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

Danang Crysnanto*, Hubert Pausch

Animal Genomics, ETH Zürich, Zürich, Switzerland

Email:

DC: danang.crysnanto@usys.ethz.ch

HP: hubert.pausch@usys.ethz.ch

*Corresponding author: danang.crysnanto@usys.ethz.ch

Abstract

Background

The current bovine genomic reference sequence was assembled from the DNA of a Hereford cow. The resulting linear assembly lacks diversity because it does not contain allelic variation. Lack of diversity is 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. We show that our novel variation-aware reference facilitates accurate read mapping and unbiased sequence variant genotyping for SNPs and Indels.

Conclusions

We developed the first variation-aware reference graph for an agricultural animal: https://doi.org/10.5281/zenodo.3570312. 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

Introduction

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

Domestication and selection for beef and milk production under various environmental conditions have led to the formation of more than thousand breeds of cattle (*Bos taurus*) with distinct genetic characteristics and high allelic variation within and between breeds [6]. The 1000 Bull Genomes Project discovered almost 100 million sequence variants that are polymorphic in 2'700 cattle from worldwide cattle breeds [7, 8]. Nucleotide diversity is higher in cattle than human populations [7, 9]. Yet, all bovine DNA sequences are aligned to the linear consensus reference sequence of a highly-inbred Hereford cow to facilitate reference-guided variant discovery and genotyping [10, 11]. A genome-wide alignment of DNA fragments from a *B. taurus* individual differs from the Hereford-based reference sequence at between 7 and 8 million single nucleotide polymorphisms (SNPs) and small (<50 bp) insertions and deletions (Indels) [12, 13]. The number of differences is higher for DNA samples with greater divergence from the reference [14, 15].

The bovine linear reference sequence does neither contain allelic variation nor nucleotides that are private to breeds other than Hereford. Lack of allelic diversity is an inherent drawback of linear reference sequences because it causes reference allele bias. DNA sequencing reads that contain only alleles that match corresponding reference nucleotides are more likely to align correctly than DNA fragments that also contain non-reference alleles [16, 17]. Reads originating from DNA fragments that are highly diverged from corresponding reference nucleotides will either obtain low alignment scores, or align at incorrect locations, or remain un-mapped [18]. Reference bias compromises analyses that are sensitive to accurately mapped reads and prevents the precise estimation of allele frequencies [16, 19–21].

Graph-based [17, 22] and personalized reference genomes [5, 23] mitigate reference allele bias. Existing linear reference coordinates can serve as backbones for variation-aware genome graphs. Nodes in the graph represent alleles at sites of variation and edges connect adjacent alleles. Once a variation-aware genome graph contains all alleles at known polymorphic sites, every haplotype can be represented as a walk through the graph [24]. However, an optimal balance between graph density and computational complexity is key to efficient whole-genome graph-based variant analysis because adding sites of variation to the graph incurs computational costs [18]. High nucleotide diversity and the separation of individuals by breeds make cattle an ideally suited species to investigate the optimal composition of genome graphs.

Here, we investigate sequence read mapping and variant genotyping accuracy using variation-aware reference structures in cattle. Using sequence variant genotypes of 288 cattle from four dairy and dual-purpose breeds, we construct breed-specific augmented and pan-genome reference graphs using the *vg toolkit* [22]. We prioritize sequence variants to be added to the graphs and assess accuracy of read mapping for variation-aware and linear references (Figure 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.

Results

Construction of bovine breed-specific augmented genome graphs

Breed-specific augmented reference graphs were constructed for four dairy (Brown Swiss – BSW, Holstein – HOL) and dual-purpose (Fleckvieh – FV, Original Braunvieh – OBV) cattle breeds using the Hereford-based linear reference sequence (ARS-UCD1.2) of chromosome 25 as a backbone. To determine the optimal composition of bovine variation-aware references, we augmented the linear reference 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 2,046 (variants had alternate allele frequency >0.9) and 293,804 (no alternate allele frequency threshold) alleles (Figure 2a).

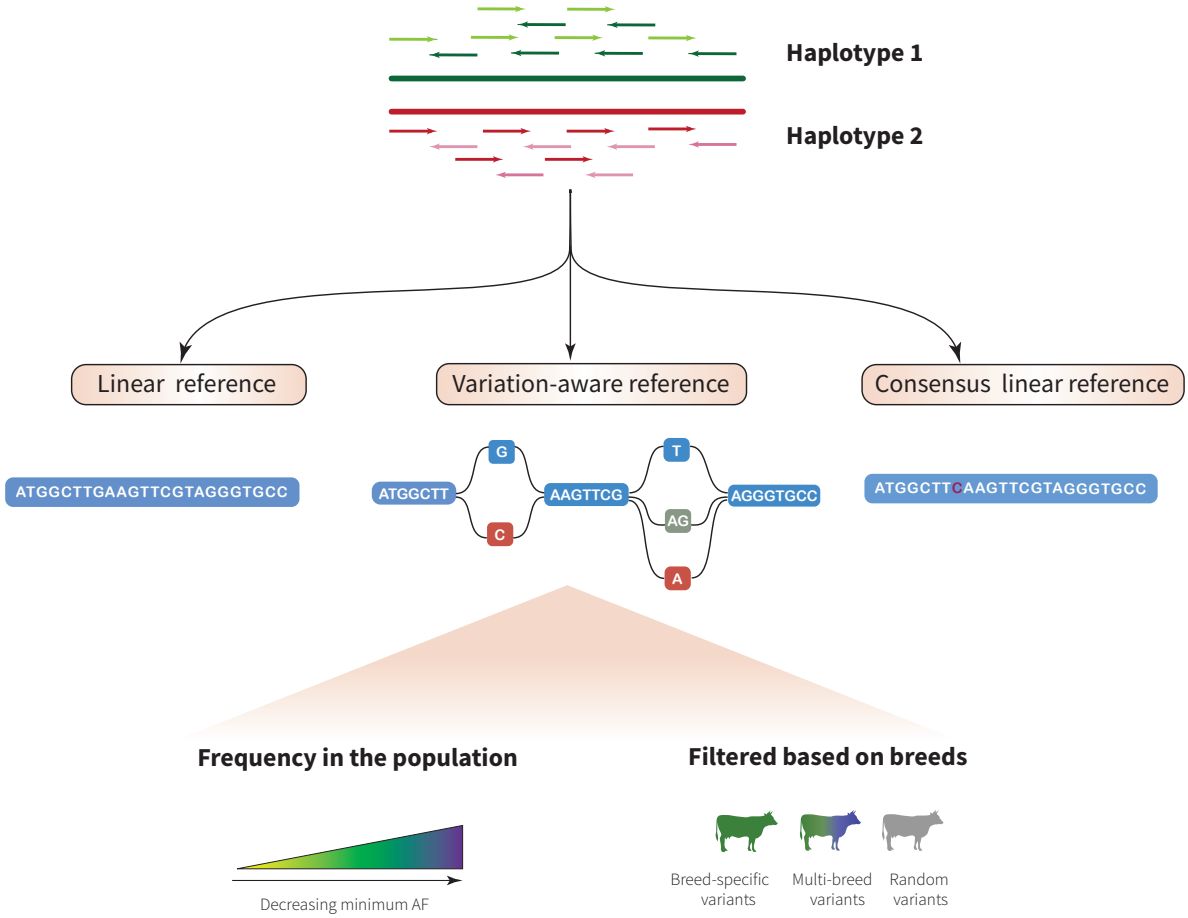

Figure 1. Schematic overview of the construction of breed-specific augmented genome graphs.
We used the *vg toolkit* to augment the bovine linear reference sequence (ARS-UCD1.2) with alleles at SNPs and Indels that were discovered in 288 cattle from four breeds. Alleles that were added to the linear 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. 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.

