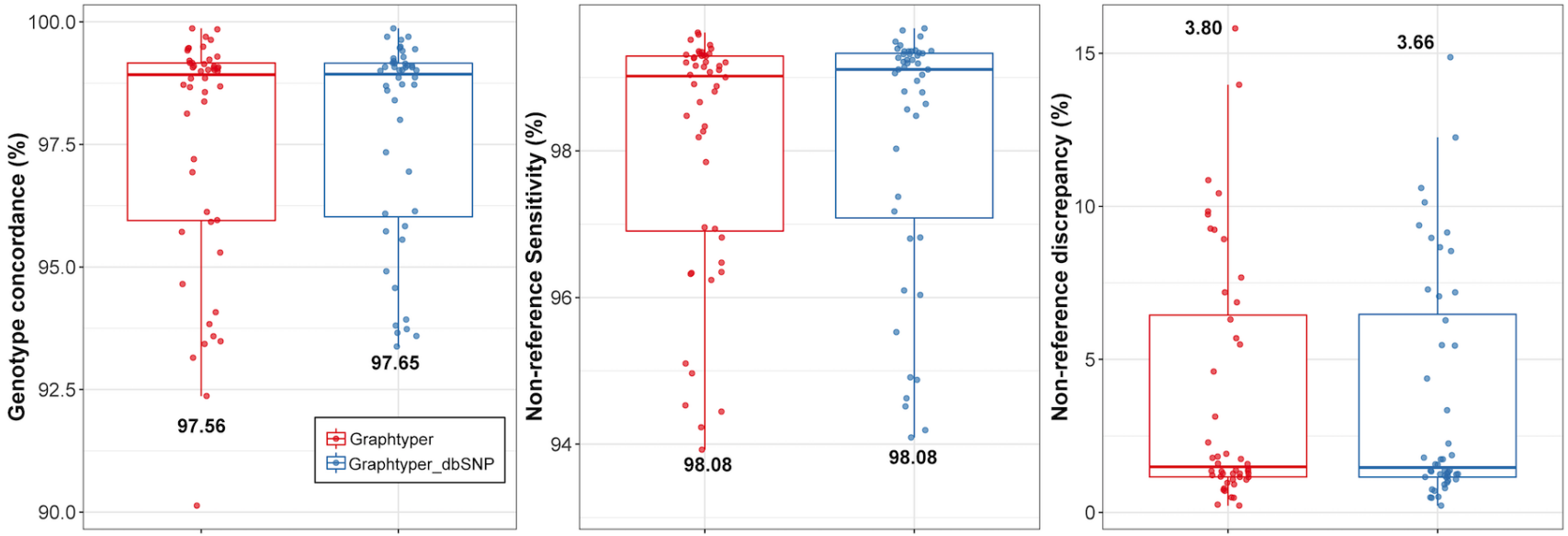
	Genotype concordance	Non-reference sensitivity	Non-reference discrepancy	Opposing Homozygous
GATK hard-filter	96.02	93.67	6.3	0.72
GATK VQSR	96.01	93.77	6.32	0.75

Figure 1 Tranche sensitivity plot of novel variants as reported by the VQSR model fitting



Additional file 2.7

Accuracy and sensitivity of sequence variant genotyping on bovine chromosome 25 from a variation-aware genome graph that incorporated 2,143,417 dbSNP variants as prior known variants.



Supplementary References

F. J. Alberto, F. Boyer, P. Orozco-terWengel, I. Streeter, B. Servin, P. De Villemereuil, B. Benjelloun, P. Librado, F. Biscarini, L. Colli, et al. Convergent genomic signatures of domestication in sheep and goats. *Nature Communications*, 9(1):1–9, 2018.

N. Chen, Y. Cai, Q. Chen, R. Li, K. Wang, Y. Huang, S. Hu, S. Huang, H. Zhang, Z. Zheng, et al. Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in east asia. *Nature Communications*, 9(1):1–13, 2018.

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.

H. Li, J. M. Bloom, Y. Farjoun, M. Fleharty, L. Gauthier, B. Neale, and D. MacArthur. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nature methods*, 15(8): 595–597, 2018.

C. Vander Jagt, A. Chamberlain, R. Schnabel, B. Hayes, and H. Daetwyler. Which is the best variant caller for large whole-genome sequencing datasets. In *Proceedings of the 11th world congress on genetics applied to livestock production*, pages 11–16, 2018.

Supplementary Materials
Chapter 2

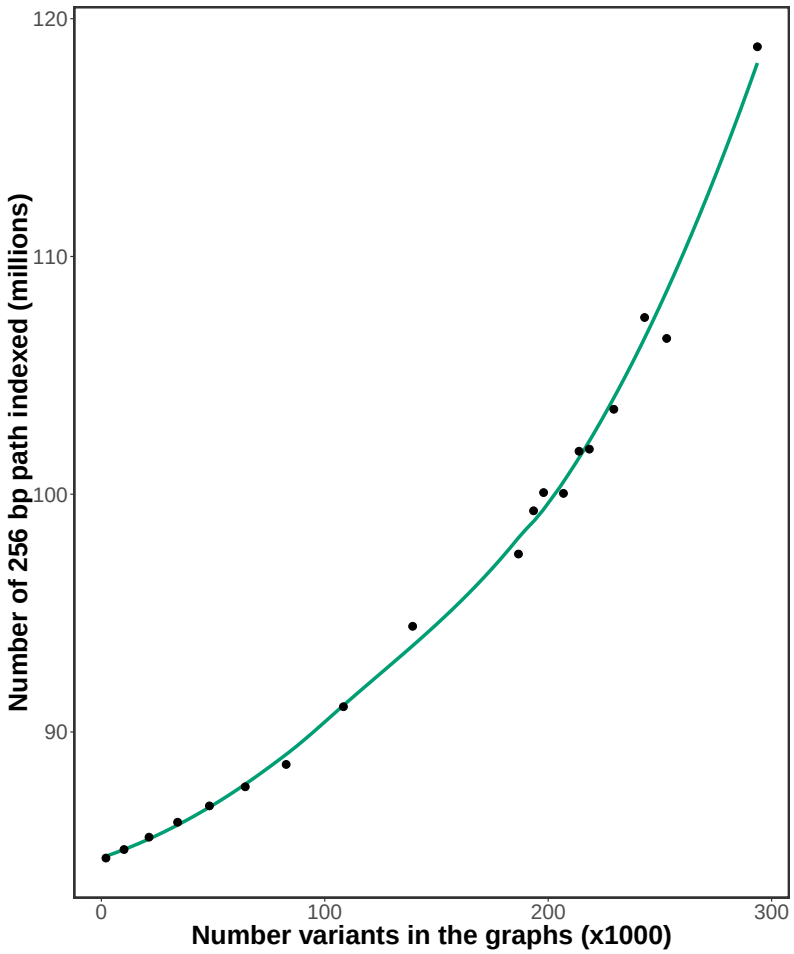


Figure S3.1: **Number of 256 bp haplotype paths in the graphs with an increasing number of variants added to the graphs.**
The line plot is fitted using loess function in *R*.

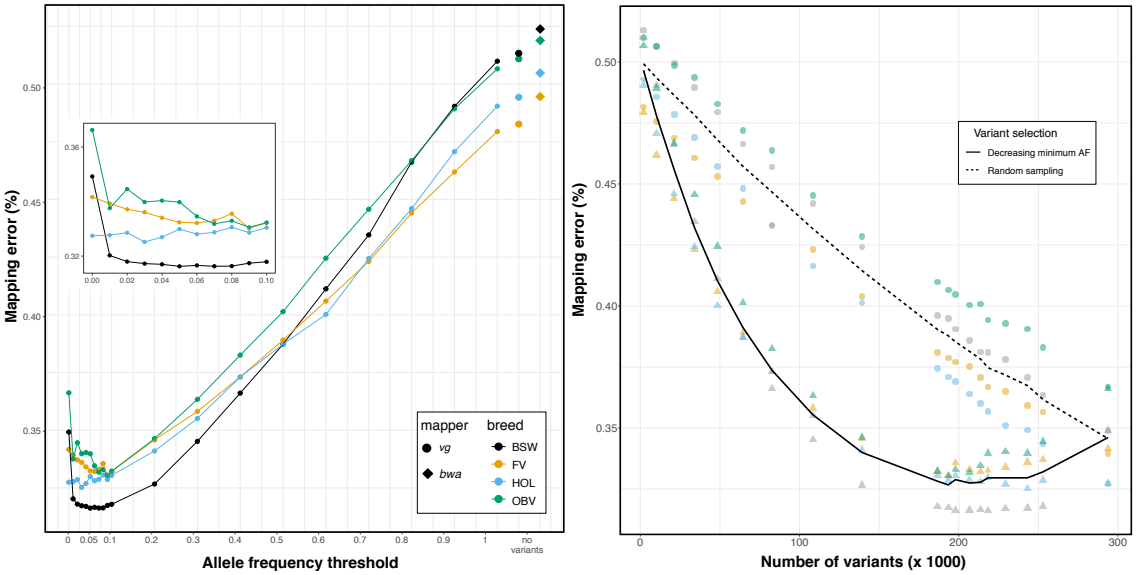


Figure S3.2: **Single-end mapping accuracy using genome graphs that contained variants filtered for allele frequency.**
(a) Proportion of incorrectly mapped reads for four breed-specific augmented genome graphs. Diamonds and large dots represent results from linear mapping using *BWA mem* and *vg*, respectively. The inset is a larger representation of the mapping accuracy for alternate allele frequency thresholds less than 0.1. (b) Read mapping accuracy for breed-specific augmented graphs that contained variants that were either filtered for alternate allele frequency (triangles) or sampled randomly (circles) from all variants detected within a breed. The dashed and solid line represents the average proportion of mapping errors across four breeds using variant prioritization and random sampling.

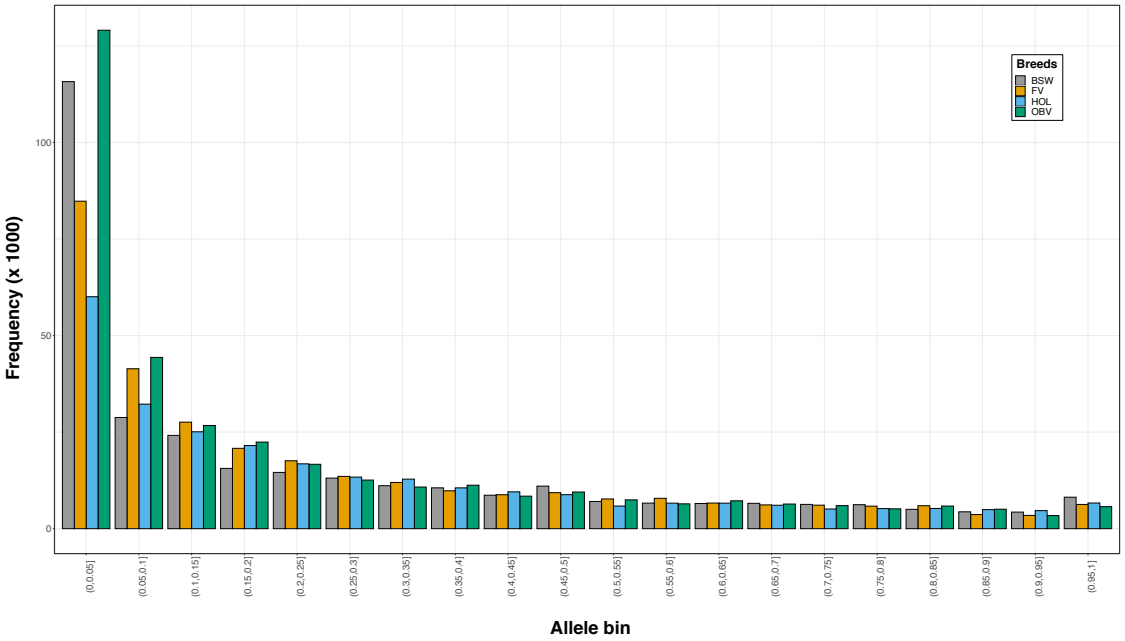


Figure S3.3: **Number of variants detected on chromosome 25 in 82 BSW, 49 FV, 49 HOL and 108 OBV cattle.**
Variants are binned according to allele frequency.

