Computational identification and validation of genetic variability: exploration in a cattle breed

Audald Lloret-Villas

DISS. ETH NO. 29794

DISS. ETH NO. 29794

Computational identification and validation of genetic variability: exploration in a cattle breed

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

Audald Lloret-Villas

M.Sc., Omics Data Analysis

Universitat de Vic

born on 05.01.1991

citizen of Flix, Catalunya

accepted on the recommendation of:

Prof. Dr. Hubert Pausch	examiner
Prof. Dr. James Prendergast	co-examiner

2023

Table of Contents

List of Figures

List of Tables

Summary

The possibility to investigate DNA variability at the population scale enables fundamental insight into the genetic architecture of heritable traits. Cutting-edge genomic approaches, methods, and computational tools are being developed at an unprecedented pace to keep up with rapidly increasing volumes of data. Such technological advances enable comprehensive exploration of genetic variability in livestock and how its intricacies impact aspects such as population management, economic traits, and diseases. Accurate detection and validation of variants (genotyping) is crucial for downstream analysis. This dissertation aims at exploring the key factors, introduced in Chapter 1, that affect the identification of variants and the genotyping accuracies using Brown Swiss cattle as an exemplary bovine population.

Chapter 2 assesses the impact of reference genomes on genomic studies. Whole genome sequencing short reads of 161 bovine samples from the BSW breed were used to compare mapping statistics, variant detection, and genomic downstream analyses when two different reference genomes were used. These genomic analyses included functional annotation, signatures of selection, genotype-phenotype association testing and genomic heritability of phenotypic traits explained by dense markers. The reference genomes used were the curated and widely accepted Hereford-based ARS-UCD1.2 and the highly continuous, haplotype-resolved Angus-based UOA_Angus_1. The results indicate that no crucial differences in read mapping, genotype detection and accuracy arise when the two different assemblies are used. However, assembly flaws (chromosomal truncation) and limitations in the annotation of the haplotype-resolved assembly were evident and affected the detection of variants which may possibly be linked to phenotypes of interest. Accounting for these assembly and annotation boundaries, breed-specific primary assemblies can be readily integrated in genomic analyses of target breeds.

Chapter 3 compares the variant genotyping accuracy of two variant callers: GATK and DeepVariant. 50 BSW samples with coverages ranging from 4- to 63-fold, for 33 of

which microarray data were available and could be used as a truth set, were used to validate the called genotypes. The variant caller that performed better for bovine short read sequencing (DeepVariant) was then used to generate a series of haplotype reference panels. The second goal was to evaluate the impact of size and composition of such reference panels, as well as the sequencing coverage (between 0.01- and 4-fold), for the imputation of low-pass data to (higher-coverage) sequence level. A set of 24 BSW samples with coverages higher than 20-fold were used to validate the imputed genotypes. The parameters to choose a suitable variant caller and composition of haplotype reference panel for imputation are thoroughly determined in this chapter. Provided a sufficient number of sequenced samples (n = 150), breed-specific haplotype panels can perform better in imputation than larger multi-breed haplotype panels.

Chapter 4 reports on a preliminary exploration of methylation patterns across bovine CpG dinucleotides and their clustering in CpG islands. Long-read sequencing data for 120 BSW bulls, with an average coverage of 14.5-fold, were screened for the most frequent epigenetic modifications (5mC). Two different tissues from the male reproductive system (105 testis samples and 15 epididymis samples) were analysed. Initial inspection of differential methylation models among tissues and individuals was performed. Considering the phenotypical effect of epigenetic modifications, the methylation status of CpG islands can be directly connected to genotypic variation.

This thesis leverages cutting-edge tools and methods to further understand variation in bovine. It produces relevant conclusions, duly expanded in Chapter 5, for optimised genomic analyses regarding what reference genome is best suited, which variant caller yields most accurate genotyping and how haplotype reference panels are composed ideally. Two additional examples are provided on how genetic variants are key for innovative approaches. First, this thesis demonstrates the use of a proper set of reference haplotypes (DNA stretches where variants are inherited together) for the imputation of low-coverage data. Second, it links properly called genotypes to methylation variability. This thesis introduces and explores a computational framework to study the importance of variation in livestock and the scope of its applications.

Zusammenfassung

Die Möglichkeit, DNA-Variabilität auf Populationsebene zu untersuchen, ermöglicht einen fundamentalen Einblick in die genetische Architektur vererbbarer Merkmale. Innovative genomische Ansätze, Methoden und rechnerische Werkzeuge werden in unvergleichlichem Tempo entwickelt, um die immer grösser werdenden Datensätze zu verarbeiten. Diese technologischen Fortschritte ermöglichen eine umfassende Untersuchung der genetischen Variabilität bei Nutztieren und ihre Auswirkungen auf Themen wie Populationsmanagement, Wirtschaftlichkeit und Krankheit. Präzise Erkennung und Validierung der genetischen Varianten (Genotypisierung) ist entscheidend für nachfolgende Analysen. Ziel dieser Dissertation ist es, die in Kapitel 1 vorgestellten Schlüsselfaktoren zu untersuchen, die die Identifizierung der Varianten und die Genauigkeit der Genotypisierung mithilfe der Daten von Brown Swiss (BSW) Rindern als exemplarische Rinderpopulation zu untersuchen.

In Kapitel 2 werden die Auswirkungen von Referenzgenomen auf genomische Studien bewertet. Mit Short-Read-Ganzgenomsequenzen von 161 Rinderproben der BSW-Rasse wurden Mapping-Statistik, Variantenerkennung und nachfolgende genomische Analysen mit zwei unterschiedlichen Referenzgenomen verglichen. Diese genomischen Analysen beinhalten funktionale Annotation, Selektionssignaturen, Genotyp-Phänotyp-Assoziationstests und durch dichte Marker erklärte genomische Vererbbarkeit phänotypischer Merkmale. Als Referenzgenom wurden der kuratierte und weithin anerkannte Hereford-basierende ARS-UCD1.2, sowie der kontinuierliche haplotyp-aufgelöste Angus-basierende UOA_Angus_1 verwendet. Die Resultate zeigen keine entscheidenden Unterschiede in Bezug auf Read-Mapping, Genotyperkennung und Genauigkeit, wenn die zwei unterschiedlichen Assemblies verglichen werden. Jedoch sind Fehler im Assembly (Chromosomenverkürzung) und Einschränkungen bei der Annotation des haplotyp-aufgelösten Assembly ersichtlich und beeinflussen die Erkennung der Varianten, die möglicherweise mit relevanten Phänotypen in Verbindung stehen. Unter Berücksichtigung dieser Assembly- und Annotationsfehler können rassenspezifische

primäre Assemblies leicht in genomische Analysen der Zielrassen integriert werden.

In Kapitel 3 wird die Genauigkeit der Genotypisierung von Varianten mit zwei Variantencallern verglichen: GATK und DeepVariant. Die abgerufenen Genotypen wurden anhand 50 BSW-Proben mit einer 4- bis 63-fach Abdeckung validiert (wovon für 33 Proben Microarray-Daten vorhanden waren und als Wahrheitssatz dienten). Der Variantencaller mit besserer Leistung in Short-Read-Sequenzierung bei Rindern (DeepVariant) wurde anschliessend gebraucht, um eine Reihe von Haplotyp-Referenzpanels zu generieren. Das zweite Ziel war es, sowohl den Einfluss der Grösse und Zusammensetzung dieser Referenzpanels als auch die Sequenzierabdeckung (zwischen 0.01- und 4-fach) für die Imputation von Low-Pass-Daten zu (höher-abgedecktem) Sequenzlevel zu evaluieren. Mit einem Set von 24 BSW-Proben mit mehr als 20-facher Abdeckung wurden die imputierten Genotypen validiert. In diesem Kapitel werden die Parameter für die Wahl eines passenden Variantencallers und die Zusammensetzung des Haplotyp-Referenzpanels für die Imputation gründlich untersucht. Unter der Voraussetzung, über genügend sequenzierte Proben (n = 150) zu verfügen, schneiden rassenspezifische Haplotyp-Panels für die Imputation besser ab als Haplotyp-Panels mehrerer Rassen.

Kapitel 4 umfasst eine vorläufige Erkundung der Methylierungsmuster von CpG Dinukleotiden und deren Cluster in CpG-Inseln bei Rindern. Es wurden Long-Read-Sequenzdaten von 120 BSW-Stieren mit einer durchschnittlichen 14.5-fach Abdeckung auf die häufigsten epigenetischen Veränderungen untersucht (5mC). Zwei unterschiedliche Gewebe des männlichen Reproduktionssystems (105 Hoden-Proben und 15 Epididymis-Proben) wurden analysiert. Erste Untersuchungen unterschiedlicher Methylierungsmodelle zwischen Gewebe und Individuen wurden durchgeführt. Unter Berücksichtigung der phänotypischen Ausprägung epigenetischer Veränderungen kann der Methylierungsstatus der CpG-Inseln direkt mit der genotypischen Variation in Verbindung gebracht werden.