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 (Figure 2a). When we added a maximum number of 293,804 variants to the linear reference sequence, 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 (Pearson R=0.99), reflecting that the number of alternative paths increases disproportionately to the number of variants in the graph.

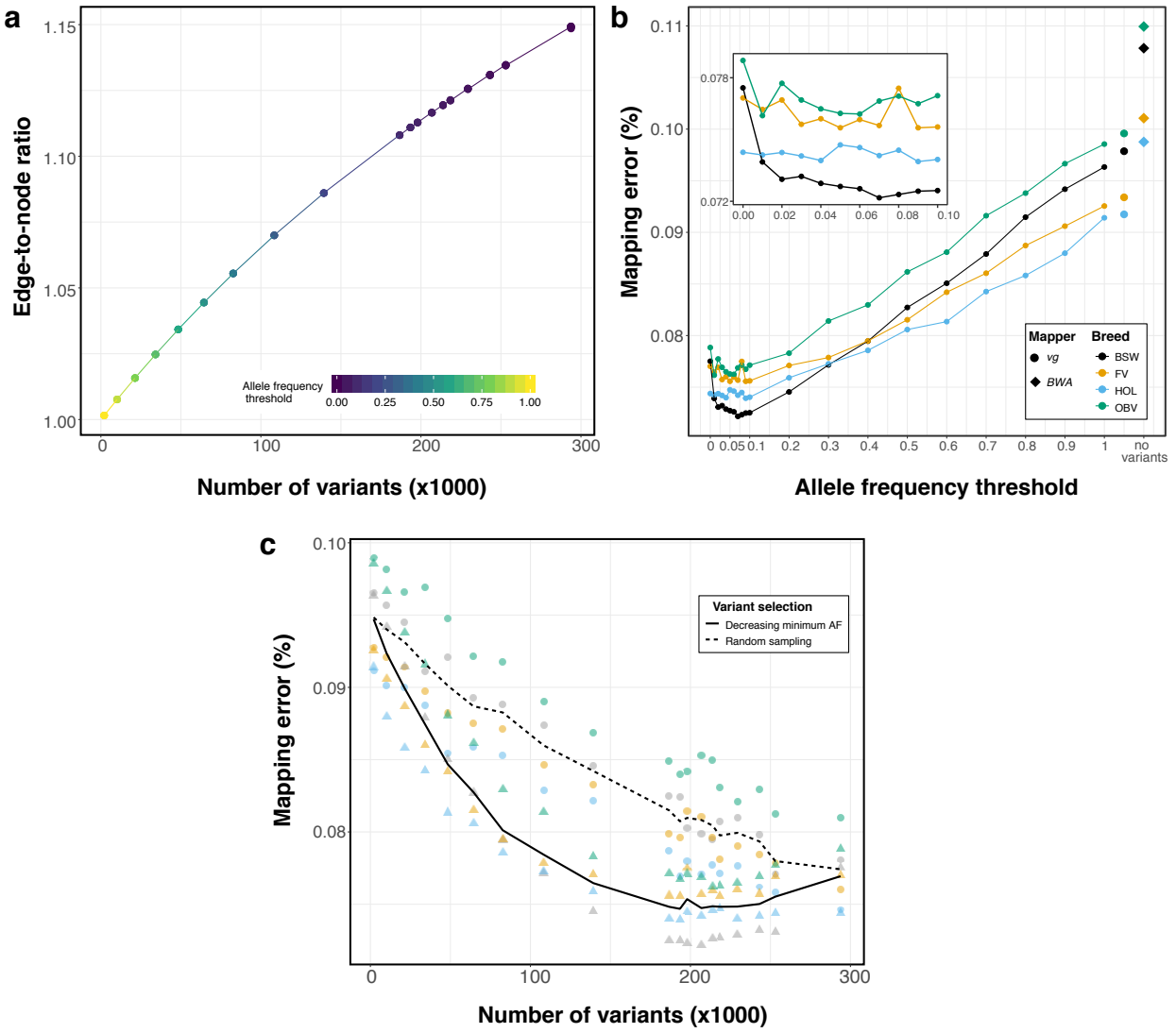

Figure 2 Paired-end mapping accuracy using genome graphs that contained variants filtered for allele frequency.
(a) Edge-to-node ratio of graphs that contained between 2,046 and 293,804 variants filtered for allele frequency. (b) Proportion of incorrectly mapped reads for four breed-specific augmented 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. (c) 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. Results for single-end mapping are presented in Additional file 1: Figure S1.

Variant prioritization based on allele frequency

We simulated 10 million paired-end reads (2x150 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 private to the 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 (Figure 2b). 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 (Figure 2b, 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 ($P = 0.13$, Figure 2b 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 (Figure 2c). This finding corroborates that variant prioritization based on alternate allele frequency is important to achieve high mapping accuracy with graph-based reference structures.

Read mapping accuracy was highly correlated (Pearson R=0.94, $P < 10^{-16}$) for single and paired-end reads (Additional file 1: Figure S1). 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 (Figure 2b). 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 ± 5K) than either FV (N=291 ± 3K), BSW (N=276 ± 6K) or HOL (N=259 ± 2K), reflecting that nucleotide diversity is higher in OBV than the other three breeds, which agrees with a recent study [25]. 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 three breeds considered [26]. 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 >0.02 was lower for BSW (84.1%) than the 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 (Additional file 1: Figure S2). An excess of private and rare variants in BSW and OBV cattle may have contributed to the decline in mapping accuracy at low alternate allele frequency thresholds (Figure 2b inset). 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.

Mapping to breed-specific augmented genome graphs

Next, we compared read mapping accuracy between breed-specific augmented and pan-genome graphs (i.e., graphs that contained variants filtered for allele frequency across multiple populations). 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 Material and 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 private to the BSW animal used for simulation were excluded from the graphs. 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 was 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 (Figure 3a). The average number of variants that were private to the breed-specific augmented graphs ranged from 8,010 in BSW to 20,392 in FV.