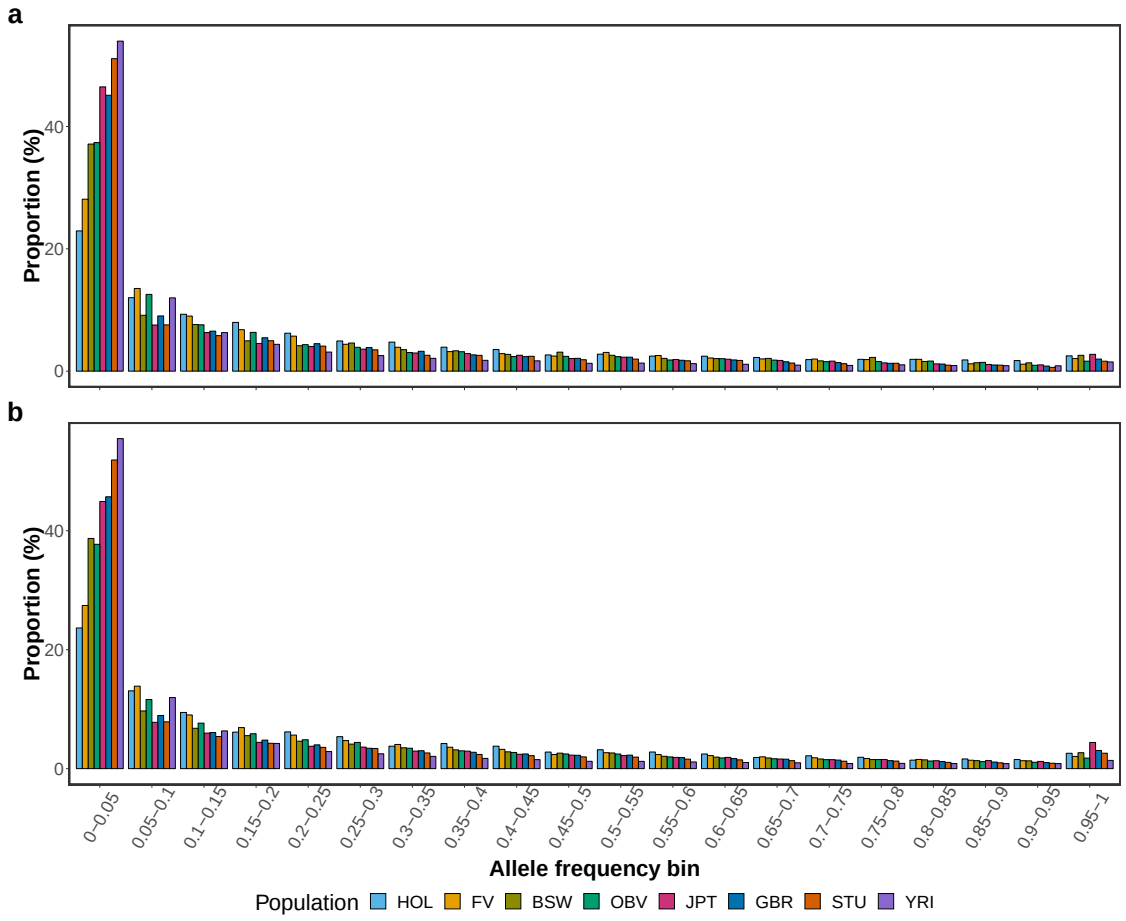


Figure S3.4: **Distribution of alternate allele frequencies in four cattle breeds and four human populations based on (a) bta25 and human chromosome 19 used for graph construction, and (b) whole genome variants.**
The bars indicate the proportion of sequence variants for 20 allele frequency classes. Different colour indicates cattle breeds (HOL, FV, BSW, OBV) and human populations (JPT, GBR, STU, YRI).

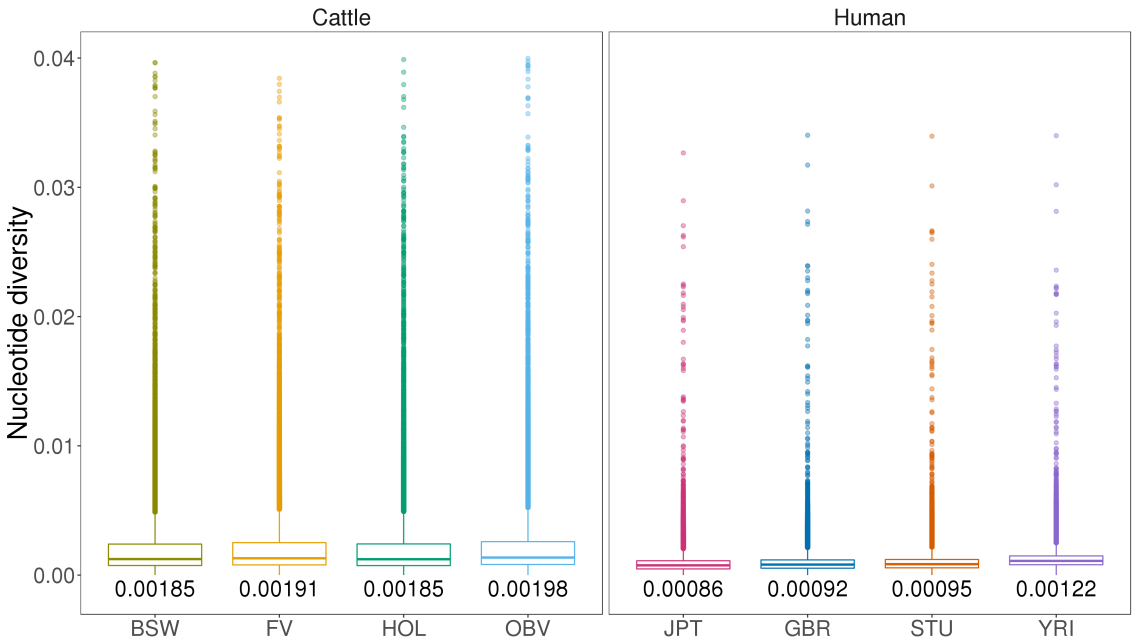


Figure S3.5: **Nucleotide diversity (π) based on whole genome autosomal variants in cattle and human.**
Nucleotide diversity (π) from each population calculated using vcftools with 10 kb non-overlapped windows based on whole genome autosomal variants. Number under the boxplot indicates average across windows.

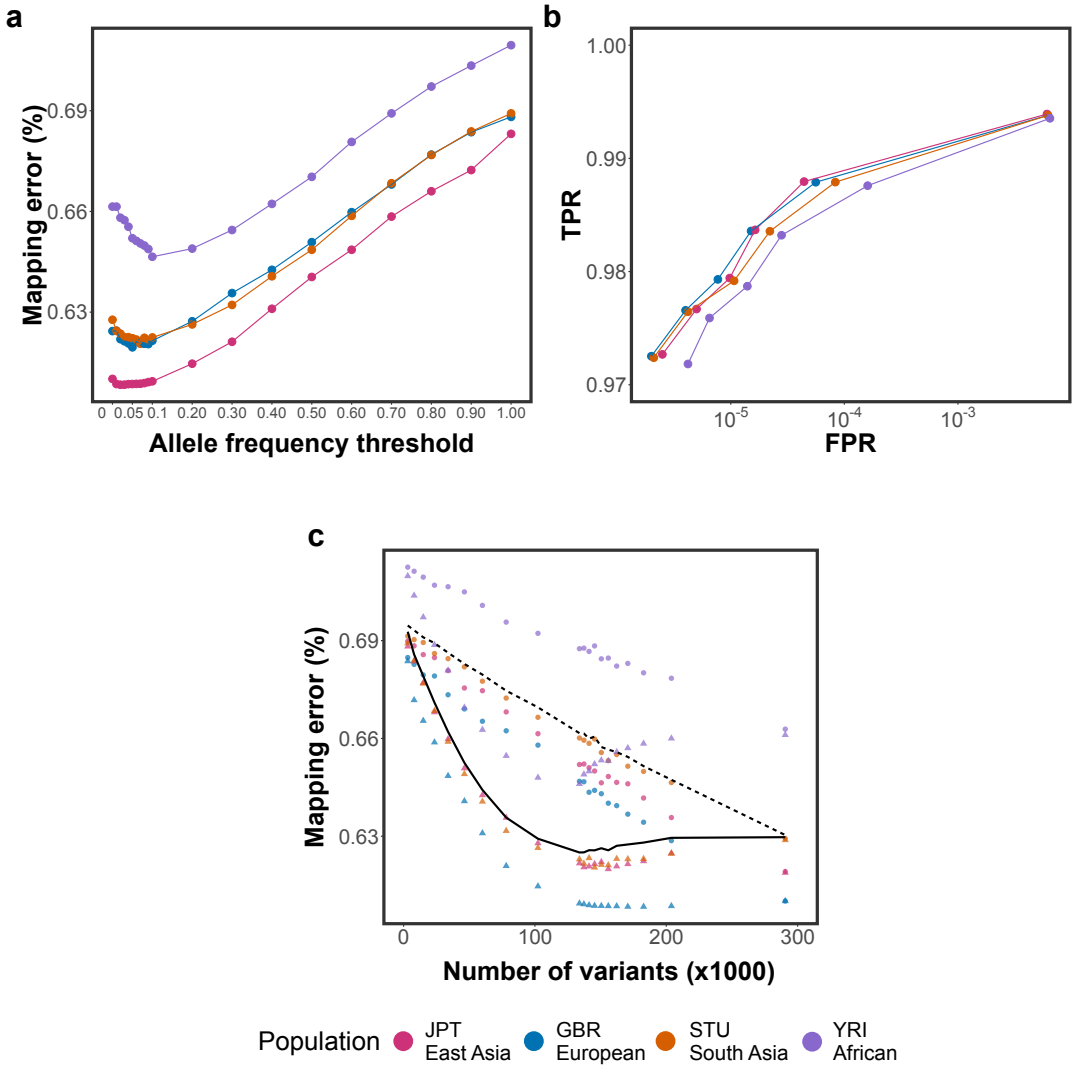


Figure S3.6: **Single-end mapping accuracy using four human population-specific augmented graphs.**
(a) Proportion of incorrectly mapped reads for four populationspecific augmented genome graphs (b) True positive (sensitivity) and false positive mapping rate (specificity) parameterized based on the mapping quality for the best performing graph from each population. (c) Read mapping accuracy for population specific augmented graphs that contained variants that were either filtered for alternate allele frequency (triangles) or sampled randomly (circles) from all variants detected within a population. The dashed and solid line represents the average proportion of mapping errors across four populations using variant prioritization and random sampling.