In dieser Arbeit wurden innovativste Instrumente und Methoden verwendet, um

Variationen bei Rindern besser zu verstehen. Für die Optimierung genomischer Analysen werden Schlussfolgerungen gemacht in Bezug auf welches Referenzgenom am besten geeignet ist, welcher Variantencaller die präziseste Genotypisierung liefert und wie die Haplotyp-Referenzpanels idealerweise zusammengesetzt sein sollten. Zusätzlich werden zwei Beispiele gemacht, wieso genetische Varianten für innovative Ansätze entscheidend sind. Erstens demonstriert diese Arbeit die Verwendung von geeigneten Referenz-Haplotypsets (DNA-Abschnitte mit zusammen vererbten Varianten) für die Imputation von nicht-dichten Daten. Zweitens können richtig abgerufene Genotypen mit der Methylierungsvariabilität in Verbindung stehen. Mit dieser Arbeit wird ein Berechnungsrahmen erarbeitet, um die Bedeutung von Variabilität bei Nutztieren und deren Anwendungsbereich zu untersuchen.

Thesis Outline

The thesis is structured as follows:

Chapter 1 provides a literature review to introduce genetic variation and the elements that intervene in its proper detection. A series of applications requiring highly accurate variants are subsequently described.

Chapter 2 assesses the impact of reference genomes for a range of genomic analyses, using the Brown Swiss cattle breed as a target species. This chapter is published in BMC Genomics.

Chapter 3 reports on the comparison of the genotype accuracy yielded by two widespread variant callers. It additionally explores how the composition of haplotype reference panels affect the imputation of low-pass data at different coverage folds. This chapter is published in Genetics Selection Evolution (GSE).

Chapter 4 introduces preliminary results of methylation patterns in two tissues of the male reproductive system of cattle. The distribution and variability of 5mC modifications across the genome and within CpG islands are explored.

Chapter 5 provides a general discussion, and outlook for future research.

The image cover was created with the assistance of DALL·E 2

Chapter 1

General Introduction

1.1 Identification of genetic variation in livestock

Bovine populations have a high nucleotide identity. The average number of homologue nucleotides (biochemical structures that form the basic constituent of DNA) that are identical between any two individuals within the same population often exceeds 99.9% [1, 2, 3]. The remaining < 0.1% determines the uniqueness of individuals. Therefore, a confident determination of such polymorphic nucleotides (genetic variants) is paramount to the success of genomic investigations [4].

Sequencing data

Obtaining the digital representation of the nucleotides for a complete genome is not trivial. Current technologies cannot read and reproduce a copy of the ∼3 billions of nucleotide bases (Gb) from a typical mammalian genome at once. Instead, as part of the whole-genome sequencing (WGS) approach, the DNA is divided into small pieces and sequencing machines provide digital copies of such subsets of the genome (reads). Next-generation sequencing (NGS) technologies enable massively parallel sequencing generating millions of reads per instrument run [5]. Short-read technologies (*e.g.*, Illumina) are the most used form of NGS. Illumina machines generate short sequencing reads that typically vary between 30 and 150 base pairs (bp) with sequencing error rates < 0.1%. Third-generation sequencing (also known as long-read sequencing) are newer technologies mainly dominated by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) which generate sequencing reads of hundreds of kilobases (Kb) in length but with higher error rates (5 - 15%) [6, 7]. PacBio's most recent "HiFi" circular consensus sequencing offers a compromise between read length (15 - 20-kbp) and accuracy (error rate of 0.1 - 1%) [8]. Given the large number of sequencing reads needed to cover a genome, even tiny error rates cause thousands of wrongly represented nucleotides. The confidence in the determination of the nucleotides depends on the coverage, the number of strings of reads that cover a certain position (locus) of the genome [9]. Regardless the technology used (long or short reads), high-coverage WGS is prohibitively costly for large-scale bovine population studies [10]. For sequence data to be used routinely in research and breeding, low-coverage whole genome sequencing (lcWGS) is a cost-effective strategy that allows population-scale screening of the entire genome [11, 12].

Reference genome

Reference genomes are a point representation of the structure and organization of the genome of a species [13]. They are an important prerequisite for many genomic anal-

yses and have been integral to the discovery of molecular genetic variants. The assembly of complex eukaryotic reference genomes has been fostered by recent technological improvements in long-read sequencing (see Fig. 1.1) [14, 15, 16]. The Vertebrate Genome Project [17], the Darwin Tree of Life project [18], the Earth Biogenome Project [19] and other efforts are underway to generate reference quality genomes for hundreds of thousands of species in the coming years.

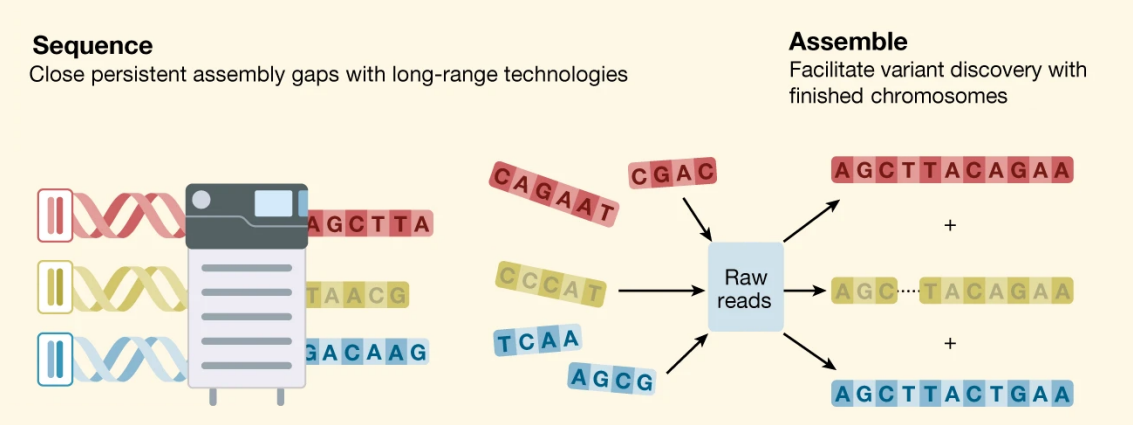


Figure 1.1: **Genome Assembly.** Long-read data generated with third-generation sequencing are used to assemble genomes that foster variant discovery - adapted from [20].

However, a single reference genome does not fully reflect the extent of genetic variation within a species [21, 22]. Reference genomes tend to derive from a single individual or a reduced subset of inbred individuals [23, 24, 25, 26]. This poses the risk that private alleles (this is, nucleotides in specific positions - loci - that are only present in the selected individual) become the reference while segregating alleles (common in the population) are considered as variants.

The current cattle reference genome (ARS-UCD1.2) [27] was derived from an inbred Hereford cow and her sire [26] and belong to one of the most common breeds in use for beef production globally. Given the decreasing error rates and increasing output of sequencing technologies, it is now possible to generate multiple reference quality genomes per species, tailored to the research needs. Trio-binning [28] is a method that uses parent-specific information to sort long sequences and simplifies the generation of reference genomes [28]. Some of the highest quality bovine genomes to date have been produced with trio-binning: yak (*Bos grunniens*), gaur (*Bos gaurus*), Brown Swiss, Original Braunvieh, Scottish Highland, Angus, Piedmontese and Simmental (*Bos taurus taurus*), and Nellore and Brahman (*Bos taurus indicus*) [28, 29, 30, 31, 32].

Genotyping

The bovine genomics community heavily relied on SNP (single nucleotide polymorphism) microarrays to characterize genetic differences between individuals. SNP microarrays comprise hundreds of thousands of probes that target genomic regions where variants have been previously discovered by sequencing and comparing to a reference genome [33]. The flanking regions of the variable position hybridise with the probes and so the presence of the variant can be detected. Microarrays are designed to interrogate common genetic markers in mainstream breeds. Such markers (SNPs) tend to be more frequent than random SNPs and may thus be not representative of all breeds [33]. Largely diverged breed-specific variants are typically less accessible and underrepresented in these arrays [34]. This effect is known as ascertainment bias [35].