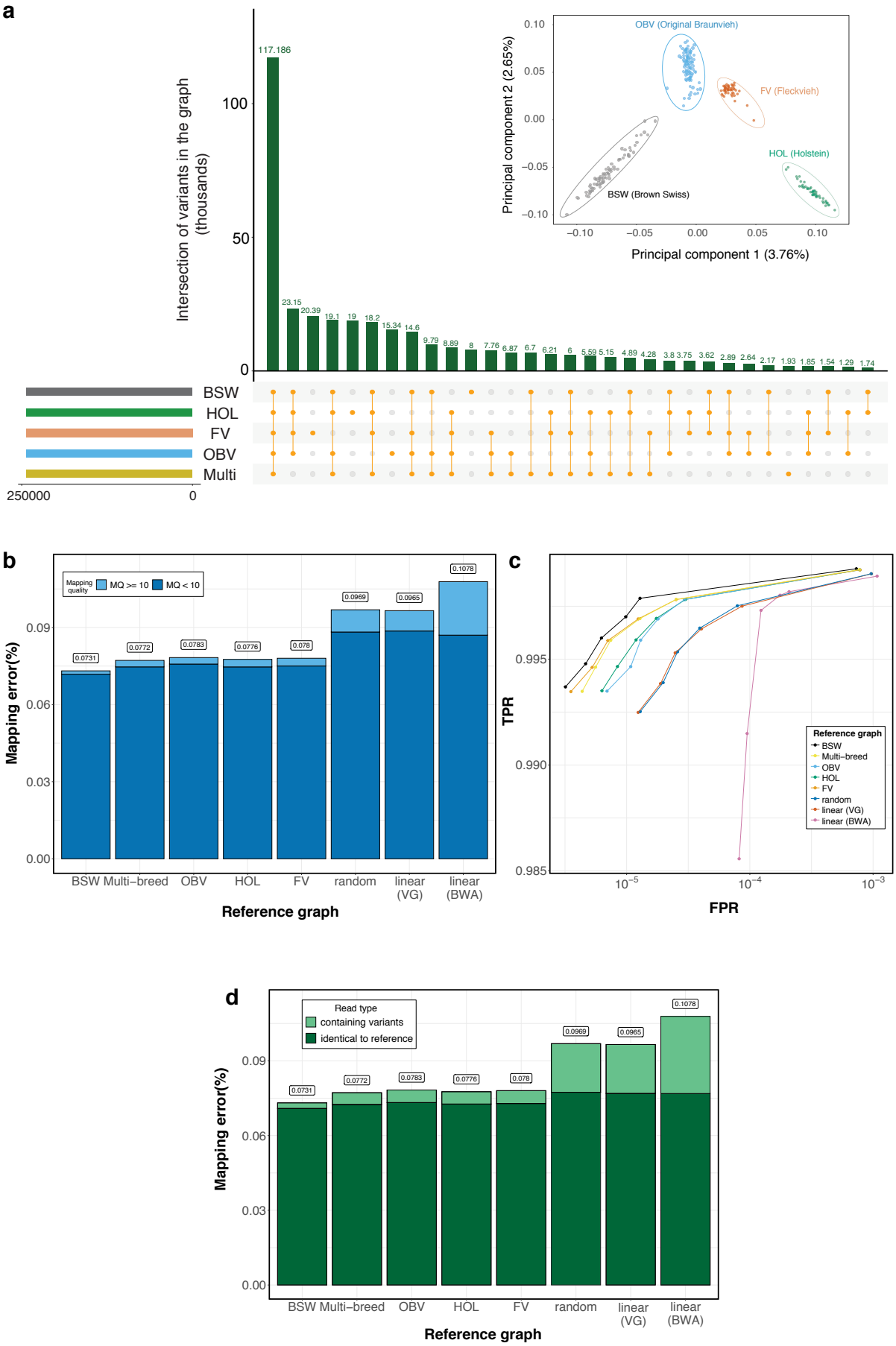
a
Intersection of variants in the graph (thousands)
117.186
23.15
20.39
19.1
19
18.2
15.34
14.6
9.79
8.89
8
7.76
6.87
6.7
6.21
6
5.59
5.15
4.89
4.28
3.8
3.75
3.62
2.89
2.64
2.17
1.93
1.85
1.54
1.29
1.74
BSW
HOL
FV
OBV
Multi
250000
0
OBV (Original Braunvieh)
FV (Fleckvieh)
HOL (Holstein)
BSW (Brown Swiss)
Principal component 2 (2.65%)
Principal component 1 (3.76%)
b
Mapping quality
MQ >= 10
MQ < 10
Mapping error(%)
0.0731
0.0772
0.0783
0.0776
0.078
0.0969
0.0965
0.1078
BSW
Multi−breed
OBV
HOL
FV
random
linear (VG)
linear (BWA)
Reference graph
c
TPR
FPR
Reference graph
BSW
Multi−breed
OBV
HOL
FV
random
linear (VG)
linear (BWA)
d
Read type
containing variants
identical to reference
Mapping error(%)
Reference graph

Figure 3 The accuracy of mapping simulated BSW paired-end reads to variation-aware and linear reference structures.
(a) We added 243,145 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. The inset is a plot of the top principal components to reflect the genetic diversity of the four cattle breeds considered. (b) Proportion of simulated BSW reads that mapped erroneously against 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 colours 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 colours 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 Additional file 1: Figure S3.

The highest mapping accuracy, sensitivity and specificity was achieved when the simulated BSW reads were aligned to a BSW-specific augmented graph (Figure 3bcd). 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 (Additional file 1: Figure S5). The number of mapping errors was higher (adjusted $P < 10^{-16}$, pairwise t test, Additional file 1: Figure S4) 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 (Additional File 1: Figure S6).

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

contained variants that were filtered for allele frequency in a different breed, likely because many common variants segregated across the four breeds considered (Figure 3a).

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, Additional file 1: Figure S4).

Compared to linear mapping using *BWA mem*, 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 1,3000,000 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 (Figure 3d, Additional file 1: Figure S3, Additional file 1: Figure S7). 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 (Figure 3d, Additional file 1: Figure S3,

Additional file 1: Figure S7). Incorrectly mapped reads that matched corresponding reference nucleotides primarily originated from repetitive and tandemly duplicated regions.

Linear mapping accuracy using a consensus reference sequence

Previous studies reported that linear mapping may be more accurate using population-specific than universal linear reference sequences [5, 27, 28]. 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 (Figure 4, Additional file 1: Figure S8). 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 (Figure 4b).