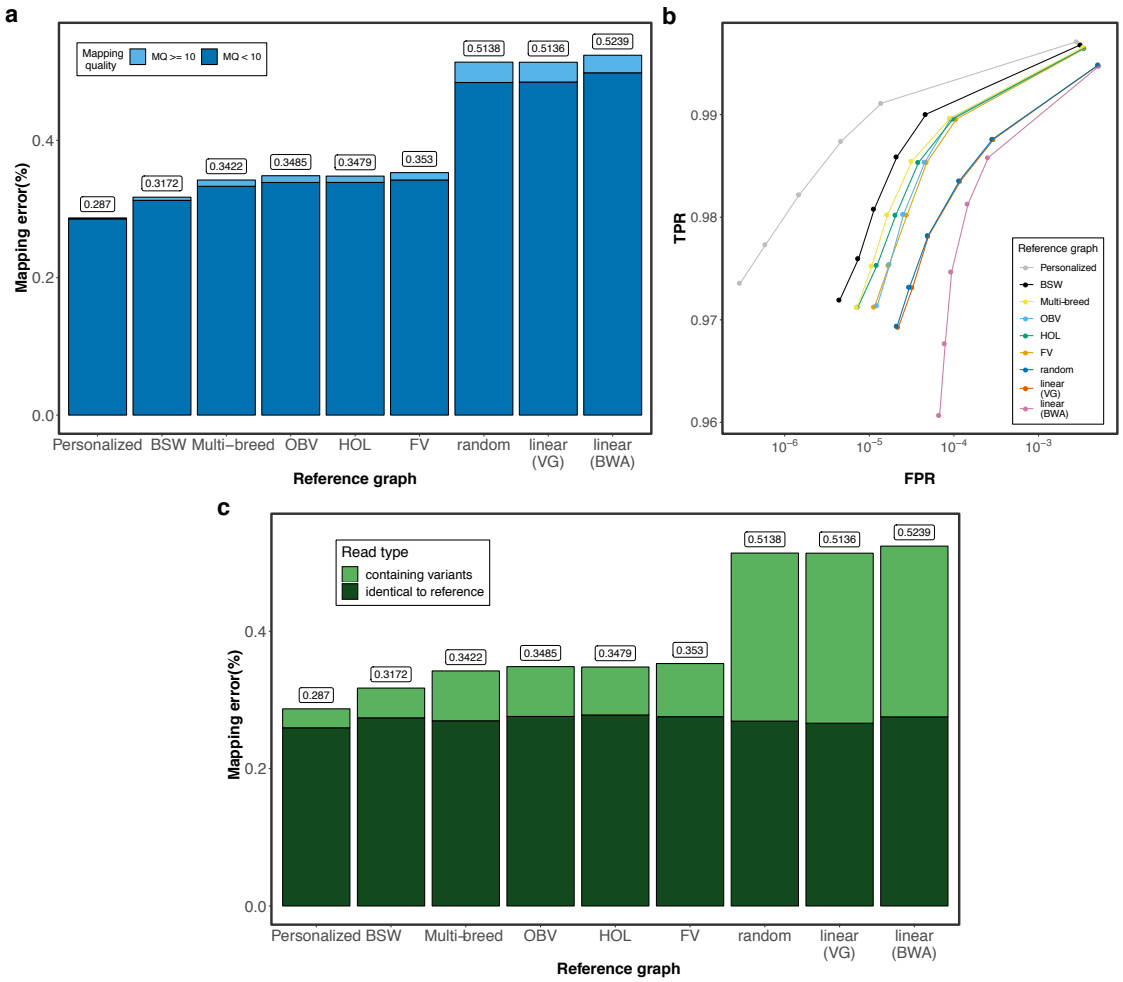


Figure S3.7: **The accuracy of mapping simulated BSW single-end reads to variation-aware and linear reference structures.**
(a) Proportion of BSW single-end reads that mapped erroneously against breed-specific augmented graphs, random graphs or linear reference sequences. Dark and light blue colours represent the proportion of incorrectly mapped reads with mapping quality (MQ)<10 and MQ>10, respectively. (b) True positive (sensitivity) and false positive mapping rate (specificity) parameterized based on the mapping quality. (c) Dark and light green colours represent the proportion of incorrectly mapped reads that matched corresponding reference nucleotides and contained non-reference alleles, respectively

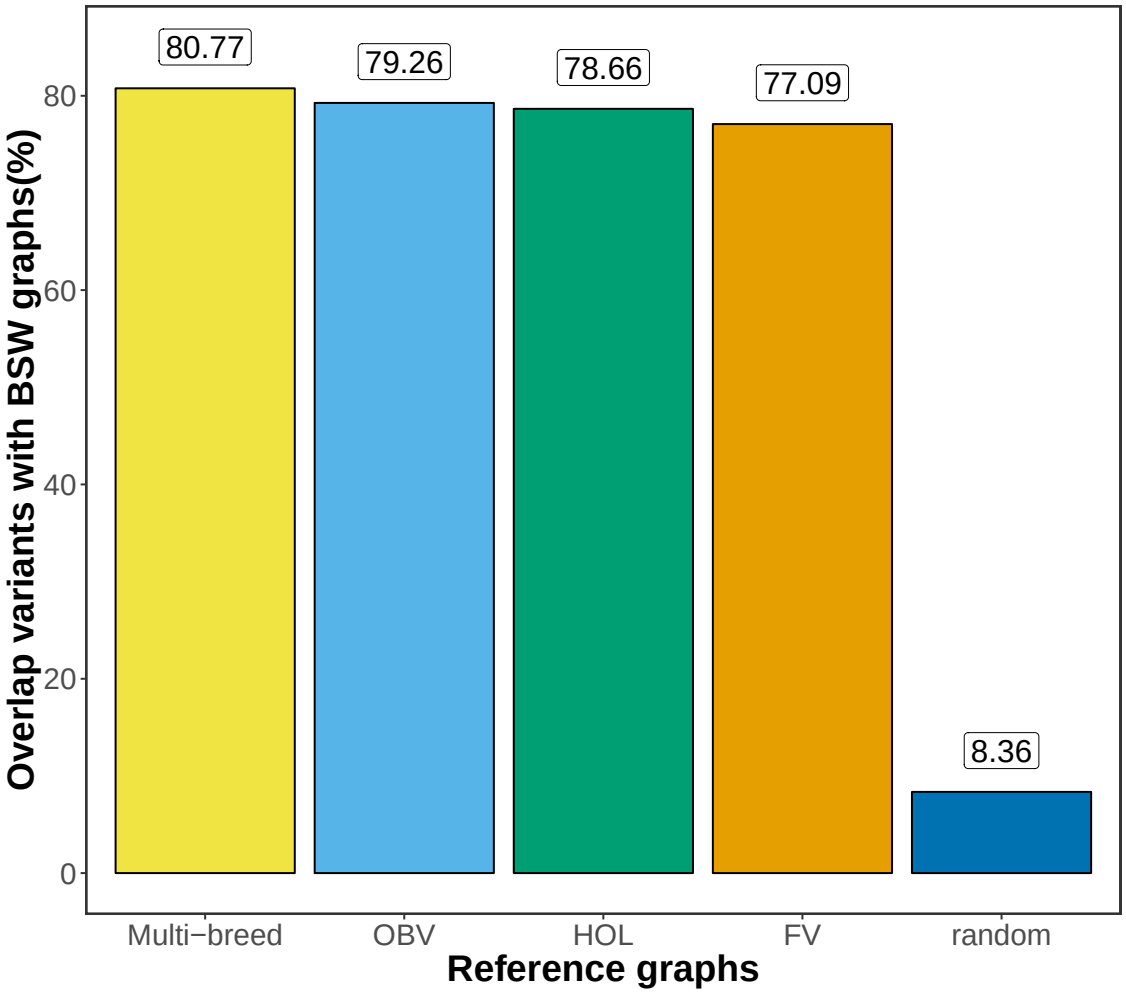


Figure S3.8: **Overlap of the variants**
(N=243,145) between the BSW-and all other variation-aware reference graphs. The values are averaged across 10 replicates.

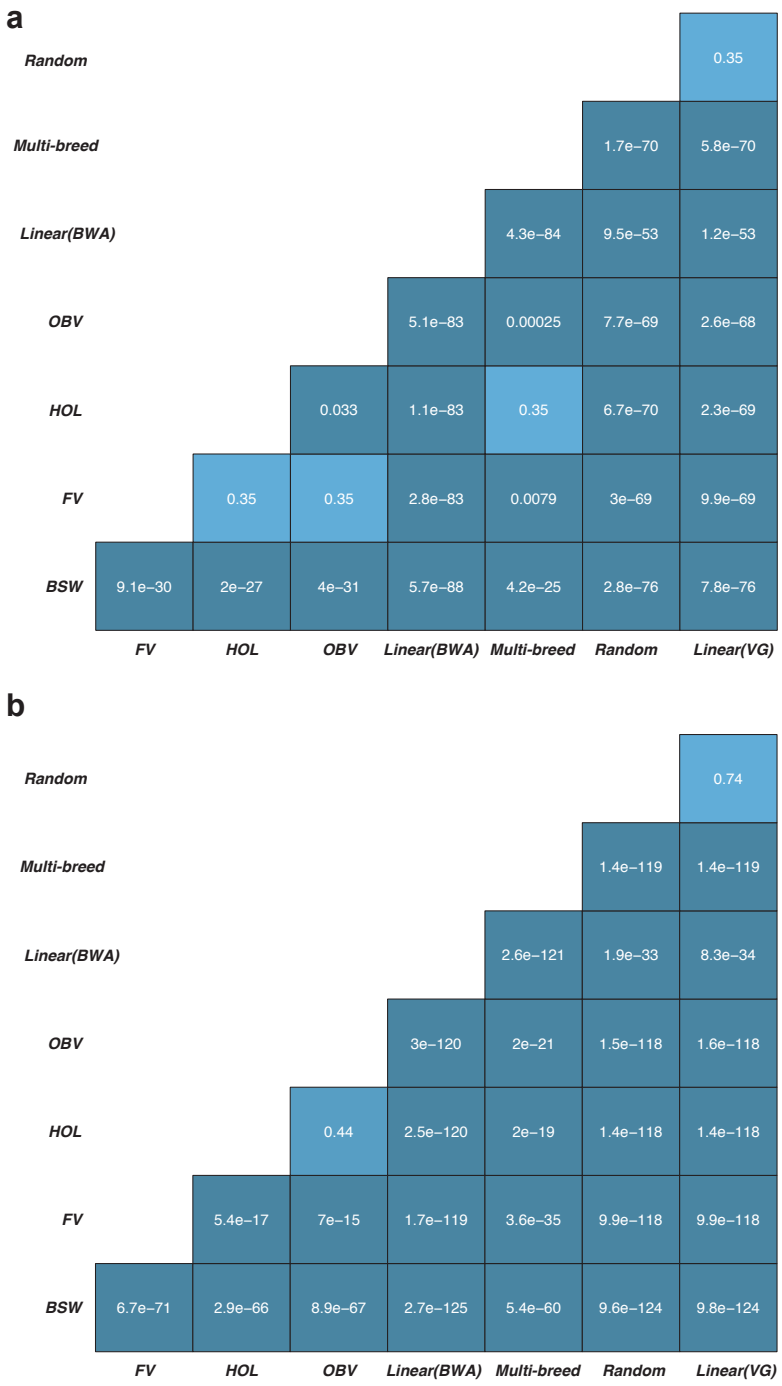


Figure S3.9: **Pairwise heatmap of *P-values* from *t tests*** comparing 8 graph-based mapping scenarios for (a) paired- and (b) single-end reads. The P-values are adjusted for multiple testing using Bonferroni-correction.

