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

Chapter 1

Analysis of the local graphs

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 is now publicly available at
https://github.com/danangcrysnanto/Graph-genotyping-paper-pipelines.

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., 2017; 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., 2017; 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 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) [46]. 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). 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 improve 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.

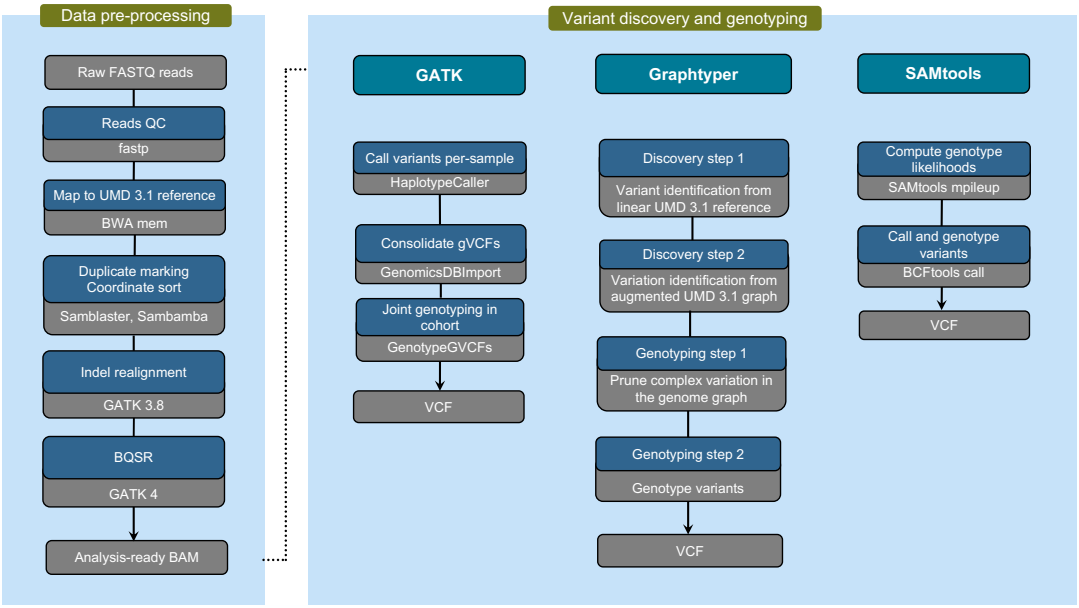

Figure 1.1: **Schematic representation of the three sequence variant discovery and genotyping methods evaluated.**
According to the best practice recommendations for sequence variant discovery using *GATK*, the VQSR module should be applied to distinguish between true and false positive variants. Because 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. 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.

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

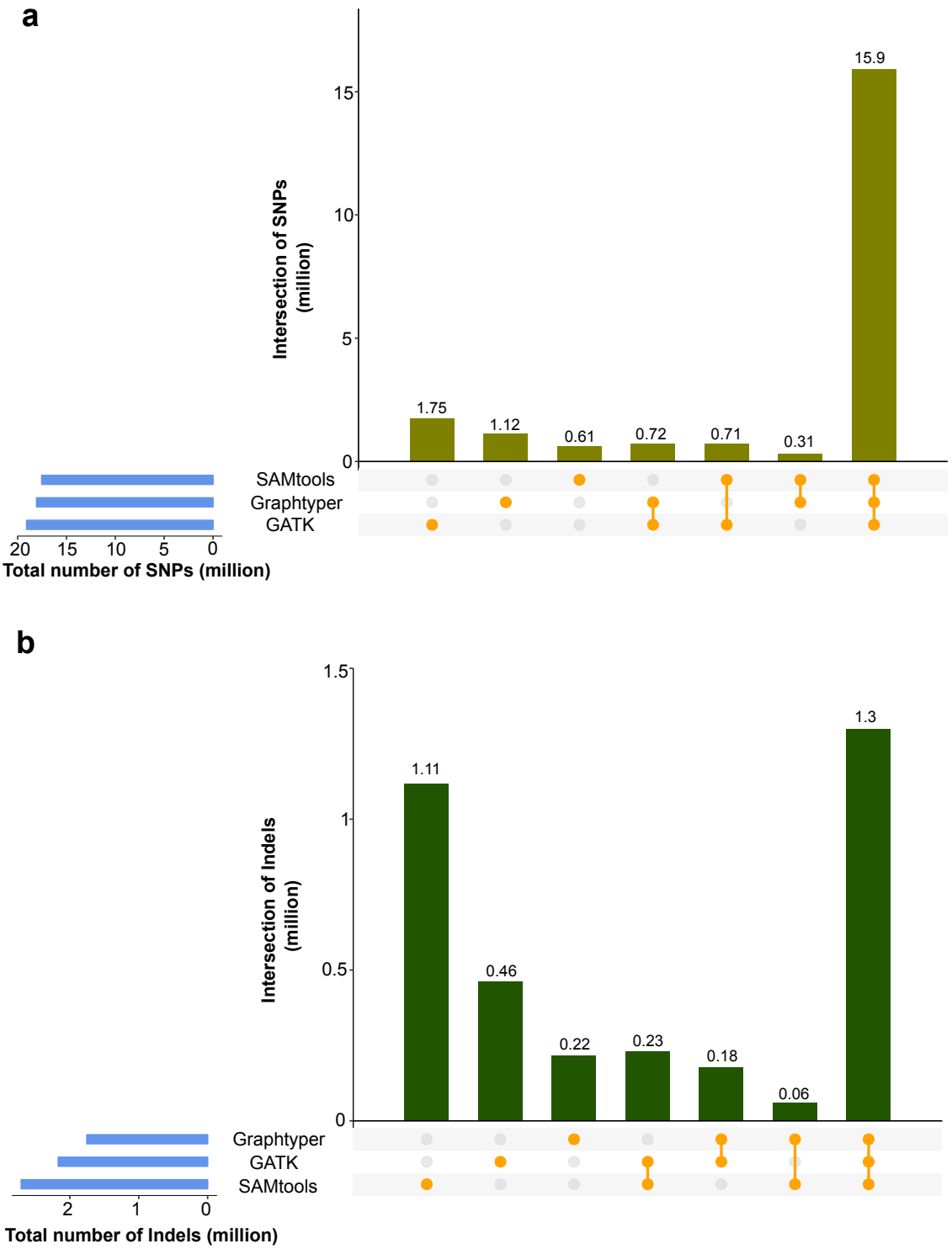

Figure 1.2: Number of biallelic SNPs (**a**) and indels (**b**) identified in 49 Original Braunvieh cattle using three sequence variant genotyping methods. Blue horizontal bars represent the total number of sites discovered for each method. Vertical bars indicate private and common variants detected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

References

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.

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.

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.

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.

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.

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.

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, 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.

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 2017. ISSN 1471-2164. doi: 10.1186/s12864-017-4263-8.

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.

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.

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.

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.