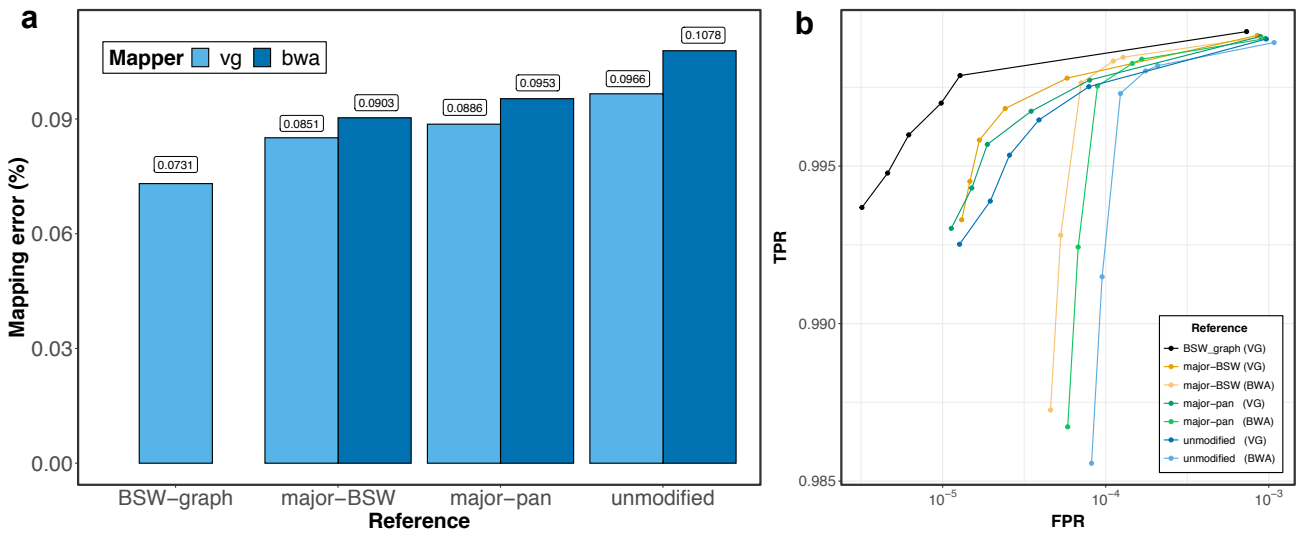

Figure 4 Paired-end read mapping accuracy using breed-specific augmented genome graphs and consensus linear reference sequences.
(a) Dark and light blue represent the proportion of reads that mapped incorrectly using *BWA mem* and *vg*, respectively, to the BSW-specific augmented reference graph (BSW-graph), 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 Additional file 1: Figure S8.

Read mapping and variant genotyping using whole genome graphs

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 yielding an edge-to-node ratio of 1.13. We also constructed a linear (empty) whole-genome graph that did not contain allelic variation. Subsequently, we mapped paired-end (2x150 bp) sequencing reads of 10 BSW cattle that had been sequenced at between 6 and 40-fold coverage (Additional file 3 Table S2) 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 to the BSW-specific augmented graph, the empty graph, and the linear reference sequence, respectively (Figure 5a). We observed slightly less uniquely mapped reads using either the whole-genome (82.46%) or empty graph (82.18%) than the linear reference sequence (83.18%) indicating that variation-aware references can increase mapping ambiguity due to providing alternative paths for read alignment.

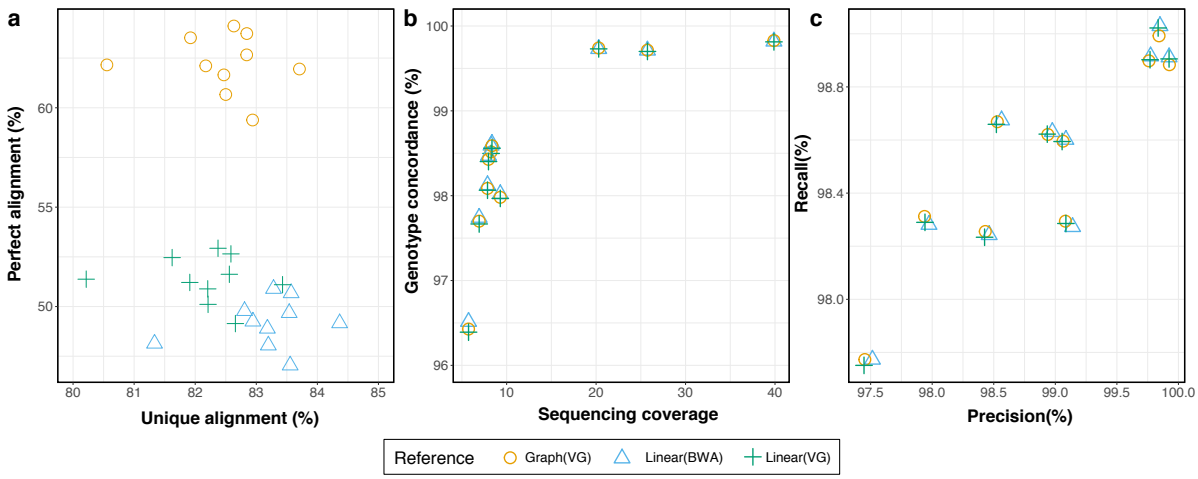

Figure 5 Sequence read mapping and variant genotyping using a breed-specific augmented whole-genome graph.
(a) Proportion of sequencing reads that mapped perfectly and uniquely to the BSW-specific augmented (circle) and Hereford-based linear (triangle, cross) reference. (b) Concordance between sequence variant and corresponding microarray-derived genotypes as a function of sequencing depth. Sequence variant genotypes were obtained using the multi-sample variant calling approach implemented in *SAMtools*. (c) Corresponding precision-recall statistic. Each symbol represents one BSW animal.

We converted (surjected) the graph-based read alignments of 10 BSW cattle to corresponding linear reference coordinates and genotyped polymorphic sites using *SAMtools mpileup*. In order to assess genotyping accuracy, we compared the sequence variant genotypes with array-called genotypes at corresponding positions. Sequence variant genotyping accuracy was correlated with sequencing coverage (Figure 5b). Genotype concordance, non-reference sensitivity, non-reference discrepancy, and precision did not differ between the graph-based and linear alignments for both raw and hard-filtered genotypes (Figure 5bc, Additional file 2: Table S1). 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* (Additional file 2: Table S1). In agreement with our previous findings [12], genotype concordance was slightly higher using *Graphtyper*, than either *SAMtools or GATK*.

Variation-aware alignment mitigates reference allele bias

To investigate reference allele bias in genotypes called from linear and graph-based alignments, we aligned sequencing reads of a BSW animal that was sequenced at 40-fold coverage (SAMEA6163185) to either the BSW-specific augmented whole-genome graph or linear reference sequence (Additional file 3: Table S2). 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 Material and Methods) 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 (Figure 6a), 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. 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 (Figure 6a). Inspection of the read alignments using the *Sequence Tube Map* graph visualization tool [29] corroborated that the support for alternate alleles is better using graph-based than linear references (Additional file 1: Figure S9).