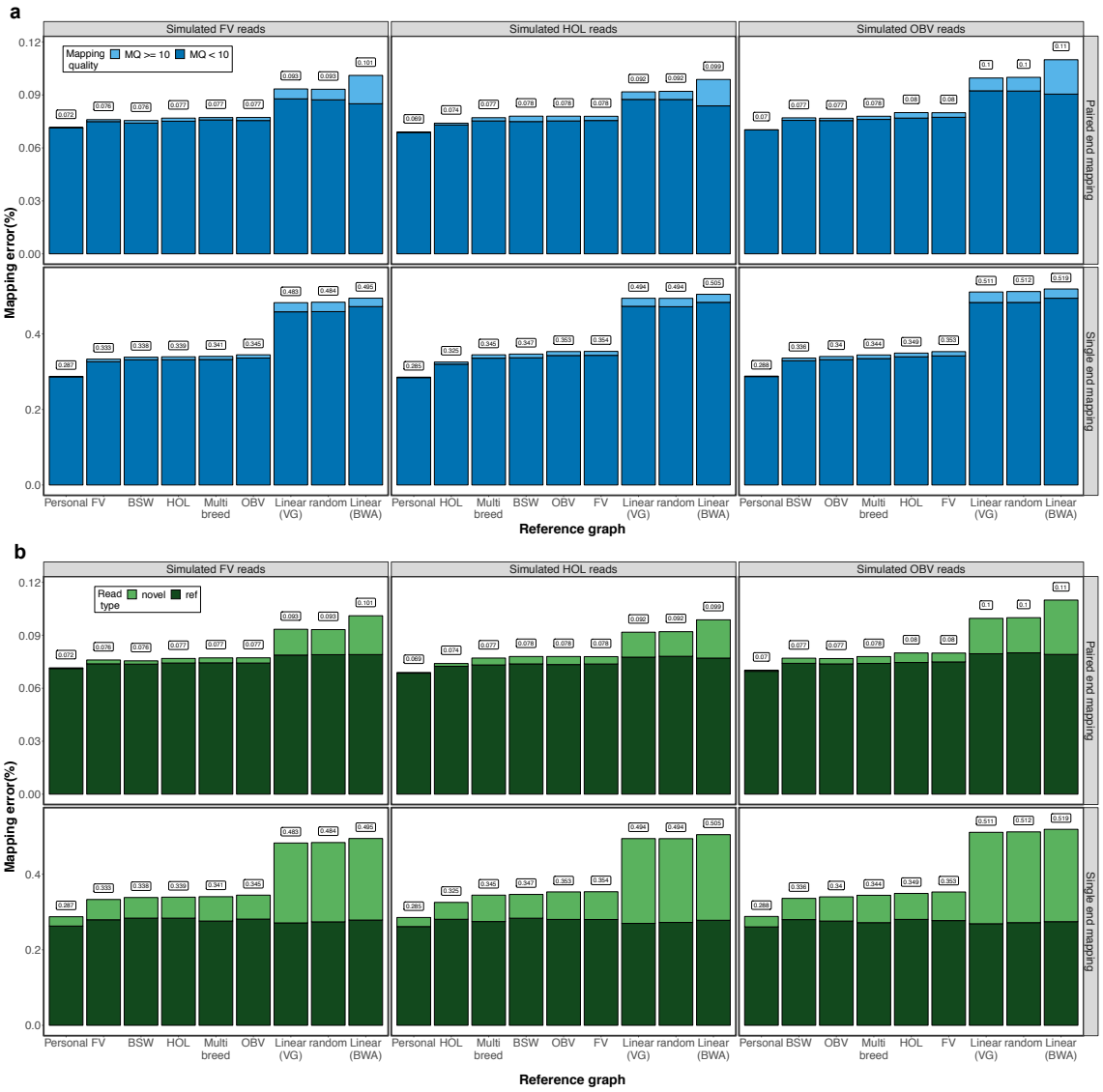


Figure S3.10: **The accuracy of mapping simulated FV, HOL and OBV reads to variation-aware and linear reference structures.**
(a) Proportion of reads that mapped erroneously against personalized graphs, breed-specific augmented graphs, random graphs or linear reference sequences. Dark and light blue colours represent the proportion of incorrectly mapped reads with mapping quality (MQ)<10 and MQ>10, respectively. The upper and lower panels reflect paired-end and single-end reads, respectively. (b) Dark and light green colours represent the proportion of incorrectly mapped reads that matched corresponding reference nucleotides and contained non-reference alleles, respectively. The upper and lower panels reflect paired-end and single-end reads, respectively

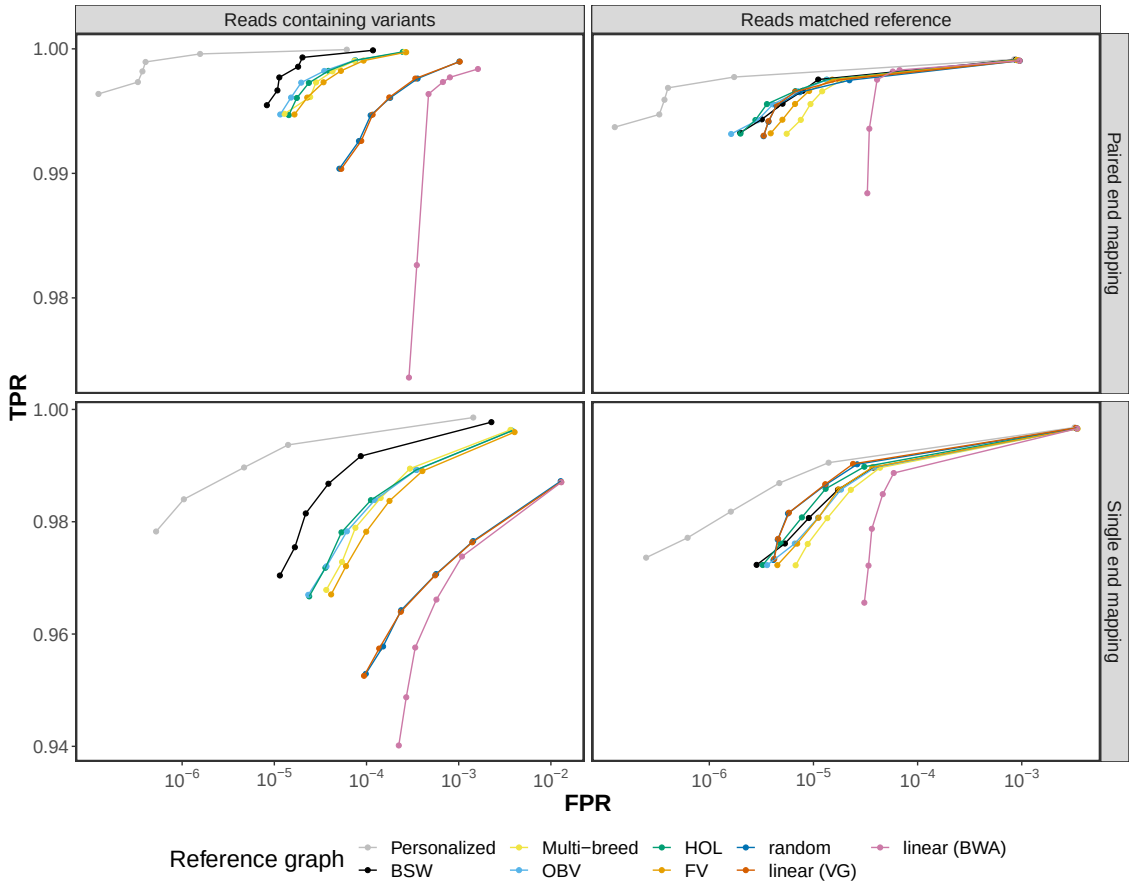


Figure S3.11: **ROC curves split by read's novelty**
Cumulative *True positive* and *False positive rate* at different mapping quality thresholds visualized as Receiver Operating Characteristic (ROC) curves for reads than contain variants and match corresponding reference alleles. The upper and lower panels represent results from paired- and single-end reads.

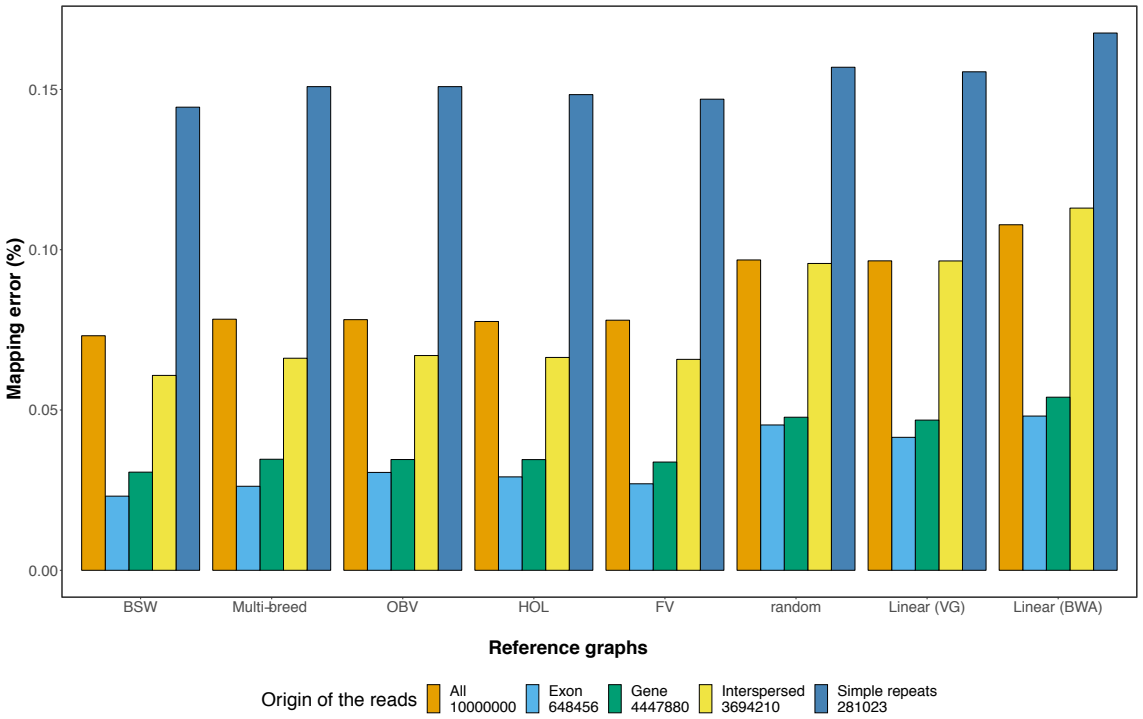


Figure S3.12: **Mapping accuracy for reads originating from different genomic features.**
The origin of 10 million simulated reads was determined based on the Bos taurus ARS-UCD1.2 ensembl 99 annotations (exonic and genic) and the ARS-UCD1.2 repeat regions labelled by Repeat Masker (Interspersed duplications including SINEs, LINEs, LTR, and DNA transposable elements, and simple repeats which contain low-complexity and simple repetitive regions). Different colour indicates the proportion of erroneously mapped reads for each annotation category. The orange bars represent the average proportion of mismapped reads for six graph-based (BSW, Multi-breed, OBV, HOL, FV, random) and two linear (VG, BWA) reference structures. Reads were simulated from haplotypes of a BSW individual.