Unlike the targeted genotyping (determination of the alleles) of SNP microarrays, WGS screens all the nucleotides of an individual genome without previous information about the position of polymorphic sites [36]. Polymorphic sites can be discovered and genotyped through the correct alignment of the sequencing reads to the relevant genomic locations in the reference genome and comparison of the differences [37, 38, 39]. These differences are due to variants present in the sequenced reads or errors either in the sequencing platform or in the reference genome [39]. The improvements in sequencing technologies and reference genomes are therefore tightly linked with the improvement of alignment rates [40]. The combination of a complete, high-quality reference genome and accurate and well-mapped reads that overlap segregating polymorphic sites in a population enable variant discovery and genotyping via resequencing (see Fig. 1.2) [37, 38, 41]. But read alignment is an important computational challenge [39, 42]. Reads in divergent regions are more prone to be misaligned, especially in highly polymorphic regions and for samples that are divergent from the reference genome as the reads span more nucleotide mismatches [11]. Reference genomes originating from individuals closer from the sequenced individuals reduce the computational burden. Additionally, the error rate of genotyping is increased by the reference bias, a preferential bias towards the reference allele during read alignment that causes misalignment for reads containing non-reference alleles [11, 43].

Short-read sequencing technologies are accurate for the genotyping of small-variants such as SNPs and insertions and deletions shorter than 50 nucleotides or base pairs (INDELs) [44]. Large fractions of the genome comprise repetitive regions which are less accessible with short reads as the mapping can be ambiguous and hence part of the variation can be missed [15]. Long-read sequencing can overcome limitations of short reads by substantially improving variant detection inside difficult-to-map parts of the

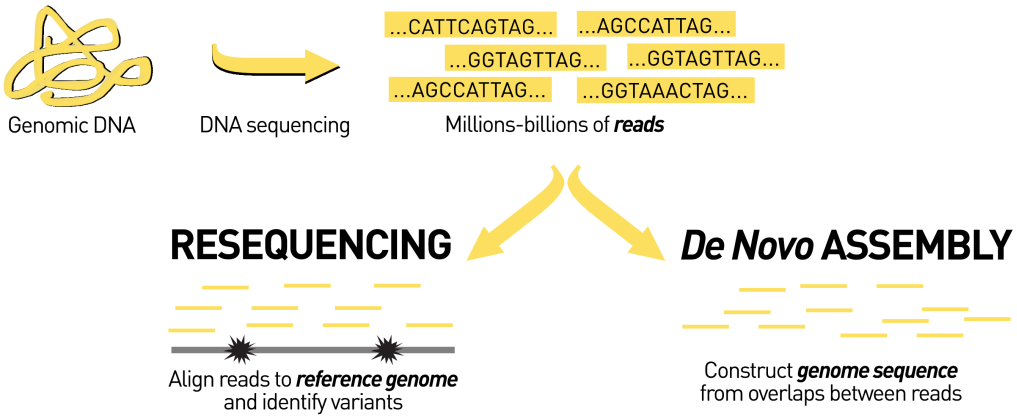


Figure 1.2: **Resequencing and *de novo* assembly.** NGS enable identification of genetic variants and assembly of genomes. Credit: Cathleen Shaw / HudsonAlpha Institute for Biotechnology: https://www.hudsonalpha.org.

genome, such as centromeres and sex chromosomes [45, 46]. With the improvement of base callers, long-read data can also simultaneously detect some epigenetic base modifications at the nucleotide level, such as methylated cytosines (5mC) [47]. Methods have been developed for both ONT (detection of deviations in the electric signal) and PacBio HiFi data (analysis of polymerase kinetics) [47, 48, 49].

Sequence variant genotyping accuracy typically decreases with lower sequencing coverage [50, 51, 52]. The identification of true genotypes from lcWGS (< 4-fold) is challenging because the impact of sequencing errors is more relevant. As a trade-off, more individuals can be sequenced and characteristics such as the frequencies of the alleles segregating in a population can be obtained [12]. For a limited amount of sequencing, the discovery of variants is maximised by sequencing samples at low coverage [53, 54].

Genotyping tools - variant callers

The choice of the software used for variant discovery and genotyping is another factor that impacts the detection of accurate callsets [55, 56, 57]. GATK [42, 58] is a widely used tool by the livestock community for variant calling purposes. It facilitates to accurately identify differences between the reads and the reference genome [36, 59]. A combination of the software DeepVariant [60] and GLnexus [61] has emerged as an alternative, especially in human-related genomic studies, with a deep learning model trained on validated human variants [36, 62]. Comparison of the performance between GATK and DeepVariant has been performed from multiple angles in humans [36, 55]. The shared goal of both GATK and DeepVariant-GLnexus strategies is to leverage mul-

tiple samples to identify and recalibrate the genotypes of the variable genomic positions in a population [42, 60]. Joint variant calling of large cohorts can help differentiating true positive variants from sequencing errors and data processing artifacts [42, 63]. It also increases variant calling sensitivity in regions with low coverage by estimating (imputing) genotypes from similar samples [42, 63]. This substantial improvement makes joint-calling of variants computationally intensive [42, 63].

1.2 Imputation of genetic variants

Even if all the above-mentioned developments, technologies and tools are considered and optimised, not all the variants identified are confidently genotyped. lcWGS is more affordable than higher-coverage strategies but the genotyping accuracy of the resulting data is suboptimal. Genotype imputation is a statistical approach to infer unknown or low-confident genotypes from observed genotypes from other samples [64]. Genomic stretches including groups of variants that are inherited together are referred to as haplotypes [9] and are at the core of genotype imputation (see a simplified example of haplotypes in Fig. 1.3). The likelihood that two nearby variants segregate together increases with the proximity of the variants and is a phenomenon so-called linkage disequilibrium (LD) [37]. Individuals in bovine populations are more related than wild populations (given the limited effective population sizes) and typically have less LD decay [65, 66, 67]. This facilitates the imputation of variants in cattle, as there are more possibilities that information is available from neighbouring variants in tight linkage [52, 67]. Taken together, appropriate sequencing strategies and powerful imputation methods enable the generation of large datasets of individuals with sequence data at a low cost [68]. Imputation can be challenging in genomic regions with a high concentration of polymorphisms or for variants that are rare in the population (with low minor allele frequencies - MAF) [67, 69, 70].

Imputation of high coverage cohort genotypes

Imputation tools have been developed for the inference of missing genotypes in high coverage data (*e.g.*, BEAGLE [71]). The set of variants refined with BEAGLE is obtained as an imputed VCF file, a row-oriented tab-delimited text file specialised for the storage of genetic variants [72]. When properly phased (separated by paternal/maternal origin), imputed VCF files can be considered as a collection (panel) of the haplotypes present in a population. High quality reference panels contain accurate variants from a large number of samples; this is, a large breadth of haplotypes well representative of the population

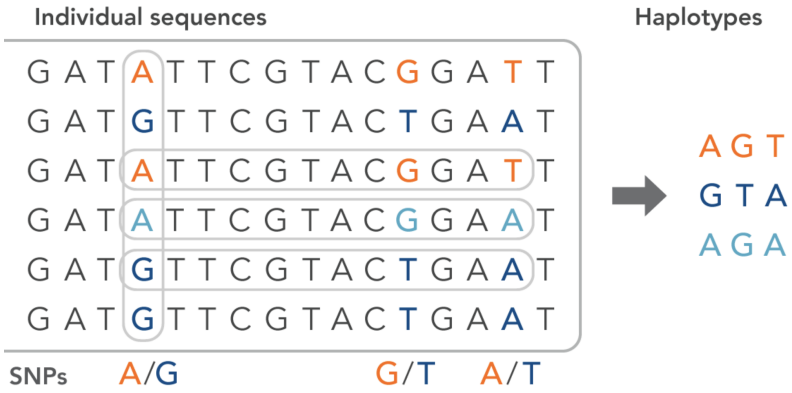


Figure 1.3: **Haplotypes.** Haplotypes made up of 3 biallelic positions. The 3 distinct haplotypes (AGT, GTA, AGA) contain biallelic SNPs (A or G, G or T, and A or T) at the 3 variant positions in this locus. Credit: Dr. Ellen Prediger / Lubio science: https://www.lubio.ch/blog/genotyping-terms/.

[63, 73, 74]. Catalogues of variants, such as the 1000 Bull Genomes consortia [4] contain information about haplotypes from different breeds and populations. Further efforts have been made to compile exhaustive catalogue of haplotypes in cattle [75, 76, 77]. The generation of reference panels was typically optimised by sequencing the key ancestors, which are expected to carry many of the high-frequency haplotypes in the population [78]. However, the access to larger scales of sequencing data enables a growing number of haplotypes from different ancestries, which increases the completeness of the panels and particularly the number of rare variants [76, 79].