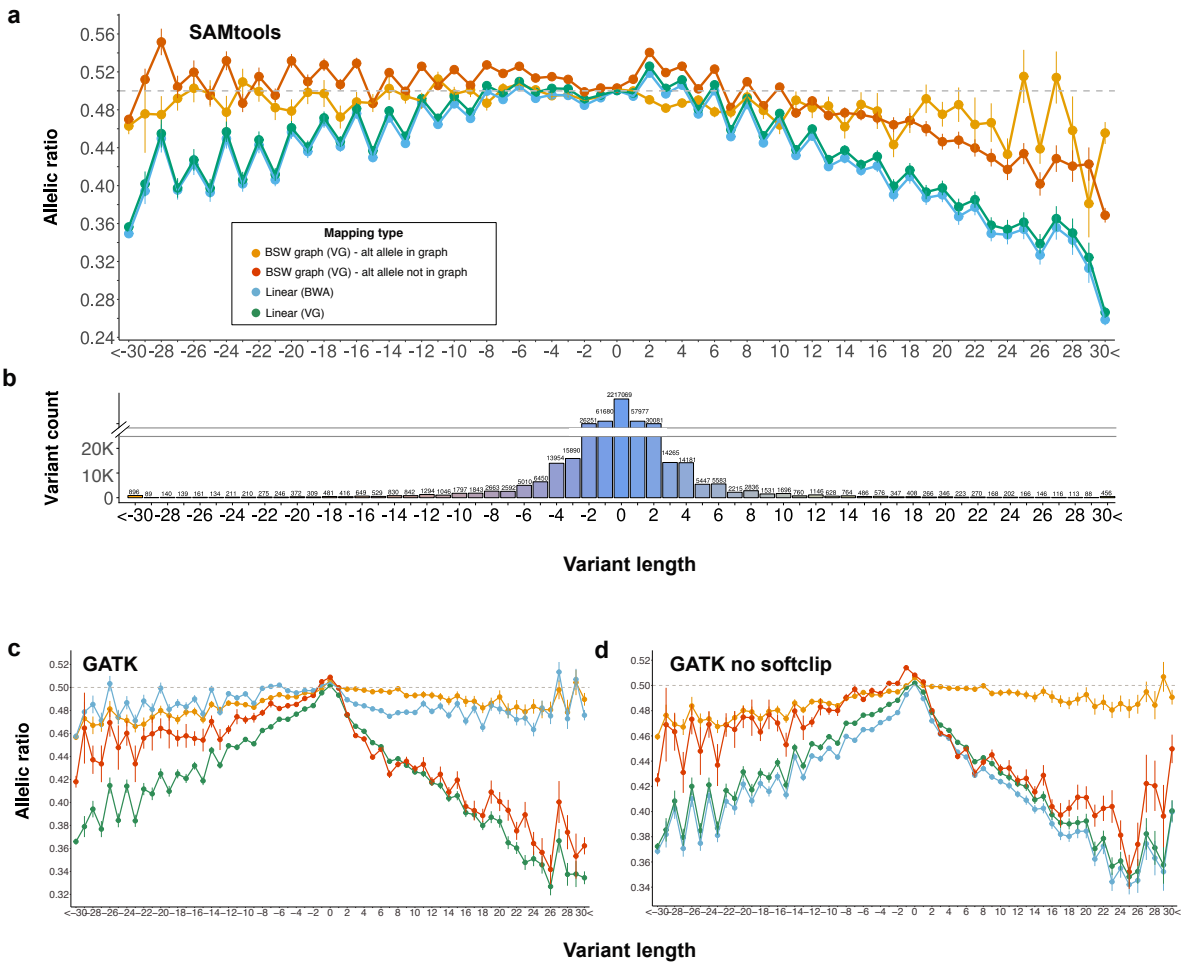

Figure 6 Reference allele bias from graph-based and linear alignments.
Reference allele bias from graph-based and linear alignments using **(a)** *SAMtools*, **(c)** *GATK*, or **(d)** *GATK* without soft-clip for variant genotyping and either *BWA mem* or *vg* for alignment. Allelic ratio reflects the proportion of mapped reads supporting the alternate allele. The grey 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 (± 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.

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 (Additional file 1: Figure S10). The difference in the number of mapped reads between graph-based and linear alignments increased with variant length. However, the number of mapped reads supporting the reference alleles did not differ between the graph-based and linear alignments. This finding indicates that reduced reference allele bias at Indel genotypes called from graph-based alignments is due to the improved mapping of reads that contain non-reference alleles.

We next investigated if these conclusions also hold for genotypes called by *GATK*. While *SAMtools mpileup* detects variants directly from the aligned reads [30], *GATK HaplotypeCaller* locally realigns the reads and calls variants from the refined alignments [31]. 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 (Figure 6c). 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* also mitigates reference alleles from linear alignments.

The percentage of soft-clipped reads increased with Indel length in the linear alignments (Additional file 1: Figure S11). 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 (Figure 6d). 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*.

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 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. Using both simulated and real short read data, our findings corroborate that a variation-aware reference facilitates accurate read mapping and unbiased sequence variant genotyping [22, 23, 32–34].

We simulated sequencing reads from haplotypes of BSW, FV, OBV and HOL cattle to obtain realistic and breed-specific base pair substitution and Indel rates. In agreement with previous studies [22, 34], read mapping was more accurate using variation-aware than linear reference sequences. Grytten *et al.* [34] observed that an adjusted parameter setting of *BWA mem* improves linear mapping. Because we compared graph-based mapping to the most-widely used linear mapping approach, we did not tune *BWA mem*. Our results show that a bovine breed-specific

consensus sequence also improves linear mapping which corroborates recent findings in humans [5, 27]. Because the adjusted consensus reference is also a linear sequence, tools that rely on a linear representation of the reference sequence can be used for read mapping and variant discovery. However, we find that read mapping is less accurate using a consensus reference sequence than variation-aware references that contain pre-selected variants.

The use of variation-aware references primarily improves the mapping of reads that contain non-reference nucleotides. This agrees with the findings of Garrison *et al.* and Grytten *et al.* [22, 34]. Reads that match corresponding reference alleles also aligned slightly better to a graph-based than linear reference. Using a bovine variation-aware reference reduced the proportion of erroneously mapped reads by more than 30% compared to the most widely-used linear mapping approach. A similar improvement in mapping accuracy over the linear reference was achieved for a human variation-aware reference genome [18]. The graph-based alignments contained 0.073% erroneously mapped reads. In order to avoid bias due to erroneous alignments, reads with low MQ are typically discarded for downstream analyses [21]. Incorrectly mapped reads that had high mapping quality (MQ >10) were less frequent in the graph-based than linear alignments. Thus, the variation-aware reference may reduce the number of flawed genotypes arising from mapping errors that would remain unnoticed in the linear alignments due to high mapping quality. However, the variation-aware reference of our study does not facilitate to accurately map short reads to repetitive and duplicated regions.

Our findings demonstrate that variant prioritization is key to accurate variation-aware read mapping. Based on investigations in four genetically distinct cattle breeds, we make three important observations: first, variation-aware references that contain random variants for which the allele frequency and haplotype phase in the target populations is unknown, do not improve read mapping accuracy over linear references. Our previous study also showed that adding many random variants does barely affect sequence variant genotyping from reference graphs [12]. Adding random unphased variants increases the number of alternative alignment paths that are not necessarily biologically plausible, 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. Further research is required to test if this observation also holds if breed-specific variants are more strongly enriched for rare alleles than in our study. Third, the highest mapping accuracy at tractable graph complexity can be achieved when variants filtered for allele frequency are added to the graph. We observed the highest mapping accuracy at allele frequency thresholds between 0.01 and 0.10. A trade-off between mapping accuracy and computational costs has been discussed previously [18]. In order to reduce the computational complexity of variation-aware read mapping, previous studies used arbitrarily chosen allele

frequency thresholds to prioritize variants to be included in the graphs (e.g., 1% [22, 35], 5% [36], 10% [37]). Using fine-grained allele frequency inclusion thresholds in four cattle breeds, we find that the read mapping accuracy does not notably differ between the 0.01 and 0.1% thresholds. Considering that most cattle breeds have an effective population size between 50 and 200 [38, 39], the vast majority of variants with allele frequency greater than 0.1 can be detected from a few sequenced key ancestor animals [13]. Most of the BSW, FV, OBV and HOL animals of our study are also key ancestor animals [7, 12, 13, 40]. As a matter of fact, key ancestor animals have been sequenced for many cattle breeds [7, 41]. Thus, the construction of variation-aware reference structures that are informative for many cattle breeds is readily possible using e.g., the sequence variant catalogue of the 1000 Bull Genomes Project [7, 8].