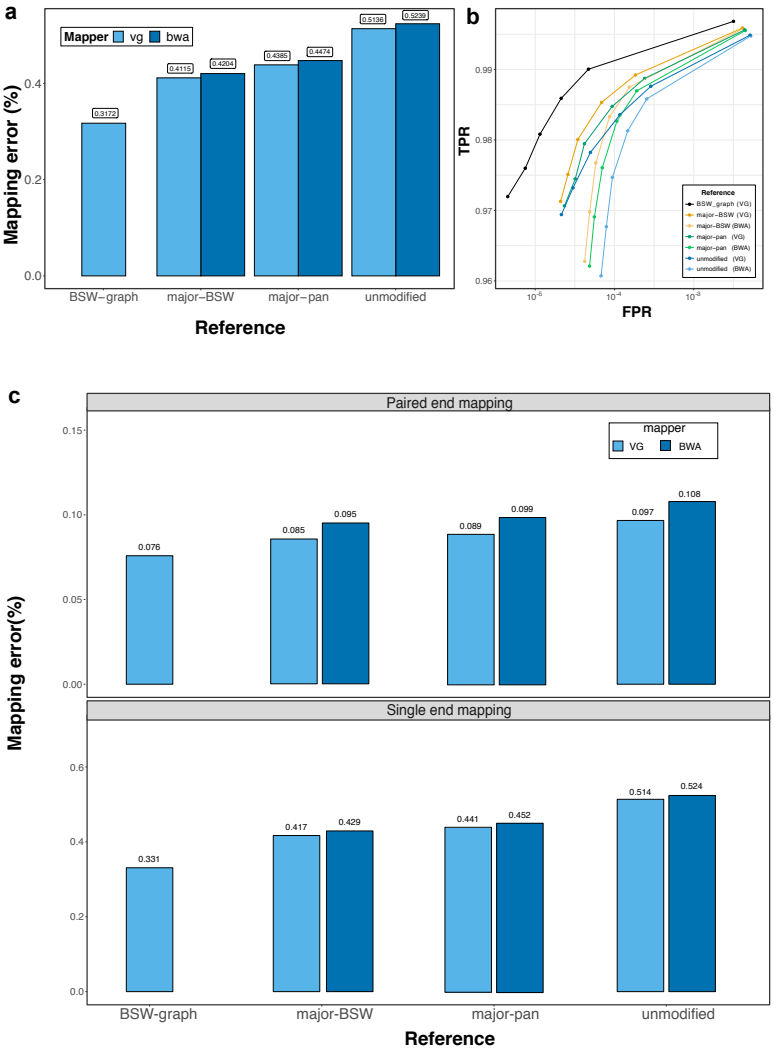


Figure S3.13: **Single-end read mapping accuracy using breed-specific augmented genome graphs and consensus linear reference sequences.**
(a) Dark and light blue represent the proportion of reads that mapped incorrectly using *BWA mem* and *vg*, respectively, to the BSW-specific augmented reference graph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

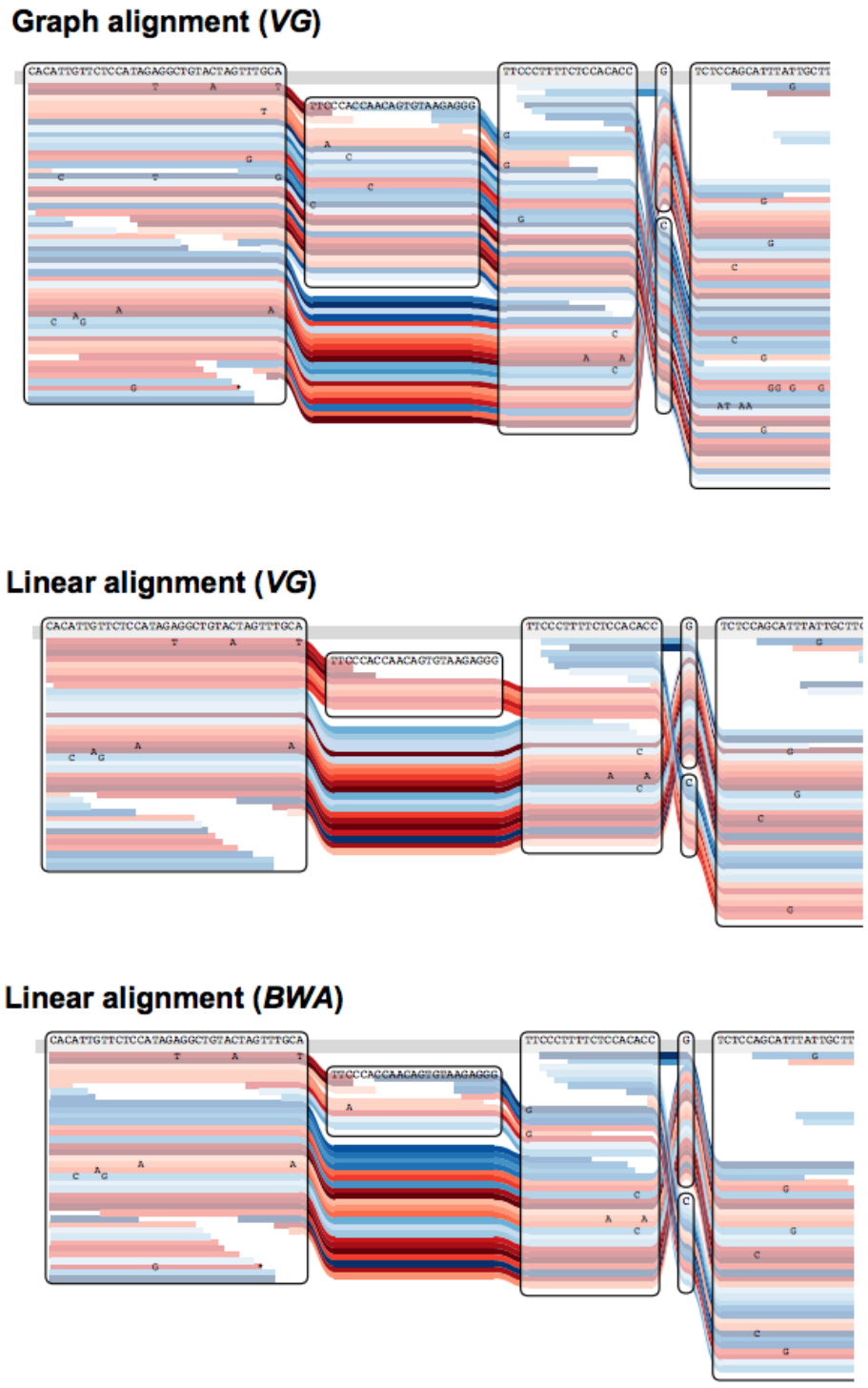


Figure S3.14: **Graph alignment visualization.**
Visualization of a 23-bp insertion at Chr10: 5,941,270 in graph and linear alignments using the *sequence tube map* tool (Beyer et al., 2019). The variant was called heterozygous from the linear alignment, but the allelic ratio was highly biased towards the reference allele. Visual inspection suggests that more reads supporting the alternate allele are present in the graph alignments. Red and blue colour indicates forward and reverse reads, respectively. The reads from the linear alignment were realigned to the variation-aware graph for the purpose of the visualisation.

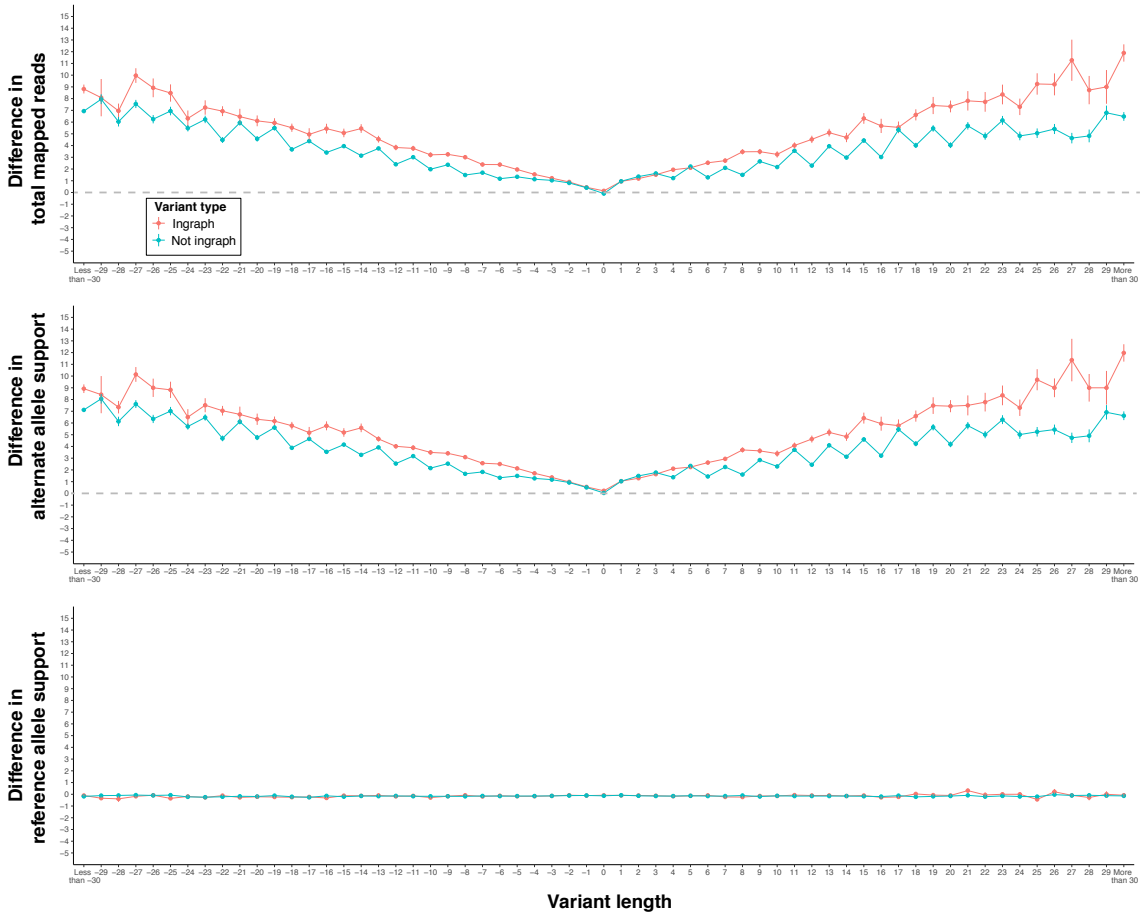


Figure S3.15: **Difference in the total of mapped reads, and reads support for reference and alternate alleles**
between the graph-based and BWA alignments for deletions, SNPs and insertions. Positive values indicate a larger number of reads for graph-based alignments. The dashed grey line indicates equal support for graphbased and linear alignments. The circles represent the mean (± standard error of mean) values at a given variant length. Red and green colour indicates that the alternate allele is included and not included in the graph, respectively.

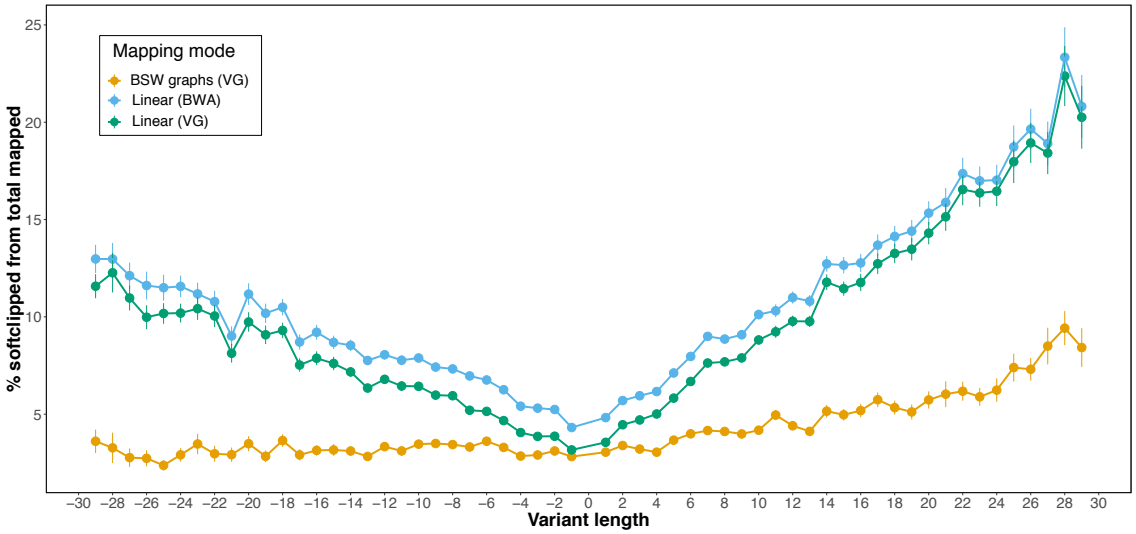


Figure S3.16: **Proportion of soft-clipped reads at heterozygous sites in graph (*vg*) and linear (*vg* and *BWA*) alignments.**
We considered only variants for which the alternate allele was already included in the graph. The circles represent the mean (± standard error of mean) values at a given variant length.