Haplotype panels and their use for imputation

Haplotype reference panels contain information about the polymorphic sites that tend to segregate together in a population [75, 80]. For samples with similar ancestry to the haplotypes, they facilitate the imputation of unobserved genotypes that have been assayed using microarray chips or lcWGS data [69, 81, 82]. Imputation from sparse microarray to higher density chips by using large haplotype reference panels has been routinely performed in animal breeding and cattle genomics and different methods are available [51, 83, 84]. lcWGS is gaining momentum as an alternative for their decreasing costs in both library preparation and sequencing [85]. However, given the shallow depth of the sequencing data, imputation of lcWGS data is required to refine the genotype likelihoods and fill the gaps between sparsely mapped reads [12, 86]. The combination of sequencing and microarray data can lead to an improved imputation [50] but availability of both is infrequent [87]. Imputation of SNPs microarray data and imputation of low-pass data (see the schematic process in Fig. 1.4) have fundamental differences:

- Imputation algorithms for lcWGS provide a higher level of uncertainty of homozygotes and heterozygotes than array genotypes. Given the limited number of reads, the stochasticity of the lecture of the different alleles may lead to miss or misread the actual genotypes and so additional probability calculations are required [51].
- Low-pass data do not require a two-step imputation process, unlike what is generally applied for low-density SNP chips, with an intermediate higher density step [82, 88].
- SNP arrays require prior knowledge about segregating variants and thus can only genotype a limited number of known variants; rare variants are often not included by design [10, 89].
- Although physical distances between variants have been optimised in microarray chips, lcWGS can leverage higher LD values (denser variant span) to increase imputation accuracy [10, 90].
- Information from genotyping microarrays is summed across paternal and maternal haplotypes, while sequencing reads come from either the paternal or maternal haplotype. If the sequencing reads can be phased, imputation becomes much simpler [85].
- The benefits of imputing sequence reads *versus* genotyping arrays appear considerably greater for populations less similar to a given reference dataset [85].

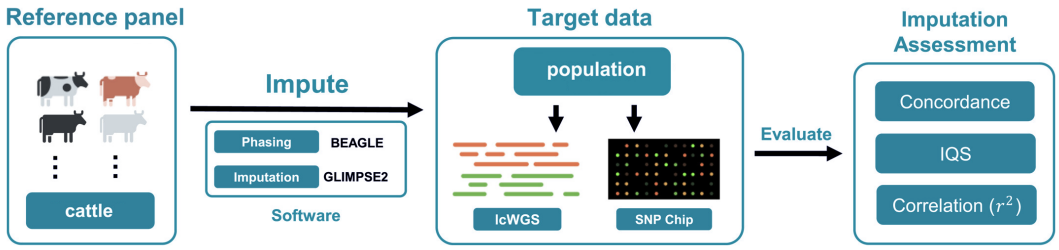


Figure 1.4: **Imputation with haplotype panels.** lcWGS and microarray data can be imputed with haplotype reference panels - adapted from [91].

Overall, in humans, lcWGS at 0.1-0.25x captures a comparable amount of variation than microarray data and lcWGS at 1-2x outperforms arrays at all frequencies, especially lower frequency variants [85, 86, 92]. The availability of large panels of reference haplotypes coupled with sequencing data of a substantial amount of the population at 2x yields high imputation accuracies [51, 76, 92]. The software GLIMPSE [86] leverages reference haplotype panels and yields slightly better performance than BEAGLE [71] for lcWGS imputation and has been successfully applied to livestock species [93]. Discrepancies exist on recommendations regarding the adequate composition of reference haplotype panels. The interplay of sequencing coverage of low-pass data, the size

and the ancestry of the haplotypes reference panels impact imputation accuracy but the extent has been hitherto underexplored in cattle [10].

Downstream analyses with imputed genotypes

High-density genotypes of hundreds or thousands of animals can be accurately obtained via imputation, be it from high- or low-coverage WGS data. Combined with the outstandingly well-documented phenotypic records from farmed species (such as fertility parameters or production traits), complex genotype-phenotype relationships can be explored at unprecedented scales [34, 94, 95]. This combination is essential for downstream analyses that enable the prediction of genomic breeding values, increase the power for genome-wide association studies (GWAS) at nucleotide resolution, fine map regions harbouring candidate variants underlying quantitative traits and provide deeper insights into function and complex trait biology [96, 97, 98, 99]. The availability of millions of imputed variants and thousands of phenotypes measured on a population scale enable powerful association tests that can detect quantitative trait loci (QTL) [100]. QTLs are genetic locations whose allele change is statistically linked to a measurable change in phenotypic traits. Numerous examples have been described in cattle [101, 102, 103]. QTL mapping is a valuable tool to elucidate the relationship between genetic variants and one of the multiple categories of molecular phenotypes (molQTL) such as gene expression (eQTL), splicing (sQTL), protein abundance, metabolomics, base methylation (meQTL), histone modification (hQTL), chromatin activity and even the interplay of such features [100, 104, 105, 106]. It is a necessary initial step towards the identification of putative causal variants that can majorly contribute to phenotypic variation [107, 108].

References

[1] M Nei and W H Li. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, 76(10): 5269–5273, October 1979. ISSN 0027-8424.

[2] Bovine HapMap Consortium, Richard A. Gibbs, Jeremy F. Taylor, Curtis P. Van Tassell, William Barendse, Kellye A. Eversole, Clare A. Gill, Ronnie D. Green, Debora L. Hamernik, Steven M. Kappes, Sigbjørn Lien, Lakshmi K. Matukumalli, John C. McEwan, et al. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science (New York, N.Y.)*, 324(5926):528–532, April 2009. ISSN 1095-9203. doi: 10.1126/science.1167936.

[3] Yangkai Liu, Haijian Cheng, Shikang Wang, Xiaoyv Luo, Xiaohui Ma, Luyang Sun, Ningbo Chen, Jicai Zhang, Kaixing Qu, Mingjin Wang, Jianyong Liu, Bizhi Huang, and Chuzhao Lei. Genomic Diversity and Selection Signatures for Weining Cattle on the Border of Yunnan-Guizhou. *Frontiers in Genetics*, 13:848951, 2022. ISSN 1664-8021. doi: 10.3389/fgene.2022.848951.

[4] Hans D. Daetwyler, Aurélien Capitan, Hubert Pausch, Paul Stothard, Rianne van Binsbergen, Rasmus F. Brøndum, Xiaoping Liao, Anis Djari, Sabrina C. Rodriguez, Cécile Grohs, Diane Esquerré, Olivier Bouchez, Marie-Noëlle Rossignol, Christophe Klopp, Dominique Rocha, Sébastien Fritz, André Eggen, Phil J. Bowman, David Coote, Amanda J. Chamberlain, Charlotte Anderson, Curt P. VanTassell, Ina Hulsegge, Mike E. Goddard, Bernt Guldbrandtsen, Mogens S. Lund, Roel F. Veerkamp, Didier A. Boichard, Ruedi Fries, and Ben J. Hayes. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genetics*, 46(8):858–865, August 2014. ISSN 1546-1718. doi: 10.1038/ng.3034.

[5] David R. Bentley, Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, Dirk J. Evers, Colin L. Barnes, Helen R. Bignell, Jonathan M. Boutell, Jason Bryant, Richard J. Carter, R. Keira Cheetham, Anthony J. Cox, Darren J. Ellis, Michael R. Flatbush, Niall A. Gormley, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, November 2008. ISSN 1476-4687. doi: 10.1038/nature07517.

[6] Miten Jain, Sergey Koren, Karen H. Miga, Josh Quick, Arthur C. Rand, Thomas A. Sasani, John R. Tyson, Andrew D. Beggs, Alexander T. Dilthey, Ian T. Fiddes, Sunir Malla, Hannah Marriott, Tom Nieto, Justin O'Grady, Hugh E. Olsen, Brent S. Pedersen, Arang Rhie, Hollian Richardson, Aaron R. Quinlan, Terrance P. Snutch, Louise Tee, Benedict Paten, Adam M. Phillippy, Jared T. Simpson, Nicholas J. Loman, and Matthew Loose. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36(4):338–345, April 2018. ISSN 1546-1696. doi: 10.1038/nbt.4060.

[7] Mick Watson and Amanda Warr. Errors in long-read assemblies can critically affect protein prediction. *Nature Biotechnology*, 37(2):124–126, February 2019. ISSN 1546-1696. doi: 10.1038/s41587-018-0004-z.

[8] Aaron M. Wenger, Paul Peluso, William J. Rowell, Pi-Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D. Olson, Armin Töpfer, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10):1155–1162, October 2019. ISSN 1546-1696. doi: 10.1038/s41587-019-0217-9.

[9] Murray Patterson, Tobias Marschall, Nadia Pisanti, Leo van Iersel, Leen Stougie, Gunnar W. Klau, and Alexander Schönhuth. WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 22(6):498–509, June 2015. ISSN 1557-8666. doi: 10.1089/cmb.2014.0157.

[10] Jun Teng, Changheng Zhao, Dan Wang, Zhi Chen, Hui Tang, Jianbin Li, Cheng Mei, Zhangping Yang, Chao Ning, and Qin Zhang. Assessment of the performance of different imputation methods for low-coverage sequencing in Holstein cattle. *Journal of Dairy Science*, 105(4):3355–3366, April 2022. ISSN 0022-0302. doi: 10.3168/jds.2021-21360.