Mapping was most accurate to graphs that contained variants filtered for allele frequency in the target breed. Interestingly, we observed almost similar accuracy improvements over linear references for reads mapped to a pan-genome graph that contained variants filtered for allele frequency across four breeds of cattle (Figure 3b). Although the principal components analysis showed that the breeds considered in our study are genetically distinct, they share many common alleles. The bovine pan-genome graph constructed in our study contained between 75.28 and 80.82% of the variants that were added to the breed-specific augmented graphs. Our findings indicate that a pan-genome graph that contains alleles that have frequency greater than 3% in important breeds might serve as a widely-applicable bovine variation-aware reference structure. 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 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.3570312. 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 alignment approach that exploits a refined search space might reduce the number of multiple mappings in dense variation-aware

graphs [34]. Compared to a human whole-genome graph, the improvement in perfect mapping over the linear reference was slightly larger in our bovine whole genome graph (9.2%) [22]. However, the proportion of reads with perfect alignments (62.19%) was lower in our BSW-specific whole-genome graph, likely because it contained only sequences that were assembled to the 29 autosomes. The graph did not contain 269.77 Mb of the sex chromosomes, mitochondrial DNA, and 2180 unplaced contigs. A more sophisticated assembly of the bovine genome with increased continuity particularly at the sex chromosomes [3, 42] might serve as a backbone for an improved variation-aware genome graph.

Sequence variant genotyping from graph-based alignments using *vg call* is currently somewhat limited to structural variants [32]. 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 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 [43, 44]. Because approximately 90% of the considered SNPs were already included in the BSW-specific whole-genome graph, they can be detected and genotyped easily from graph-based alignments [17]. These variants can also be detected and genotyped accurately from linear alignments [12, 45].

Bias towards the reference allele was less in graph-based than linear alignments particularly if the variants were included in the graph. Unbiased genotyping of heterozygous variants from graph-based alignments is possible because reads supporting alternate alleles align better to variation-aware than linear references. Thus, our bovine whole-genome graph offers an appealing novel reference for investigations that either rely on low-coverage sequencing or are sensitive to unbiased allele frequencies [16, 19, 46]. Because a benchmark dataset for an unbiased evaluation of sequence variant genotyping performance [47] is not available in cattle, our assessment was restricted to heterozygous variants that were identified from both linear and graph-based alignments. This set of variants is possibly enriched for variants that can be called confidently from linear alignments, thus underestimating the graph-based genotyping performance (e.g., [22]).

Our study has three limitations. First, variants used to construct the breed-specific augmented genome graphs might be biased because they were detected from linear alignments of short sequencing reads. Variant discovery from an independent variation-aware reference structure might

allow for a more complete assessment of genetic variation [47]. 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 [48–51]. 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 BV cattle were not available. Haplotype-resolved genome assemblies of cattle from different breeds will facilitate the construction of more informative genome graphs and make non-reference sequences and their sites of variation amenable to genetic investigations [2, 3]. Third, we did not investigate the impact of large sequence variation on sequence read mapping and variant genotyping performance because a high-quality set of large structural variants (cf. [52]) is not available in cattle. Recent studies indicated that large structural variants can be identified accurately from genome graphs [32, 53–55]. 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 [56, 57].

Conclusions

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

Methods

Whole-genome sequencing data

We used short paired-end sequencing reads of 288 cattle from dairy (n=82 Brown Swiss – BSW, n=49 Holstein – HOL) and dual-purpose (n=49 Fleckvieh – FV, n=108 Original Braunvieh – OBV) breeds to detect variants that segregate in these populations. The average sequencing depth of the 288 cattle was 12.71-fold and it ranged from 3.49 to 70.04. Most of the sequencing data were generated previously [7, 12, 13, 40, 58]. Accession numbers for all animals are available in Additional file 3: Table S2.

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 at 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* [59]. Duplicates were marked and the aligned reads were coordinate sorted using the *Picard tools* software suite (http://broadinstitute.github.io/picard) and *Sambamba* [60], respectively. We discovered and genotyped polymorphic sites from the linear read alignments using the Best Practices Workflow descriptions for multi-sample variant calling with *GATK* (version 4.1.0) [1]. Because a truth set of variants required for variant quality score recalibration (VQSR) is not available for *Bos taurus*, we followed the recommendations for sequence variant discovery and filtration when applying VQSR is not possible. Genotypes of the hard-filtered variants were subsequently refined and sporadically missing genotypes were imputed with *BEAGLE* v4 [61] using the genotype likelihoods from the *GATK HaplotypeCaller* model as input values. Additional information on the sequence variant genotyping workflow and the expected genotyping accuracy can be found in Crysnanto *et al.* [12].

The principal components of a genomic relationship matrix constructed from whole-genome sequence variant genotypes were calculated using *PLINK* v1.9 [62]. The top principal components separated the animals by breeds, corroborating that the four breeds are genetically distinct (Figure 3a). 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 [63].

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

Haplotype-aware simulation of short sequencing reads

We simulated 40 million reads (150 bp) from reference haplotypes of one animal per breed that had sequencing coverage greater than 20-fold (see Additional file 3 Table S2). 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 (2x150 nt) from each haplotype using *vg sim*, yielding 10 million 150 bp reads per breed corresponding to approximately 35-fold sequencing coverage of bovine chromosome 25. The simulation parameter setting for read and fragment length was 150 and 500 (± 50), respectively. The substitution and indel error rate was 0.01 and 0.002, respectively, according to the settings used in Garrison *et al.* [22].

Read mapping to graphs augmented with variants filtered for allele frequency

The alternate allele frequency of 541,876 variants of bovine chromosome 25 was calculated separately for the BSW, FV, HOL and OBV breeds using sequence variant genotypes of 82, 49, 49 and 108 sequenced cattle, respectively. We added to each breed-specific genome graph 20 sets of variants that were filtered for alternate allele frequency using thresholds between 0 and 1 with increments of 0.01 and 0.1 for frequency below and above 0.1, respectively. For instance, at an alternate allele frequency threshold of 0.05, the graph was constructed with variants that had alternate allele frequency greater than 5%. Alleles that were private to 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 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 private to 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. 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 was 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 [46]. 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 [64] that was guided using a *chain file* produced by *vcf2diploid*. In order to prevent possible errors arising from coordinate shifts when reference nucleotides are either deleted or inserted at Indels, we repeated the analysis when only the alleles at SNPs were replaced.