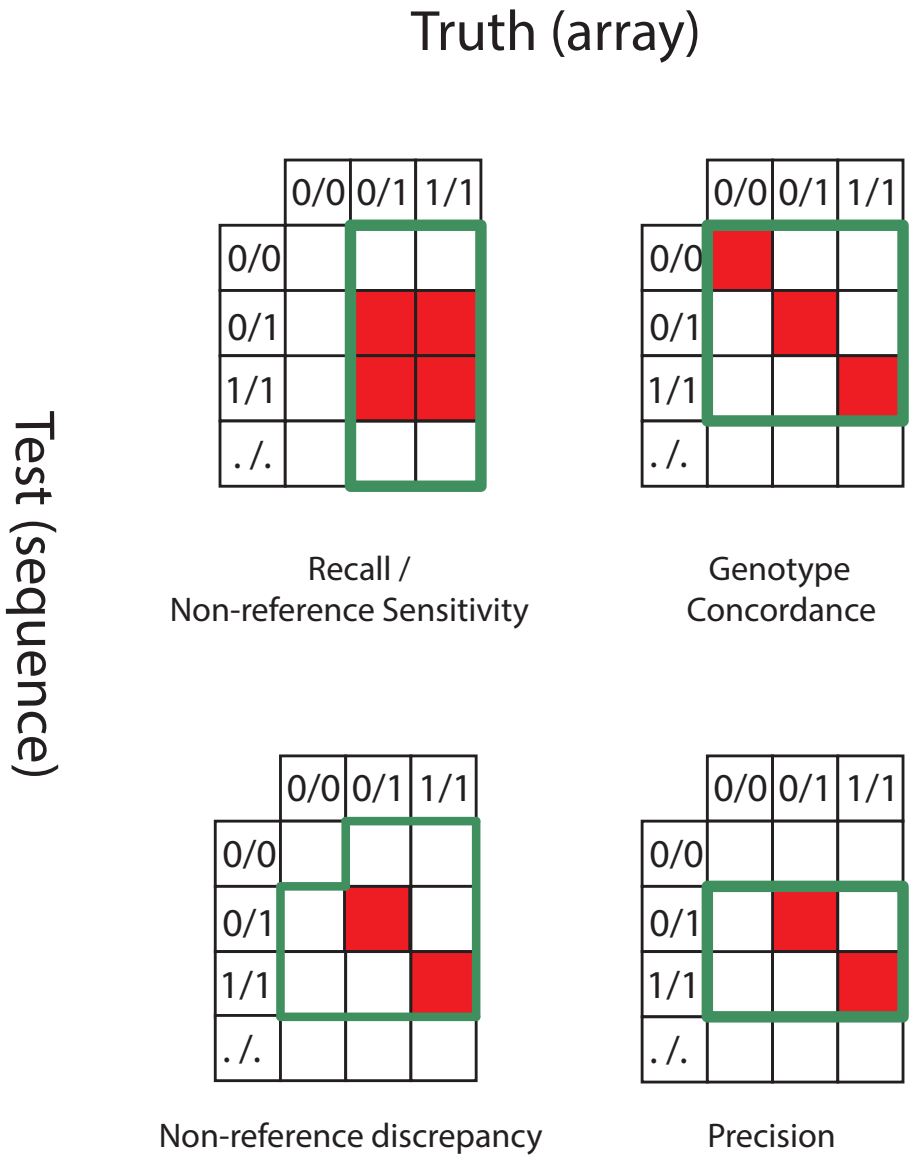


Figure S3.17: **Genotype concordance matrices for four quality parameters.** For each metric, we divided the sum of the red cells by the sum of the cells within the green frame.

Note S3.1

Comparison of variant prioritization approaches
We applied FORGe (Pritt et al., 2018) to prioritize variants to be added to the Brown Swiss reference graph for chromosome 25. Specifically, we considered the four variant ranking approaches implemented in FORGe and compared the mapping accuracy from the resulting graphs with a graph that was constructed with variants selected based on an allele frequency threshold.

The following prioritization approaches were investigated:

1. Pop Cov: variants ranked based on allele frequency
2. Pop Cov + blowup: variants ranked based on allele frequency and proximity (variants that are nearby receive lower scores)
3. Hybrid: variants ranked based on allele frequency and how the variants affect the resulting k-mer profile of the genome graph (variants that would increase the repetiveness of the resulting graph receive lower scores)
4. Hybrid + blowup: hybrid methods + considering variant proximity
5. AF threshold: variants ranked based on allele frequency (AF, as applied in our paper).

We refer to the FORGe paper (Pritt et al., 2018) for a detailed description on the implementation of the variant prioritization methods 1-4. For each prioritization approach, we constructed a number of graphs that included the top x% of the ranked variants, where x ranged from 1 to 100 with steps of 10 (e.g., a graph constructed with x=10 included 34,715 out of 347,147 bta25 Brown Swiss variants). We then mapped paired-end reads simulated form a Brown Swiss animal (as detailed in the Material and Methods part of the main manuscript) to the graphs in order to calculate mapping accuracy.

Graphs constructed with variants that were prioritized solely using allele frequency (as applied in our current paper and the Pop Cov method of FORGe) enable the most accurate mapping of reads (Table SN31 and Figure SN31). Considering additional factors other than allele frequency did not lead to further accuracy improvements. The mapping accuracy of the Pop Cov and AF threshold strategies was virtually identical when the same number of variants was used. The most accurate Pop Cov approach corresponds to an alternate allele frequency threshold of 0.06.

Table SN31: Comparison of the most accurate graph from each ranking method

Ranking methods	Minimum mapping error	Number of variants in the graphs with maximum accuracy
PopCov	0.0722	208288
PopCov + blowup	0.0730	208288
Variant frequency	0.0723	208288
Hybrid	0.0749	347147
Hybrid + blowup	0.0749	347147

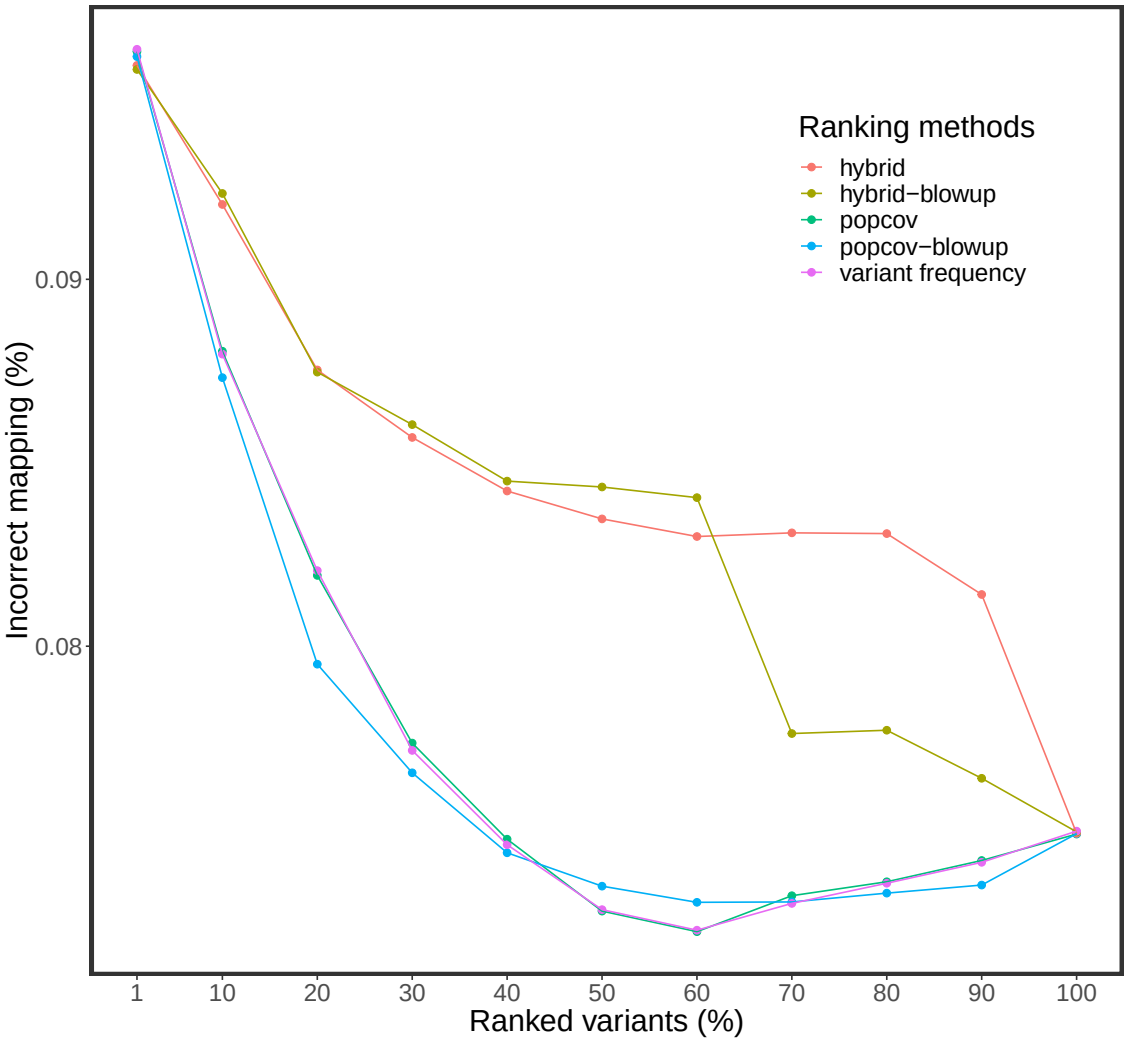


Figure SN31: Comparison of different variant prioritization strategies. Proportion of incorrectly mapped reads for graphs constructed with five variant prioritization approaches.

Note S3.2

Adjusted (tuned) linear mapping approach

We followed the proposed approach outlined by (Grytten et al., 2020) to adjust the default parameters of *BWA mem* in order to also consider sub-optimal alignments. First, we reduce the D value (default 0.5) to consider more alternative alignment positions. However, the mapping performance changed only marginally.

Second, we ran *Minimap2* in short read mode (-ax sr) to find all suboptimal alignments. Subsequently, we retained for each read the read placement from either *BWA mem* or *Minimap2* that had the higher alignment score. For reads that had identical alignment score and position for both linear mappers, we retained the lower mapping quality score. For all other cases, we retained the *BWA mem* alignment.

We made two observations (Figure SN32):

1. The overall mapping accuracy increased mainly due to a smaller number of incorrectly placed reads that had high mapping quality (MQ $>$ 10). This indicates that the tuned linear mapping approach assigns the quality of the alignments better.
2. We found an improvement in mapping accuracy only on reads that are identical to the reference, but not on reads that contain variants.

While Grytten et al. observed that an adjusted parameter setting of *BWA mem* and subsequent application of *Minimap2* led to considerable accuracy improvements, the gain in accuracy was low in our study. The proportion of simulated reads with variants was twice as high (19.16% vs. 10.6%) in our study than in Grytten et al., because the average number of polymorphic sites per genome was almost two-fold higher in cattle than humans.