[11] Roger Ros-Freixedes, Mara Battagin, Martin Johnsson, Gregor Gorjanc, Alan J. Mileham, Steve D. Rounsley, and John M. Hickey. Impact of index hopping and bias towards the reference allele on accuracy of genotype calls from low-coverage sequencing. *Genetics, Selection, Evolution : GSE*, 50:64, December 2018. ISSN 0999-193X. doi: 10.1186/s12711-018-0436-4.

[12] Runyang Nicolas Lou, Arne Jacobs, Aryn P. Wilder, and Nina Overgaard Therkildsen. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology*, 30(23):5966–5993, 2021. ISSN 1365-294X. doi: 10.1111/mec.16077.

[13] Giulio Formenti, Kathrin Theissinger, Carlos Fernandes, Iliana Bista, Aureliano Bombarely, Christoph Bleidorn, Claudio Ciofi, Angelica Crottini, José A. Godoy, Jacob Höglund, Joanna Malukiewicz, Alice Mouton, Rebekah A. Oomen, Sadye Paez, Per J. Palsbøll, Christophe Pampoulie, María J. Ruiz-López, Hannes Svardal, Constantina Theofanopoulou, Jan de Vries, Ann-Marie Waldvogel, Guojie Zhang, Camila J. Mazzoni, Erich D. Jarvis, Miklós Bálint, and European Reference Genome Atlas (ERGA) Consortium. The era of reference genomes in conservation genomics. *Trends in Ecology & Evolution*, 37(3):197–202, March 2022. ISSN 1872-8383. doi: 10.1016/j.tree.2021.11.008.

[14] Ann M. Mc Cartney, Kishwar Shafin, Michael Alonge, Andrey V. Bzikadze, Giulio Formenti, Arkarachai Fungtammasan, Kerstin Howe, Chirag Jain, Sergey Koren, Glennis A. Logsdon, Karen H.

Miga, Alla Mikheenko, Benedict Paten, Alaina Shumate, Daniela C. Soto, Ivan Sović, Jonathan M. D. Wood, Justin M. Zook, Adam M. Phillippy, and Arang Rhie. Chasing perfection: Validation and polishing strategies for telomere-to-telomere genome assemblies. *Nature Methods*, March 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01440-3.

[15] Mikhail Kolmogorov, Kimberley J. Billingsley, Mira Mastoras, Melissa Meredith, Jean Monlong, Ryan Lorig-Roach, Mobin Asri, Pilar Alvarez Jerez, Laksh Malik, Ramita Dewan, Xylena Reed, Rylee M. Genner, Kensuke Daida, Sairam Behera, Kishwar Shafin, Trevor Pesout, Jeshuwin Prabakaran, Paolo Carnevali, Jianzhi Yang, Arang Rhie, Sonja W. Scholz, Bryan J. Traynor, Karen H. Miga, Miten Jain, Winston Timp, Adam M. Phillippy, Mark Chaisson, Fritz J. Sedlazeck, Cornelis Blauwendraat, and Benedict Paten. Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nature Methods*, pages 1–10, September 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01993-x.

[16] Sam Kovaka, Shujun Ou, Katharine M. Jenike, and Michael C. Schatz. Approaching complete genomes, transcriptomes and epi-omes with accurate long-read sequencing. *Nature Methods*, 20(1): 12–16, January 2023. ISSN 1548-7105. doi: 10.1038/s41592-022-01716-8.

[17] Arang Rhie, Shane A. McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, William Chow, Arkarachai Fungtammasan, Juwan Kim, Chul Lee, Byung June Ko, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, 592 (7856):737–746, April 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03451-0.

[18] The Darwin Tree of Life Project Consortium. Sequence locally, think globally: The Darwin Tree of Life Project. *Proceedings of the National Academy of Sciences*, 119(4):e2115642118, January 2022. doi: 10.1073/pnas.2115642118.

[19] Harris A. Lewin, Gene E. Robinson, W. John Kress, William J. Baker, Jonathan Coddington, Keith A. Crandall, Richard Durbin, Scott V. Edwards, Félix Forest, M. Thomas P. Gilbert, Melissa M. Goldstein, Igor V. Grigoriev, Kevin J. Hackett, David Haussler, Erich D. Jarvis, Warren E. Johnson, Aristides Patrinos, Stephen Richards, Juan Carlos Castilla-Rubio, Marie-Anne van Sluys, Pamela S. Soltis, Xun Xu, Huanming Yang, and Guojie Zhang. Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of America*, 115(17):4325–4333, April 2018. ISSN 1091-6490. doi: 10.1073/pnas.1720115115.

[20] Ting Wang, Lucinda Antonacci-Fulton, Kerstin Howe, Heather A. Lawson, Julian K. Lucas, Adam M. Phillippy, Alice B. Popejoy, Mobin Asri, Caryn Carson, Mark J. P. Chaisson, Xian Chang, Robert Cook-Deegan, Adam L. Felsenfeld, Robert S. Fulton, Erik P. Garrison, Nanibaa' A. Garrison, Tina A. Graves-Lindsay, Hanlee Ji, Eimear E. Kenny, Barbara A. Koenig, Daofeng Li, Tobias Marschall, Joshua F. McMichael, Adam M. Novak, Deepak Purushotham, Valerie A. Schneider, Baergen I. Schultz, Michael W. Smith, Heidi J. Sofia, Tsachy Weissman, Paul Flicek, Heng Li, Karen H. Miga, Benedict Paten, Erich D. Jarvis, Ira M. Hall, Evan E. Eichler, and David Haussler. The Human Pangenome Project: A global resource to map genomic diversity. *Nature*, 604(7906):437–446, April 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04601-8.

[21] Sara Ballouz, Alexander Dobin, and Jesse A. Gillis. Is it time to change the reference genome? *Genome Biology*, 20(1):159, August 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1774-4.

[22] Xiaofei Yang, Wan-Ping Lee, Kai Ye, and Charles Lee. One reference genome is not enough. *Genome Biology*, 20(1):104, May 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1717-0.

[23] LaDeana W. Hillier, Webb Miller, Ewan Birney, Wesley Warren, Ross C. Hardison, Chris P. Ponting, Peer Bork, David W. Burt, Martien A. M. Groenen, Mary E. Delany, Jerry B. Dodgson, Asif T. Chinwalla, Paul F. Cliften, Sandra W. Clifton, Kimberly D. Delehaunty, Catrina Fronick, et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018):695–716, December 2004. ISSN 1476-4687. doi: 10.1038/nature03154.

[24] Lawrence B. Schook, Jonathan E. Beever, Jane Rogers, Sean Humphray, Alan Archibald, Patrick Chardon, Denis Milan, Gary Rohrer, and Kellye Eversole. Swine Genome Sequencing Consortium (SGSC): A strategic roadmap for sequencing the pig genome. *Comparative and Functional Genomics*, 6 (4):251–255, 2005. ISSN 1531-6912. doi: 10.1002/cfg.479.

[25] Christine G. Elsik, Ross L. Tellam, Kim C. Worley, and The Bovine. The Genome Sequence of Taurine Cattle: A window to ruminant biology and evolution. *Science (New York, N.Y.)*, 324(5926):522–528, April 2009. ISSN 0036-8075. doi: 10.1126/science.1169588.

[26] D. M. Bickhart, J. C. McClure, R. D. Schnabel, B. D. Rosen, J. F. Medrano, and T. P. L. Smith. Symposium review: Advances in sequencing technology herald a new frontier in cattle genomics and genome-enabled selection. *Journal of Dairy Science*, 103(6):5278–5290, June 2020. ISSN 0022-0302. doi: 10.3168/jds.2019-17693.

[27] Benjamin D. Rosen, Derek M. Bickhart, Robert D. Schnabel, Sergey Koren, Christine G. Elsik, Elizabeth Tseng, Troy N. Rowan, Wai Y. Low, Aleksey Zimin, Christine Couldrey, Richard Hall, Wenli Li, Arang Rhie, Jay Ghurye, Stephanie D. McKay, Françoise Thibaud-Nissen, Jinna Hoffman, Brenda M. Murdoch, Warren M. Snelling, Tara G. McDaneld, John A. Hammond, John C. Schwartz, Wilson Nandolo, Darren E. Hagen, Christian Dreischer, Sebastian J. Schultheiss, Steven G. Schroeder, Adam M. Phillippy, John B. Cole, Curtis P. Van Tassell, George Liu, Timothy P. L. Smith, and Juan F. Medrano. De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience*, 9(3): giaa021, March 2020. ISSN 2047-217X. doi: 10.1093/gigascience/giaa021.

[28] Sergey Koren, Arang Rhie, Brian P. Walenz, Alexander T. Dilthey, Derek M. Bickhart, Sarah B. Kingan, Stefan Hiendleder, John L. Williams, Timothy P. L. Smith, and Adam M. Phillippy. De novo assembly of haplotype-resolved genomes with trio binning. *Nature biotechnology*, page 10.1038/nbt.4277, October 2018. ISSN 1087-0156. doi: 10.1038/nbt.4277.