Assessment of the read mapping accuracy

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

Using *vg annotate*, we annotated the simulated reads with respect to the linear reference coordinates and determined if they contained non-reference alleles. Comparing the true and mapped positions of the simulated reads enabled us to differentiate between correctly and incorrectly mapped reads. Following the approach of Garrison *et al.* [22] and taking into account the possibility that aligned reads may be clipped at Indels, we considered reads as incorrectly mapped if their starting positions were more than *k=150* (k=read length) bases distant from true positions.

In order to assess mapping sensitivity and specificity, we calculated the cumulative TPR (*True positive rate*) and FPR (*False positive rate*) at different mapping quality thresholds and visualized it as pseudo-ROC (*Receiver Operating Characteristic*) curve [22] using:

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

and

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

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 (2x150 bp) sequencing reads from 10 BSW cattle (SAMEA6163185, SAMEA6163188, SAMEA6163187, SAMEA6163177, SAMEA6163178, SAMEA6163176, SAMEA6163179, SAMEA6163183, SAMEA6163181, SAMEA6163182, Additional file 3 Table S2) 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* [65]. 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* [31], *Graphtyper* [35], or *SAMtools* [30], as stated above and detailed in [12].

In order to assess the read mapping accuracy from real sequencing data, we calculated the proportion of reads that aligned (i) perfectly and (ii) uniquely [18, 27, 66]. 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). 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 been genotyped at between 24,512 and 683,752 positions using Illumina SNP Bead chips. The sequence variant genotypes were compared to microarray-called genotypes at corresponding positions to calculate recall/non-reference sensitivity, genotype concordance, precision and non-reference discrepancy [1, 67]. The concordance metrics are explained in Additional file 1: Figure S12.

Snakemake workflows [68] 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.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

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.3570312 and scripts: https://doi.org/10.5281/zenodo.3592660). 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). Accession numbers are provided in Additional file 3 Table S2.

Funding

This study was supported by a grant from the Swiss National Science Foundation (310030_185229). The sequencing of Original Braunvieh cattle was partially funded from the Federal Office for Agriculture (FOAG), Bern, Switzerland. The sequencing data of the Brown Swiss bulls used for sequence variant genotyping were funded from Swissgenetics. The funders had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Competing interests

The authors declare that they have no competing interests.

Author contributions

Analyzed data: DC HP; Wrote the paper: DC HP; Read and approved the final version of the paper: DC HP.

Acknowledgements

We thank the many people who made data and software publicly available, particularly *vgteam* for providing the *vg toolkit* as an open source software. We thank Braunvieh Schweiz for providing pedigree and genotype data of Original Braunvieh and Brown Swiss cattle. Semen samples of the sequenced bulls were kindly provided by Swissgenetics.

References

1. DePristo MA, Banks E, Poplin R, Garimella K V., Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8. doi:10.1038/ng.806.
2. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. De novo assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol. 2018. doi:10.1038/nbt.4277.
3. Rice ES, Koren S, Rhie A, Heaton MP, Kalbfleisch TS, Hardy T, et al. Chromosome-length haplotigs for yak and cattle from trio binning assembly of an F1 hybrid. bioRxiv. 2019. doi:10.1101/737171.
4. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, et al. Telomere-to-telomere assembly of a complete human X chromosome. bioRxiv. 2019. doi:10.1101/735928.
5. Ballouz S, Dobin A, Gillis JA. Is it time to change the reference genome? Genome Biol. 2019;20:159. doi:10.1186/s13059-019-1774-4.
6. FAO. The second report on the state of the world´s animal genetic resources for food and agriculcure. Rome: Food and Agriculture Organization (FAO); 2015.
7. Daetwyler HD, Capitan A, Pausch H, Stothard P, Van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46:858–65. doi:10.1038/ng.3034.
8. Hayes BJ, Daetwyler HD. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. Annu Rev Anim Biosci. 2019;7:annurev-animal-020518-115024. doi:10.1146/annurev-animal-020518-115024.
9. Charlier C, Li W, Harland C, Littlejohn M, Coppieters W, Creagh F, et al. NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. Genome Res. 2016;26:1333–41.
10. Bovine Genome Sequencing and Analysis Consortium. The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. Science (80-). 2009;324:522–8. doi:10.1126/science.1169588.
11. Worley K, Gibbs R. Sequencing the Bovine Genome. John Wiley & Sons, Ltd; 2012.
12. Crysnanto D, Wurmser C, Pausch H. Accurate sequence variant genotyping in cattle using variation-aware genome graphs. Genet Sel Evol. 2019;51:21. doi:10.1186/s12711-019-0462-x.
13. Jansen S, Aigner B, Pausch H, Wysocki M, Eck S, Benet-Pagès A, et al. Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. BMC Genomics. 2013;14:446. doi:10.1186/1471-2164-14-446.
14. Kim J, Hanotte O, Mwai OA, Dessie T, Salim B, Diallo B, et al. The genome landscape of indigenous African cattle. Genome Biol. 2017;18:1–14.
15. Koufariotis L, Hayes BJ, Kelly M, Burns BM, Lyons R, Stothard P, et al. Sequencing the mosaic genome of Brahman cattle identifies historic and recent introgression including polled. Sci Rep. 2018;8:17761. doi:10.1038/s41598-018-35698-5.
16. Van De Geijn B, Mcvicker G, Gilad Y, Pritchard JK. WASP: Allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods. 2015;12:1061–3.
17. Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome inference. Genome Res. 2017;27:665–76. doi:10.1101/gr.214155.116.
18. Pritt J, Chen N-C, Langmead B. FORGe: prioritizing variants for graph genomes. Genome Biol. 2018;19:220. doi:10.1186/s13059-018-1595-x.
19. Günther T, Nettelblad C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. PLOS Genet. 2019;15:e1008302.
20. Salavati M, Bush SJ, Palma-Vera S, McCulloch MEB, Hume DA, Clark EL. Elimination of Reference Mapping Bias Reveals Robust Immune Related Allele-Specific Expression in Crossbred Sheep. Front Genet. 2019;10:1–16.
21. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics. 2009;25:3207–12. doi:10.1093/bioinformatics/btp579.
22. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018;36:875–81. doi:10.1038/nbt.4227.
23. Groza C, Kwan T, Soranzo N, Pastinen T, Bourque G. Personalized and graph genomes reveal missing signal in epigenomic data. bioRxiv. 2019;:457101. doi:10.1101/457101.
24. Sirén J, Garrison E, Novak AM, Paten B, Durbin R. Haplotype-aware graph indexes. Bioinformatics. 2019;:1–8.
25. Bhati M, Kadri NK, Crysnanto D, Pausch H. Assessing genomic diversity and signatures of selection in