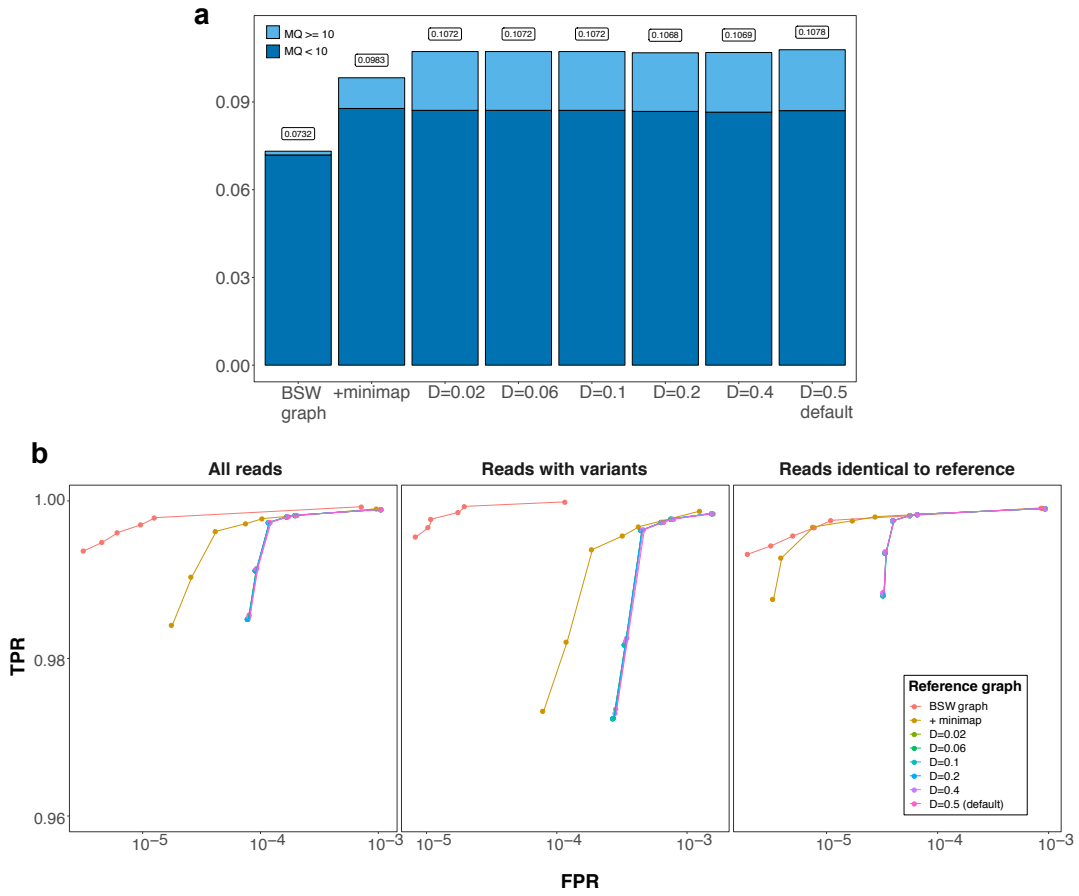


Figure SN32: Mapping accuracy of paired-end reads simulated form a Brown Swiss animal using different mapping approaches.
(a) Proportion of simulated reads with mapping errors for different mapping scenarios. (b) True positive and false positive rate parameterized on mapping quality for the different scenarios.

Note S3.3

Integrating structural variants into the graphs

We investigated the effect of including longer (structural) variants. For this purpose, we first called and genotyped structural variants using *Delly* (Rausch et al., 2012) from 82 Brown Swiss samples that had been sequenced using short-reads (see Material and Methods part of the main manuscript). We discovered 157 precise SVs on bovine chromosome 25 that had an average length of 178 bp. We then combined these variants with 243,145 SNPs and Indels that were discovered using *GATK*. We used the bta25 ARS-UCD1.2 reference as a backbone and constructed four graphs: (i) SNPs (+Indels) from *GATK*, (i) SVs from *Delly*, (iii) SNPs (+Indels) from *GATK* + SVs from *Delly*, (iv) empty (only the backbone, no variants). We simulated 10 million paired end reads from haplotypes of one Brown Swiss animal (SAMEA6272105, that had 121,996 SNPs + Indels and 57 SVs that were included in the graph). The simulated reads were mapped to the different graphs using *vg*.

Table SN32: Mapping accuracy for graphs that contained different variant types
MQ=0 and MQ < 10 indicates the proportion of reads mapped with mapping quality 0 and less than 10, respectively.

Graphs	**Variants in the graphs**	**MQ=0 (%)**	**MQ<10 (%)**	**Mapping error (%)**
Linear	0	0.15474	0.22310	0.08599
SNP	243,145	0.15366	0.21804	0.07995
SV	157	0.15508	0.22390	0.08629
SNP + SV	243,145 + 157	0.15458	0.21900	0.08003

Adding SVs that were detect from short sequencing reads to the graph marginally affected the mapping performance. Actually, the mapping accuracy decreased slightly when SVs were added. Read mapping accuracy improvements were attributable to the SNPs and Indels detected using *GATK*.

Table S3.1: **Properties of autosomal variants detected in human (JPT, GBR, STU, YRI) and bovine (HOL, FV, BSW, OBV) populations**

Species	Population	Number of samples	Variant count	Average per sample	Singleton variants	Variants with allele frequency < 0.05
Human	JPT	104	12,433,397	4,020,815	2,836,542 (22.81%)	5,580,288 (44.88%)
	GBR	91	13,148,448	4,011,102	2,878,144 (21.88 %)	6,005,303 (45.67%)
	STU	102	15,264,479	4,096,457	4,024,478 (26.34%)	7,915,678 (51.85%)
	YRI	108	22,420,039	4,863,955	4,702,120 (20.97%)	12,431,887 (55.45%)
Cattle	HOL	49	16,762,842	6,841,965	1,713,642 (10.22%)	3,964,699 (23.65%)
	FV	49	18,638,951	6,955,100	2,272,546 (12.19%)	5,112,547 (27.42%)
	BSW	82	20,446,693	6,983,517	3,957,703 (19.35%)	7,913,226 (38.70%)
	OBV	104	21,875,164	7,111,562	3,124,950 (14.28%)	8,250,961 (37.71%)

Table S3.2: **Properties of variants detected on human chromosome 19 and bovine chromosome 25 in human (JPT, GBR, STU, YRI) and bovine (HOL, FV, BSW, OBV) populations**

Species	Population	Number of samples	Variant count	Average per sample	Singleton variants	Variants with allele frequency < 0.05
Human	JPT	104	291,303	88,945	66,944 (22.98%)	135,289 (46.44%)
	GBR	91	306,304	90,988	64,119 (20.93 %)	138,076 (45.07%)
	STU	102	355,107	94,253	93,116 (26.22%)	181,300 (51.05%)
	YRI	108	521,021	118,429	106,734 (20.49%)	280,960 (53.92%)
Cattle	HOL	49	295,801	121,114	30,543 (10.32%)	67827 (22.92%)
	FV	49	336,390	125,597	43,783 (13.01%)	94,577 (28.11%)
	BSW	82	347,402	124,209	53,773 (15.47%)	128,990 (37.12%)
	OBV	104	387,855	126,158	47,498 (12.24%)	144,958 (37.37%)

Table S3.3: **Concordance between array-called and sequence variant genotypes that were discovered from either graph or linear alignments using *Samtools*, *GATK*, or *Graphtyper*.**

Numbers represent average values (± standard deviation) of 10 BSW animals for the raw (Full) and hard-filtered (Filtered) genotypes.

	Full			Filtered		
Samtools						
	Graph	Linear	Linear	Graph	Linear	Linear
	VG	VG	BWA	VG	VG	BWA
Genotype concordance	98.50(1.07)	98.47(1.07)	98.53(1.03)	98.53(1.07)	98.50(1.07)	98.55(1.04)
NR-sensitivity (Recall)	98.53(0.37)	98.52(0.39)	98.53(0.39)	97.48(0.36)	97.45(0.35)	97.53(0.36)
NR-discrepancy	2.21(1.60)	2.24(1.60)	2.17(1.55)	2.17(1.60)	2.20(1.61)	2.13(1.56)
Precision	98.90(0.83)	98.89(0.83)	98.93(0.81)	98.91(0.83)	98.90(0.83)	98.94(0.82)
GATK						
Genotype concordance	97.26(2.24)	97.24(2.25)	97.38(2.15)	97.26(2.25)	97.25(2.25)	97.39(2.15)
NR-sensitivity (Recall)	98.17(0.94)	98.16(0.94)	98.23(0.87)	98.14(0.94)	98.12(0.94)	98.18(0.87)
NR-discrepancy	4.09(3.38)	4.10(3.39)	3.89(3.23)	4.08(3.38)	4.09(3.39)	3.88(3.23)
Precision	98.90(0.83)	98.90(0.83)	98.94(0.80)	98.91(0.83)	98.91(0.83)	98.95(0.80)
Graphtyper						
Genotype concordance	98.57(1.01)	98.57(1.01)	98.61(0.97)	98.61(1.03)	98.61(1.03)	98.64(0.99)
NR-sensitivity (Recall)	98.34(0.54)	98.36(0.55)	98.37(0.53)	96.14(0.54)	96.13(0.54)	96.17(0.52)
NR-discrepancy	2.08(1.49)	2.08(1.50)	2.02(1.44)	2.01(1.50)	2.01(1.50)	1.97(1.45)
Precision	98.85(0.80)	98.84(0.81)	98.87(0.79)	98.89(0.82)	98.89(0.82)	98.91(0.80)

Table S3.4: **Accession numbers of the animals** used for variant detection, read simulation, sequence read mapping and genotyping

Accession	Breed	Graph construction	Simulation	Validation genotyping	Validation Mapping bias	Coverage
SAMEA4827645	OBV	x				14.41
SAMEA4827646	OBV	x				12.9
SAMEA4827647	OBV	x				14.79
SAMEA4827648	OBV	x				10.76
SAMEA4827649	OBV	x				11.55
SAMEA4827650	OBV	x				10.29
SAMEA4827651	OBV	x				14.76
SAMEA4827652	OBV	x				10.65
SAMEA4827653	OBV	x				9.69
SAMEA4827654	OBV	x				10.72
SAMEA4827655	OBV	x				11.32
SAMEA4827656	OBV	x				11.83
SAMEA4827657	OBV	x				8.47
SAMEA4827658	OBV	x				9.69
SAMEA4827659	OBV	x				9.52
SAMEA4827660	OBV	x				10.04
SAMEA4827661	OBV	x				9.68
SAMEA4827662	OBV	x				17.37
SAMEA4827663	OBV	x				11.2
SAMEA4827664	OBV	x				11.29
SAMEA4827665	OBV	x				13.07
SAMEA4827666	OBV	x				11.23
SAMEA4827667	OBV	x				10.99
SAMEA4827668	OBV	x				10.93
SAMEA4827669	OBV	x				12.89
SAMEA4827670	OBV	x				12.18
SAMEA4827671	OBV	x				11.35
SAMEA4827672	OBV	x				10.49
SAMEA4827673	OBV	x				10.31
SAMEA4827674	OBV	x				12.58
SAMEA5059741	OBV	x				4.58
SAMEA5059742	OBV	x				3.76
SAMEA5059743	OBV	x	x			22.33
SAMEA5059744	OBV	x				3.93
SAMEA5059745	OBV	x				4.31
SAMEA5059746	OBV	x				4.29
SAMEA5059747	OBV	x				4.58
SAMEA5059748	OBV	x				5.08
SAMEA5059749	OBV	x				5.19
SAMEA5059750	OBV	x				3.91
SAMEA5059751	OBV	x				5.59
SAMEA5059752	OBV	x				3.89
SAMEA5059753	OBV	x				4.18
SAMEA5059754	OBV	x				3.49
SAMEA5059755	OBV	x				7.49
SAMEA5059756	OBV	x				6.65
SAMEA5059757	OBV	x				5.74
SAMEA5059758	OBV	x				5.1