[29] Wai Yee Low, Rick Tearle, Ruijie Liu, Sergey Koren, Arang Rhie, Derek M. Bickhart, Benjamin D. Rosen, Zev N. Kronenberg, Sarah B. Kingan, Elizabeth Tseng, Françoise Thibaud-Nissen, Fergal J. Martin, Konstantinos Billis, Jay Ghurye, Alex R. Hastie, Joyce Lee, Andy W. C. Pang, Michael P. Heaton, Adam M. Phillippy, Stefan Hiendleder, Timothy P. L. Smith, and John L. Williams. Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. *Nature Communications*, 11(1):2071, April 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-15848-y.

[30] Edward S. Rice, Sergey Koren, Arang Rhie, Michael P. Heaton, Theodore S. Kalbfleisch, Timothy Hardy, Peter H. Hackett, Derek M. Bickhart, Benjamin D. Rosen, Brian Vander Ley, Nicholas W. Maurer, Richard E. Green, Adam M. Phillippy, Jessica L. Petersen, and Timothy P. L. Smith. Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. *GigaScience*, 9(4):giaa029, April 2020. ISSN 2047-217X. doi: 10.1093/gigascience/giaa029.

[31] Michael P. Heaton, Timothy P. L. Smith, Derek M. Bickhart, Brian L. Vander Ley, Larry A. Kuehn, Jonas Oppenheimer, Wade R. Shafer, Fred T. Schuetze, Brad Stroud, Jennifer C. McClure, Jennifer P. Barfield, Harvey D. Blackburn, Theodore S. Kalbfleisch, Kimberly M. Davenport, Kristen L. Kuhn, Richard E. Green, Beth Shapiro, and Benjamin D. Rosen. A Reference Genome Assembly of Simmental Cattle, Bos taurus taurus. *The Journal of Heredity*, 112(2):184–191, March 2021. ISSN 1465-7333. doi: 10.1093/jhered/esab002.

[32] Alexander S. Leonard, Danang Crysnanto, Zih-Hua Fang, Michael P. Heaton, Brian L. Vander Ley, Carolina Herrera, Heinrich Bollwein, Derek M. Bickhart, Kristen L. Kuhn, Timothy P. L. Smith, Benjamin D. Rosen, and Hubert Pausch. Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. *Nature Communications*, 13(1):3012, May 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-30680-2.

[33] Anders Albrechtsen, Finn Cilius Nielsen, and Rasmus Nielsen. Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. *Molecular Biology and Evolution*, 27(11):2534–2547, November 2010. ISSN 0737-4038. doi: 10.1093/molbev/msq148.

[34] Lakshmi K. Matukumalli, Cynthia T. Lawley, Robert D. Schnabel, Jeremy F. Taylor, Mark F. Allan, Michael P. Heaton, Jeff O'Connell, Stephen S. Moore, Timothy P. L. Smith, Tad S. Sonstegard, and Curtis P. Van Tassell. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. *PLOS ONE*, 4(4):e5350, April 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0005350.

[35] R. Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154(2):931–942, February 2000. ISSN 0016-6731. doi: 10.1093/genetics/154.2.931.

[36] Sen Zhao, Oleg Agafonov, Abdulrahman Azab, Tomasz Stokowy, and Eivind Hovig. Accuracy and efficiency of germline variant calling pipelines for human genome data. *Scientific Reports*, 10(1):20222, November 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-77218-4.

[37] Richard M. Durbin, David Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Francis S. Collins, Francisco M. De La Vega, Peter Donnelly, Michael Egholm, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, October 2010. ISSN 1476-4687. doi: 10.1038/nature09534.

[38] Justin Bohling. Evaluating the effect of reference genome divergence on the analysis of empirical RADseq datasets. *Ecology and Evolution*, 10(14):7585–7601, July 2020. ISSN 2045-7758. doi: 10.1002/ece3.6483.

[39] Mohammed Alser, Jeremy Rotman, Dhrithi Deshpande, Kodi Taraszka, Huwenbo Shi, Pelin Icer Baykal, Harry Taegyun Yang, Victor Xue, Sergey Knyazev, Benjamin D. Singer, Brunilda Balliu, David Koslicki, Pavel Skums, Alex Zelikovsky, Can Alkan, Onur Mutlu, and Serghei Mangul. Technology dictates algorithms: Recent developments in read alignment. *Genome Biology*, 22(1):249, August 2021. ISSN 1474-760X. doi: 10.1186/s13059-021-02443-7.

[40] Yan Guo, Yulin Dai, Hui Yu, Shilin Zhao, David C. Samuels, and Yu Shyr. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, 109(2):83–90, March 2017. ISSN 0888-7543. doi: 10.1016/j.ygeno.2017.01.005.

[41] Raphael O. Betschart, Alexandre Thiéry, Domingo Aguilera-Garcia, Martin Zoche, Holger Moch, Raphael Twerenbold, Tanja Zeller, Stefan Blankenberg, and Andreas Ziegler. Comparison of calling pipelines for whole genome sequencing: An empirical study demonstrating the importance of mapping and alignment. *Scientific Reports*, 12(1):21502, December 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-26181-3.

[42] Mark A. DePristo, Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, Guillermo del Angel, Manuel A. Rivas, Matt Hanna, Aaron McKenna, Tim J. Fennell, Andrew M. Kernytsky, Andrey Y. Sivachenko, Kristian Cibulskis, Stacey B. Gabriel, David Altshuler, and Mark J. Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498, May 2011. ISSN 1546-1718. doi: 10.1038/ng.806.

[43] Mao-Jan Lin, Sheila Iyer, Nae-Chyun Chen, and Ben Langmead. Measuring, visualizing and diagnosing reference bias with biastools, September 2023.

[44] Alexander T. Dilthey, Sebastian A. Meyer, and Achim J. Kaasch. Ultraplexing: Increasing the efficiency of long-read sequencing for hybrid assembly with k-mer-based multiplexing. *Genome Biology*, 21(1):68, March 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-01974-9.

[45] Miten Jain, Hugh E. Olsen, Daniel J. Turner, David Stoddart, Kira V. Bulazel, Benedict Paten, David Haussler, Huntington F. Willard, Mark Akeson, and Karen H. Miga. Linear assembly of a human centromere on the Y chromosome. *Nature Biotechnology*, 36(4):321–323, April 2018. ISSN 1546-1696. doi: 10.1038/nbt.4109.

[46] Karen H. Miga, Sergey Koren, Arang Rhie, Mitchell R. Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, Edmund Howe, David Porubsky, Glennis A. Logsdon, Valerie A. Schneider, Tamara Potapova, et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, 585 (7823):79–84, September 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2547-7.

[47] Scott Gigante, Quentin Gouil, Alexis Lucattini, Andrew Keniry, Tamara Beck, Matthew Tinning, Lavinia Gordon, Chris Woodruff, Terence P. Speed, Marnie E. Blewitt, and Matthew E. Ritchie. Using long-read sequencing to detect imprinted DNA methylation. *Nucleic Acids Research*, 47(8):e46, May 2019. ISSN 1362-4962. doi: 10.1093/nar/gkz107.

[48] Jared T. Simpson, Rachael E. Workman, P. C. Zuzarte, Matei David, L. J. Dursi, and Winston Timp. Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, 14(4):407–410, April 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4184.

[49] O. Y. Olivia Tse, Peiyong Jiang, Suk Hang Cheng, Wenlei Peng, Huimin Shang, John Wong, Stephen L. Chan, Liona C. Y. Poon, Tak Y. Leung, K. C. Allen Chan, Rossa W. K. Chiu, and Y. M. Dennis Lo. Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *Proceedings of the National Academy of Sciences*, 118(5):e2019768118, February 2021. doi: 10.1073/pnas.2019768118.

[50] Androniki Menelaou and Jonathan Marchini. Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics*, 29(1):84–91, January 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts632.

[51] Paul M. VanRaden, Chuanyu Sun, and Jeffrey R. O'Connell. Fast imputation using medium or low-coverage sequence data. *BMC genetics*, 16:82, July 2015. ISSN 1471-2156. doi: 10.1186/s12863-015-0243-7.

[52] Ruoyun Hui, Eugenia D'Atanasio, Lara M. Cassidy, Christiana L. Scheib, and Toomas Kivisild. Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. *Scientific Reports*, 10(1):18542, October 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-75387-w.

[53] Michael C. Wendl and Richard K. Wilson. The theory of discovering rare variants via DNA sequencing. *BMC Genomics*, 10(1):485, October 2009. ISSN 1471-2164. doi: 10.1186/1471-2164-10-485.

[54] Si Quang Le and Richard Durbin. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Research*, 21(6):952–960, June 2011. ISSN 1549-5469. doi: 10.1101/gr.113084.110.