Original Braunvieh cattle using whole-genome sequencing data. BMC Genomics. 2020;21. doi:10.1186/s12864-020-6446-y.
26. Signer-Hasler H, Burren A, Neuditschko M, Frischknecht M, Garrick D, Stricker C, et al. Population structure and genomic inbreeding in nine Swiss dairy cattle populations. Genet Sel Evol. 2017;49:1–13.
27. Shukla HG, Bawa PS, Srinivasan S. hg19KIndel: ethnicity normalized human reference genome. BMC Genomics. 2019;20:459. doi:10.1186/s12864-019-5854-3.
28. Dewey FE, Chen R, Cordero SP, Ormond KE, Caleshu C, Karczewski KJ, et al. Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. PLoS Genet. 2011;7:e1002280. doi:10.1371/journal.pgen.1002280.
29. Beyer W, Novak AM, Hickey G, Chan J, Tan V, Paten B, et al. Sequence tube maps: making graph genomes intuitive to commuters. Bioinformatics. 2019. doi:10.1093/bioinformatics/btz597.
30. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9. doi:10.1093/bioinformatics/btp352.
31. Poplin R, Ruano-Rubio V, Depristo MA, Fennell TJ, Carneiro MO, Van Der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2017. doi:10.1101/201178.
32. Glenn H, Heller D, Monlong J, Andreas Sibbesen J, Siren J, Eizenga J, et al. Genotyping structural variants in pangenome graphs using the vg toolkit. bioRxiv. 2019. doi:10.1101/654566.
33. Martiniano R, Garrison E, Jones ER, Manica A, Durbin R. Removing reference bias in ancient DNA data analysis by mapping to a sequence variation graph. bioRxiv. 2019.
34. Grytten I, Rand KD, Nederbragt AJ, Sandve GK. Assessing graph-based read mappers against a novel baseline approach highlights strengths and weaknesses of the current generation of methods. bioRxiv. 2019. doi:10.1101/538066.
35. Eggertsson HP, Jonsson H, Kristmundsdottir S, Hjartarson E, Kehr B, Masson G, et al. Graphtyper enables population-scale genotyping using pangenome graphs. Nat Genet. 2017;49:1654–60. doi:10.1038/ng.3964.
36. Maciuca S, Elias CDO, McVean G, Iqbal Z. A natural encoding of genetic variation in a burrows-wheeler transform to enable mapping and genome inference. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag; 2016. p. 222–33.
37. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15. doi:10.1038/s41587-019-0201-4.
38. Hall SJG. Effective population sizes in cattle, sheep, horses, pigs and goats estimated from census and herdbook data. Animal. 2016;10:1778–85.
39. Leroy G, Mary-Huard T, Verrier E, Danvy S, Charvolin E, Danchin-Burge C. Methods to estimate effective population size using pedigree data: Examples in dog, sheep, cattle and horse. Genet Sel Evol. 2013;45:1–10.
40. Baes CF, Dolezal MA, Koltes JE, Bapst B, Fritz-Waters E, Jansen S, et al. Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. BMC Genomics. 2014;15:948. doi:10.1186/1471-2164-15-948.
41. Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel FS, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nat Genet. 2018;50:362–7. doi:10.1038/s41588-018-0056-5.
42. Liu R, Low WY, Tearle R, Koren S, Ghurye J, Rhie A, et al. New insights into mammalian sex chromosome structure and evolution using high-quality sequences from bovine X and Y chromosomes. BMC Genomics. 2019;20:1–11. doi:10.1186/s12864-019-6364-z.
43. Li H, Wren J. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30:2843–51. doi:10.1093/bioinformatics/btu356.
44. Malomane DK, Reimer C, Weigend S, Weigend A, Sharifi AR, Simianer H. Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. BMC Genomics. 2018;19:22. doi:10.1186/s12864-017-4416-9.
45. Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, et al. An open resource for accurately benchmarking small variant and reference calls. Nat Biotechnol. 2019;:1. doi:10.1038/s41587-019-0074-6.
46. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, et al. AlleleSeq: Analysis of allele-specific expression and binding in a network framework. Mol Syst Biol. 2011;7:1–15. doi:10.1038/msb.2011.54.
47. Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. Nat Methods. 2018;15:595–7.
48. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet. 2019;51:30–5. doi:10.1038/s41588-018-0273-y.
49. Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ, et al. A high-quality

human reference panel reveals the complexity and distribution of genomic structural variants. Nat Commun. 2016;7:1–10.
50. Holden LA, Arumilli M, Hytönen MK, Hundi S, Salojärvi J, Brown KH, et al. Assembly and Analysis of Unmapped Genome Sequence Reads Reveal Novel Sequence and Variation in Dogs. Sci Rep. 2018;8:1–11.
51. Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, et al. Building the sequence map of the human pan-genome. Nat Biotechnol. 2010;28:57–62.
52. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10:1784. doi:10.1038/s41467-018-08148-z.
53. Chen S, Krusche P, Dolzhenko E, Sherman RM, Petrovski R, Schlesinger F, et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. Genome Biol. 2019;20:1–13.
54. Rakocevic G, Semenyuk V, Lee WP, Spencer J, Browning J, Johnson IJ, et al. Fast and accurate genomic analyses using genome graphs. Nat Genet. 2019;51:354–62.
55. Eggertsson HP, Kristmundsdottir S, Beyter D, Jonsson H, Skuladottir A, Hardarson MT, et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. Nat Commun. 2019;10:5402. doi:10.1038/s41467-019-13341-9.
56. Duan Z, Qiao Y, Lu J, Lu H, Zhang W, Yan F, et al. HUPAN: A pan-genome analysis pipeline for human genomes. Genome Biol. 2019;20. doi:10.1186/s13059-019-1751-y.
57. Beyter D, Ingimundardottir H, Eggertsson HP, Bjornsson E, Kristmundsdottir S, Mehringer S, et al. Long read sequencing of 1,817 Icelanders provides insight into the role of structural variants in human disease. bioRxiv. 2019.
58. Hofstetter S, Seefried F, Häfliger IM, Jagannathan V, Leeb T, Drögemüller C. A non-coding regulatory variant in the 5′-region of the MITF gene is associated with white-spotted coat in Brown Swiss cattle. Anim Genet. 2019;50:27–32.
59. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013. http://arxiv.org/abs/1303.3997.
60. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31:2032–4. doi:10.1093/bioinformatics/btv098.
61. Browning BL, Browning SR. Genotype Imputation with Millions of Reference Samples. Am J Hum Genet. 2016;98:116–26. doi:10.1016/j.ajhg.2015.11.020.
62. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience. 2015;4:1–16.
63. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 2018;103:338–48. doi:10.1016/j.ajhg.2018.07.015.
64. Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, et al. The UCSC Genome Browser database: 2019 update. Nucleic Acids Res. 2019;47:D853–8.
65. Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90. doi:10.1093/bioinformatics/bty560.
66. Novak AM, Hickey G, Garrison E, Blum S, Connelly A, Dilthey A, et al. Genome Graphs. bioRxiv. 2017. doi:10.1101/101378.
67. Linderman MD, Brandt T, Edelmann L, Jabado O, Kasai Y, Kornreich R, et al. Analytical validation of whole exome and whole genome sequencing for clinical applications. BMC Med Genomics. 2014;7:20. doi:10.1186/1755-8794-7-20.
68. Köster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics. 2012;28:2520–2. doi:10.1093/bioinformatics/bts480.

Additional files

Additional file 1

File format: pdf

Description: Integrated supplementary figures: S1-S12.

Additional file 2 Table S1

File format: xlsx

Description: 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.

Additional file 3 Table S2

File format: xlsx

Description: Accession numbers for the animals used for variant detection, short read simulation, sequence read mapping and genotyping, and reference allele bias assessment.