Continuation of Table S3.4

Accession	Breed	Graph construction	Simulation	Validation genotyping	Validation Mapping bias	Coverage
SAMEA6272117	OBV	x				6.43
SAMEA5059759	OBV	x				3.97
SAMEA5159792	BSW	x				10.68
SAMEA5159791	BSW	x				10.22
SAMEA5159788	BSW	x				10.71
SAMEA5159783	BSW	x				11.91
SAMEA5159785	BSW	x				11.94
SAMEA5159799	BSW	x				10.25
SAMEA5159787	BSW	x				13.63
SAMEA5159761	BSW	x				16.46
SAMEA5159782	BSW	x				11.47
SAMEA5159775	BSW	x				10.14
SAMEA5159786	BSW	x				12.04
SAMEA5159784	BSW	x				11.88
SAMEA5159798	BSW	x				12.79
SAMEA5159781	BSW	x				12.65
SAMEA5159780	BSW	x				12.41
SAMEA5159777	BSW	x				9.8
SAMEA5159797	BSW	x				11.98
SAMEA5159774	BSW	x				9.46
SAMEA5159769	BSW	x				12.3
SAMEA5159778	BSW	x				13.03
SAMEA5159771	BSW	x				10.92
SAMEA5159779	BSW	x				10.63
SAMEA5159772	BSW	x				11.88
SAMEA5159773	BSW	x				10.77
SAMEA5159793	BSW	x				12.6
SAMEA5159770	BSW	x				10.01
SAMEA5159795	OBV	x				12.58
SAMEA5159768	OBV	x				8.69
SAMEA5159796	OBV	x				11.39
SAMEA5159789	OBV	x				10.27
SAMEA5159790	OBV	x				10.52
SAMEA5159794	OBV	x				11.46
SAMEA5159776	OBV	x				9.71
SAMEA5159767	OBV	x				10.17
SAMN05216093	OBV	x				10.85
SAMN05216095	OBV	x				11.12
SAMN05216094	OBV	x				10.64
SAMN05216096	OBV	x				11.51
SAMEA6272131	FV	x				13.4
SAMEA6272130	FV	x				10.41
SAMEA4644727	BSW	x				14.86
SAMEA4644728	BSW	x				14.86
SAMEA19864918	BSW	x				9.23
SAMEA4644765	BSW	x				12.14
SAMEA4644766	BSW	x				16.48
SAMEA4644768	OBV	x				13.41
SAMEA4644769	BSW	x				16.04
SAMEA19312918	BSW	x				4.43

Continuation of Table S3.4

Accession	Breed	Graph construction	Simulation	Validation genotyping	Validation Mapping bias	Coverage
SAMEA19313668	BSW	x				7.13
SAMEA19314418	BSW	x				10.99
SAMEA19315168	BSW	x				9.7
SAMEA19318918	BSW	x				6.9
SAMEA19323418	BSW	x				18.83
SAMEA4644754	BSW	x				15.25
SAMEA4644755	BSW	x				13.58
SAMEA4644756	BSW	x				13.88
SAMEA4644730	OBV	x				14.85
SAMEA4644734	OBV	x				15.3
SAMEA4644735	BSW	x				9.43
SAMEA4644757	BSW	x				11.36
SAMEA4644739	BSW	x				14.13
SAMEA4644740	OBV	x				15.73
SAMEA4644741	BSW	x				15.57
SAMEA4644742	BSW	x				15.68
SAMEA4644758	BSW	x				13
SAMEA4644743	BSW	x				15.46
SAMEA4644749	OBV	x				13.85
SAMEA4644750	OBV	x				15.25
SAMEA4644762	BSW	x				13.92
SAMEA4644763	BSW	x				11.62
SAMEA4644764	OBV	x				10.57
SAMN07692225	BSW	x				10.72
SAMN02671625	FV	x				5.06
SAMN02671626	FV	x	x			23.24
SAMN02671627	FV	x				6.32
SAMN02671628	FV	x				4.95
SAMN02671629	FV	x				8.41
SAMN02671630	FV	x				4.88
SAMN02671631	FV	x				4.77
SAMN02671632	FV	x				7.64
SAMN02671633	FV	x				3.59
SAMN02671634	FV	x				7.67
SAMN02671635	FV	x				6.37
SAMN02671636	FV	x				6.26
SAMN02671637	FV	x				3.79
SAMN02671638	FV	x				3.95
SAMN02671639	FV	x				7.21
SAMN02671640	FV	x				8.62
SAMN02671641	FV	x				6.08
SAMN02671642	FV	x				5.47
SAMN02671643	FV	x				5.03
SAMN02671644	FV	x				4.35
SAMN02671645	FV	x				5.06
SAMN02671646	FV	x				5.79
SAMN02671647	FV	x				5.2
SAMN02671648	FV	x				5.81
SAMN02671649	FV	x				5.32
SAMN02671650	FV	x				5.34

Continuation of Table S3.4

Accession	Breed	Graph construction	Simulation	Validation genotyping	Validation Mapping bias	Coverage
SAMN02671651	FV	x				4.51
SAMN02671652	FV	x				7.48
SAMN02671653	FV	x				7.5
SAMN02671654	FV	x				7.6
SAMN02671655	FV	x				7.19
SAMN02671656	FV	x				5.4
SAMN02671657	FV	x				5.61
SAMN02671658	FV	x				4.91
SAMN02671659	FV	x				4.83
SAMN02671661	FV	x				5.58
SAMN02671662	FV	x				6.08
SAMN02671663	FV	x				5.06
SAMN02671664	FV	x				7.95
SAMN02671665	FV	x				6.53
SAMN02671666	FV	x				6.06
SAMN02671667	FV	x				8.13
SAMN02671572	HOL	x				6.79
SAMN02671574	HOL	x				10.25
SAMN02671576	HOL	x				5.02
SAMN02671578	HOL	x				19.78
SAMN02671580	HOL	x				10.52
SAMN02671582	HOL	x				15.22
SAMN02671584	HOL	x	x			29.97
SAMN02671586	HOL	x				17.21
SAMN02671588	HOL	x				16.99
SAMN02671590	HOL	x				13.79
SAMN02671592	HOL	x				16.31
SAMN02671594	HOL	x				19.56
SAMN02671596	HOL	x				16.43
SAMN02671455	HOL	x				9.23
SAMN02671457	HOL	x				10.28
SAMN02671459	HOL	x				8.4
SAMN02671461	HOL	x				9.47
SAMN02671463	HOL	x				6.36
SAMN02671465	HOL	x				10.61
SAMN02671467	HOL	x				9.78
SAMN02671469	HOL	x				9.13
SAMN02671471	HOL	x				6.49
SAMN02671473	HOL	x				8.71
SAMN02671475	HOL	x				9.57
SAMN02671477	HOL	x				10.89
SAMN02671479	HOL	x				8.81
SAMN02671481	HOL	x				8.59
SAMN02671483	HOL	x				10.79
SAMN02671485	HOL	x				9.18
SAMN02671487	HOL	x				10.1
SAMN02671489	HOL	x				10.06
SAMN02671491	HOL	x				9.83
SAMN02671493	HOL	x				10.1
SAMN02671495	HOL	x				8.58

Continuation of Table S3.4

Accession	Breed	Graph construction	Simulation	Validation genotyping	Validation Mapping bias	Coverage
SAMN02671613	HOL	x				23.58
SAMN02671615	HOL	x				20.36
SAMN02671617	HOL	x				20.36
SAMN02671619	HOL	x				12.54
SAMN02671621	HOL	x				12.86
SAMN02671623	HOL	x				4.73
SAMN02671668	HOL	x				11.92
SAMN02671670	HOL	x				11.35
SAMN02671672	HOL	x				10.21
SAMN02671674	HOL	x				10.4
SAMN02671676	HOL	x				11.21
SAMN02671725	HOL	x				11.54
SAMN02671727	HOL	x				5.43
SAMN02671729	HOL	x				13.68
SAMN02671731	HOL	x				13.58
SAMEA6272085	OBV	x				8.01
SAMEA6272091	OBV	x				9.55
SAMEA6272090	OBV	x				10.74
SAMEA6272089	OBV	x				8.25
SAMEA6272088	OBV	x				10.97
SAMEA6272093	OBV	x				11.3
SAMEA6272087	OBV	x				11.62
SAMEA6272086	OBV	x				12.58
SAMEA6272092	OBV	x				9.38
SAMEA6272094	OBV	x				8.31
SAMEA6272115	OBV	x				8.65
SAMEA6272114	OBV	x				8.06
SAMEA6272112	OBV	x				9.51
SAMEA6272113	OBV	x				10.61
SAMEA6272110	OBV	x				7.99
SAMEA6272103	OBV	x				9.09
SAMEA6272109	OBV	x				7.97
SAMEA6272107	OBV	x				10.34
SAMEA6272102	OBV	x				7.25
SAMEA6272100	OBV	x				8.55
SAMEA6272133	FV	x				12.73
SAMEA6272134	FV	x				10.25
SAMEA6272128	FV	x				11.09
SAMEA6163196	BSW	x				11.48
SAMEA6163197	BSW	x				9.86
SAMEA6163198	BSW	x				11.63
SAMEA6163199	BSW	x				13.68
SAMEA6272129	FV	x				14.9
SAMEA6272132	FV	x				15.25
SAMEA6272119	OBV	x				19.58
SAMEA6272123	OBV	x				16.93
SAMEA6272118	OBV	x				18.66
SAMEA6272120	OBV	x				18.5
SAMEA6272121	OBV	x				16.58
SAMEA6272126	OBV	x				61.9

Continuation of Table S3.4

Accession	Breed	Graph construction	Simulation	Validation genotyping	Validation Mapping bias	Coverage
SAMEA6272124	OBV	x				18.82
SAMEA6272122	OBV	x				18.33
SAMEA6272127	OBV	x				53.65
SAMEA6272125	OBV	x				23.01
SAMEA6272084	OBV	x				11.78
SAMEA6272083	OBV	x				31.95
SAMEA6272082	OBV	x				23.39
SAMEA6272095	BSW	x				25.36
SAMEA6272096	BSW	x				20.6
SAMEA6272097	BSW	x				10.68
SAMEA6272098	BSW	x				15.25
SAMEA6272099	BSW	x				12.32
SAMEA6272101	BSW	x				10.4
SAMEA6272104	BSW	x				12.63
SAMEA6272105	BSW	x	x			33.7
SAMEA6272106	BSW	x				15.76
SAMEA6272108	BSW	x				20.46
SAMEA6272111	BSW	x				28.82
SAMEA6272116	BSW	x				70.04
SAMEA5159861	BSW	x				24.84
SAMEA5159863	BSW	x				23.64
SAMEA5159864	BSW	x				24.92
SAMEA5159865	BSW	x				25.99
SAMEA5159866	BSW	x				25.11
SAMEA5159867	BSW	x				26.28
SAMEA5159868	BSW	x				26.73
SAMEA5159869	BSW	x				27.62
SAMEA5159870	BSW	x				32.64
SAMEA5159871	BSW	x				34.49
SAMEA5159872	BSW	x				27.96
SAMEA5159873	BSW	x				24.08
SAMEA5159874	BSW	x				33.8
SAMEA5159875	BSW	x				22.66
SAMEA5159885	BSW	x				23.1
SAMEA5159837	OBV	x				28.12
SAMEA5159843	OBV	x				22.81
SAMEA5159848	OBV	x				22.5
SAMEA5159849	OBV	x				26.32
SAMEA5159850	OBV	x				27.69
SAMEA5159886	OBV	x				35.51
SAMEA6163185	BSW			x	x	39.88
SAMEA6163188	BSW			x		25.74
SAMEA6163187	BSW			x		20.29
SAMEA6163177	BSW			x		8.26
SAMEA6163178	BSW			x		5.74
SAMEA6163176	BSW			x		9.29
SAMEA6163179	BSW			x		6.93
SAMEA6163183	BSW			x		7.86
SAMEA6163181	BSW			x		7.97
SAMEA6163182	BSW			x		8.36

Supplementary References

W. Beyer, A. M. Novak, G. Hickey, J. Chan, V. Tan, B. Paten, and D. R. Zerbino. Sequence tube maps: making graph genomes intuitive to commuters. *Bioinformatics*, 35(24):5318, 2019.

I. Grytten, K. D. Rand, A. J. Nederbragt, and G. K. Sandve. Assessing graph-based read mappers against a baseline approach highlights strengths and weaknesses of current methods. *BMC genomics*, 21:1–9, 2020.

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

T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel. Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18): i333–i339, 2012.