[55] Anna Supernat, Oskar Valdimar Vidarsson, Vidar M. Steen, and Tomasz Stokowy. Comparison of three variant callers for human whole genome sequencing. *Scientific Reports*, 8(1):17851, December 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-36177-7.

[56] Manojkumar Kumaran, Umadevi Subramanian, and Bharanidharan Devarajan. Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. *BMC Bioinformatics*, 20(1):342, June 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-2928-9.

[57] Omar Abdelwahab, Francois Belzile, and Davoud Torkamaneh. Performance analysis of conventional and AI-based variant callers using short and long reads, August 2023.

[58] Geraldine A. Van der Auwera, Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo Del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Kiran V. Garimella, David Altshuler, Stacey Gabriel, and Mark A. DePristo. From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 43:11.10.1–11.10.33, 2013. ISSN 1934-340X. doi: 10.1002/0471250953.bi1110s43.

[59] Yi-Lin Lin, Pi-Chuan Chang, Ching Hsu, Miao-Zi Hung, Yin-Hsiu Chien, Wuh-Liang Hwu, FeiPei Lai, and Ni-Chung Lee. Comparison of GATK and DeepVariant by trio sequencing. *Scientific Reports*, 12:1809, February 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-05833-4.

[60] Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T. Afshar, Sam S. Gross, Lizzie Dorfman, Cory Y. McLean, and Mark A. DePristo. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10):983–987, November 2018. ISSN 1546-1696. doi: 10.1038/nbt.4235.

[61] Michael F. Lin, Ohad Rodeh, John Penn, Xiaodong Bai, Jeffrey G. Reid, Olga Krasheninina, and William J. Salerno. GLnexus: Joint variant calling for large cohort sequencing, June 2018.

[62] Taedong Yun, Helen Li, Pi-Chuan Chang, Michael F. Lin, Andrew Carroll, and Cory Y. McLean. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics (Oxford, England)*, page btaa1081, January 2021. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaa1081.

[63] Ryan Poplin, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, Laura D. Gauthier, Ami Levy-Moonshine, David Roazen, Khalid Shakir, Joel Thibault, Sheila Chandran, Chris Whelan, Monkol Lek, Stacey Gabriel, Mark J. Daly, Ben Neale, Daniel G. MacArthur, and Eric Banks. Scaling accurate genetic variant discovery to tens of thousands of samples, July 2018.

[64] Paul Scheet and Matthew Stephens. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, 78(4):629–644, April 2006. ISSN 0002-9297. doi: 10.1086/502802.

[65] Didier Boichard, Hoyoung Chung, Romain Dassonneville, Xavier David, André Eggen, Sébastien Fritz, Kimberly J. Gietzen, Ben J. Hayes, Cynthia T. Lawley, Tad S. Sonstegard, Curtis P. Van Tassell, Paul M. VanRaden, Karine A. Viaud-Martinez, George R. Wiggans, and for the Bovine LD Consortium. Design of a Bovine Low-Density SNP Array Optimized for Imputation. *PLOS ONE*, 7(3):e34130, March 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0034130.

[66] Roger Ros-Freixedes, Serap Gonen, Gregor Gorjanc, and John M. Hickey. A method for allocating low-coverage sequencing resources by targeting haplotypes rather than individuals. *Genetics, selection, evolution: GSE*, 49(1):78, October 2017. ISSN 1297-9686. doi: 10.1186/s12711-017-0353-y.

[67] Sanne van den Berg, Jérémie Vandenplas, Fred A. van Eeuwijk, Aniek C. Bouwman, Marcos S. Lopes, and Roel F. Veerkamp. Imputation to whole-genome sequence using multiple pig populations and its use in genome-wide association studies. *Genetics, Selection, Evolution : GSE*, 51:2, January 2019. ISSN 0999-193X. doi: 10.1186/s12711-019-0445-y.

[68] Roger Ros-Freixedes, Andrew Whalen, Gregor Gorjanc, Alan J. Mileham, and John M. Hickey. Evaluation of sequencing strategies for whole-genome imputation with hybrid peeling. *Genetics, selection, evolution: GSE*, 52(1):18, April 2020. ISSN 1297-9686. doi: 10.1186/s12711-020-00537-7.

[69] Warren M. Snelling, Jesse L. Hoff, Jeremiah H. Li, Larry A. Kuehn, Brittney N. Keel, Amanda K. Lindholm-Perry, and Joseph K. Pickrell. Assessment of Imputation from Low-Pass Sequencing to Predict Merit of Beef Steers. *Genes*, 11(11):E1312, November 2020. ISSN 2073-4425. doi: 10.3390/genes11111312.

[70] Muhammad Yasir Nawaz, Priscila Arrigucci Bernardes, Rodrigo Pelicioni Savegnago, Dajeong Lim, Seung Hwan Lee, and Cedric Gondro. Evaluation of Whole-Genome Sequence Imputation Strategies in Korean Hanwoo Cattle. *Animals*, 12(17):2265, January 2022. ISSN 2076-2615. doi: 10.3390/ani12172265.

[71] Brian L. Browning and Sharon R. Browning. Genotype Imputation with Millions of Reference Samples. *American Journal of Human Genetics*, 98(1):116–126, January 2016. ISSN 1537-6605. doi: 10.1016/j.ajhg.2015.11.020.

[72] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):giab008, February 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab008.

[73] Roger Ros-Freixedes, Andrew Whalen, Ching-Yi Chen, Gregor Gorjanc, William O. Herring, Alan J. Mileham, and John M. Hickey. Accuracy of whole-genome sequence imputation using hybrid peeling in large pedigreed livestock populations. *Genetics, selection, evolution: GSE*, 52(1):17, April 2020. ISSN 1297-9686. doi: 10.1186/s12711-020-00536-8.

[74] Jared O'Connell, Taedong Yun, Meghan Moreno, Helen Li, Nadia Litterman, Alexey Kolesnikov, Elizabeth Noblin, Pi-Chuan Chang, Anjali Shastri, Elizabeth H. Dorfman, Suyash Shringarpure, Adam Auton, Andrew Carroll, and Cory Y. McLean. A population-specific reference panel for improved genotype imputation in African Americans. *Communications Biology*, 4:1269, November 2021. ISSN 2399-3642. doi: 10.1038/s42003-021-02777-9.

[75] Wenqian Yang, Yanbo Yang, Cecheng Zhao, Kun Yang, Dongyang Wang, Jiajun Yang, Xiaohui Niu, and Jing Gong. Animal-ImputeDB: A comprehensive database with multiple animal reference panels for genotype imputation. *Nucleic Acids Research*, 48(D1):D659–D667, January 2020. ISSN 1362-4962. doi: 10.1093/nar/gkz854.

[76] Hoff J. Gencove's improved cattle low-pass sequencing solution, 2021. URL https://medium.com/the-gencove-blog/gencove-v3-cattle-panel-a26c88ccb01a.

[77] Li J. Gencove's new pig haplotype reference panel, 2021. URL https://medium.com/the-gencove-blog/gencoves-new-pig-haplotype-reference-panel-6860b849f01f.

[78] Serap Gonen, Roger Ros-Freixedes, Mara Battagin, Gregor Gorjanc, and John M. Hickey. A method for the allocation of sequencing resources in genotyped livestock populations. *Genetics Selection Evolution*, 49(1):47, May 2017. ISSN 1297-9686. doi: 10.1186/s12711-017-0322-5.

[79] Zhe Zhang, Peipei Ma, Zhenyang Zhang, Zhen Wang, Qishan Wang, and Yuchun Pan. The construction of a haplotype reference panel using extremely low coverage whole genome sequences and its application in genome-wide association studies and genomic prediction in Duroc pigs. *Genomics*, 114 (1):340–350, January 2022. ISSN 1089-8646. doi: 10.1016/j.ygeno.2021.12.016.

[80] International HapMap Consortium, Kelly A. Frazer, Dennis G. Ballinger, David R. Cox, David A. Hinds, Laura L. Stuve, Richard A. Gibbs, John W. Belmont, Andrew Boudreau, Paul Hardenbol, Suzanne M. Leal, Shiran Pasternak, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, October 2007. ISSN 1476-4687. doi: 10.1038/nature06258.

[81] Shane McCarthy, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R. Wood, Alexander Teumer, Hyun Min Kang, Christian Fuchsberger, Petr Danecek, Kevin Sharp, Yang Luo, Carlo Sidore, Alan Kwong, Nicholas Timpson, Seppo Koskinen, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, 48(10):1279–1283, October 2016. ISSN 1546-1718. doi: 10.1038/ng.3643.

[82] Troy N. Rowan, Jesse L. Hoff, Tamar E. Crum, Jeremy F. Taylor, Robert D. Schnabel, and Jared E. Decker. A multi-breed reference panel and additional rare variants maximize imputation accuracy in cattle. *Genetics Selection Evolution*, 51(1):77, December 2019. ISSN 1297-9686. doi: 10.1186/s12711 -019-0519-x.

[83] John M Hickey, Brian P Kinghorn, Bruce Tier, Julius HJ van der Werf, and Matthew A Cleveland. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. *Genetics, Selection, Evolution : GSE*, 44(1):9, June 2012. ISSN 0999-193X. doi: 10.1186/1297 -9686-44-9.

[84] Mehdi Sargolzaei, Jacques P. Chesnais, and Flavio S. Schenkel. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics*, 15(1):478, June 2014. ISSN 1471-2164. doi: 10.1186/1471-2164-15-478.

[85] Robert W. Davies, Marek Kucka, Dingwen Su, Sinan Shi, Maeve Flanagan, Christopher M. Cunniff, Yingguang Frank Chan, and Simon Myers. Rapid genotype imputation from sequence with reference panels. *Nature Genetics*, 53(7):1104–1111, July 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00877-0.

[86] Simone Rubinacci, Diogo M. Ribeiro, Robin J. Hofmeister, and Olivier Delaneau. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nature Genetics*, 53(1): 120–126, January 2021. ISSN 1546-1718. doi: 10.1038/s41588-020-00756-0.

[87] NeoGen. Neogen® and gencove launch infiniseek™ — the first whole genome and genotyping sequencing solution for cattle breeders, 2022. URL https://www.neogen.com/en-gb/neocenter/press-releases/neogen-gencove-launch-infiniseek-first-whole-genome-genotyping-sequencing-solution-cattle-breeders/.

[88] Hubert Pausch, Iona M. MacLeod, Ruedi Fries, Reiner Emmerling, Phil J. Bowman, Hans D. Daetwyler, and Michael E. Goddard. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. *Genetics, selection, evolution: GSE*, 49(1):24, February 2017. ISSN 1297-9686. doi: 10.1186/s12711-017-0301-x.

[89] Jérôme Nicod, Robert W. Davies, Na Cai, Carl Hassett, Leo Goodstadt, Cormac Cosgrove, Benjamin K. Yee, Vikte Lionikaite, Rebecca E. McIntyre, Carol Ann Remme, Elisabeth M. Lodder, Jennifer S. Gregory, Tertius Hough, Russell Joynson, Hayley Phelps, Barbara Nell, Clare Rowe, et al. Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing. *Nature Genetics*, 48(8):912–918, August 2016. ISSN 1546-1718. doi: 10.1038/ng.3595.

[90] P. M. VanRaden, D. J. Null, M. Sargolzaei, G. R. Wiggans, M. E. Tooker, J. B. Cole, T. S. Sonstegard, E. E. Connor, M. Winters, J. B. C. H. M. van Kaam, A. Valentini, B. J. Van Doormaal, M. A. Faust, and G. A. Doak. Genomic imputation and evaluation using high-density Holstein genotypes. *Journal of Dairy Science*, 96(1):668–678, January 2013. ISSN 1525-3198. doi: 10.3168/jds.2012-5702.

[91] Zhuangbiao Zhang, Ao Wang, Honghong Hu, Lulu Wang, Mian Gong, Qimeng Yang, Anguo Liu, Ran Li, Huanhuan Zhang, Qianqian Zhang, Ali Mujtaba Shah, Xihong Wang, Yachun Wang, Quanzhong Liu, Liutao Gao, Zhipeng Zhang, Congyong Wang, Yun Ma, Yudong Cai, and Yu Jiang. The efficient phasing and imputation pipeline of low-coverage whole genome sequencing data using a high-quality and publicly available reference panel in cattle. *Animal Research and One Health*, n/a (n/a), 2023. ISSN 2835-5075. doi: 10.1002/aro2.8.

[92] Bogdan Pasaniuc, Nadin Rohland, Paul J. McLaren, Kiran Garimella, Noah Zaitlen, Heng Li, Namrata Gupta, Benjamin M. Neale, Mark J. Daly, Pamela Sklar, Patrick F. Sullivan, Sarah Bergen, Jennifer L. Moran, Christina M. Hultman, Paul Lichtenstein, Patrik Magnusson, Shaun M. Purcell, David W. Haas, Liming Liang, Shamil Sunyaev, Nick Patterson, Paul I. W. de Bakker, David Reich, and Alkes L. Price. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genetics*, 44(6):631–635, May 2012. ISSN 1546-1718. doi: 10.1038/ng.2283.

[93] Adéla Nosková, Meenu Bhati, Naveen Kumar Kadri, Danang Crysnanto, Stefan Neuenschwander, Andreas Hofer, and Hubert Pausch. Characterization of a haplotype-reference panel for genotyping by low-pass sequencing in Swiss Large White pigs. *BMC genomics*, 22(1):290, April 2021. ISSN 1471-2164. doi: 10.1186/s12864-021-07610-5.

[94] Leif Andersson, Alan L. Archibald, Cynthia D. Bottema, Rudiger Brauning, Shane C. Burgess, Dave W. Burt, et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biology*, 16:57, March 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0622-4.

[95] James A. Roth and Christopher K. Tuggle. Livestock Models in Translational Medicine. *ILAR Journal*, 56(1):1–6, May 2015. ISSN 1084-2020. doi: 10.1093/ilar/ilv011.

[96] Martijn F. L. Derks, Arne B. Gjuvsland, Mirte Bosse, Marcos S. Lopes, Maren van Son, Barbara Harlizius, Beatrice F. Tan, Hanne Hamland, Eli Grindflek, Martien A. M. Groenen, and Hendrik-Jan Megens. Loss of function mutations in essential genes cause embryonic lethality in pigs. *PLoS genetics*, 15(3):e1008055, March 2019. ISSN 1553-7404. doi: 10.1371/journal.pgen.1008055.

[97] Maya Hiltpold, Fredi Janett, Xena Marie Mapel, Naveen Kumar Kadri, Zih-Hua Fang, Hermann Schwarzenbacher, Franz R. Seefried, Mirjam Spengeler, Ulrich Witschi, and Hubert Pausch. A 1-bp deletion in bovine QRICH2 causes low sperm count and immotile sperm with multiple morphological abnormalities. *Genetics Selection Evolution*, 54(1):18, March 2022. ISSN 1297-9686. doi: 10.1186/s12711-022-00710-0.

[98] Roger Ros-Freixedes, Martin Johnsson, Andrew Whalen, Ching-Yi Chen, Bruno D. Valente, William O. Herring, Gregor Gorjanc, and John M. Hickey. Genomic prediction with whole-genome sequence data in intensely selected pig lines. *Genetics Selection Evolution*, 54(1):65, September 2022. ISSN 1297-9686. doi: 10.1186/s12711-022-00756-0.

[99] Emily L. Clark, Alan L. Archibald, Hans D. Daetwyler, Martien A. M. Groenen, Peter W. Harrison, Ross D. Houston, Christa Kühn, Sigbjørn Lien, Daniel J. Macqueen, James M. Reecy, Diego Robledo, Mick Watson, Christopher K. Tuggle, and Elisabetta Giuffra. From FAANG to fork: Application of highly annotated genomes to improve farmed animal production. *Genome Biology*, 21:285, November 2020. ISSN 1474-7596. doi: 10.1186/s13059-020-02197-8.

[100] Olivier Delaneau, Halit Ongen, Andrew A. Brown, Alexandre Fort, Nikolaos I. Panousis, and Emmanouil T. Dermitzakis. A complete tool set for molecular QTL discovery and analysis. *Nature Communications*, 8(1):15452, May 2017. ISSN 2041-1723. doi: 10.1038/ncomms15452.

[101] Maya Hiltpold, Guanglin Niu, Naveen Kumar Kadri, Danang Crysnanto, Zih-Hua Fang, Mirjam Spengeler, Fritz Schmitz-Hsu, Christian Fuerst, Hermann Schwarzenbacher, Franz R. Seefried, Frauke Seehusen, Ulrich Witschi, Angelika Schnieke, Ruedi Fries, Heinrich Bollwein, Krzysztof Flisikowski, and Hubert Pausch. Activation of cryptic splicing in bovine WDR19 is associated with reduced semen quality and male fertility. *PLoS genetics*, 16(5):e1008804, May 2020. ISSN 1553-7404. doi: 10.1371/journal.pgen.1008804.

[102] Maya Hiltpold, Naveen Kumar Kadri, Fredi Janett, Ulrich Witschi, Fritz Schmitz-Hsu, and Hubert Pausch. Autosomal recessive loci contribute significantly to quantitative variation of male fertility in a dairy cattle population. *BMC genomics*, 22(1):225, March 2021. ISSN 1471-2164. doi: 10.1186/s12864-021-07523-3.

[103] Xena Marie Mapel, Maya Hiltpold, Naveen Kumar Kadri, Ulrich Witschi, and Hubert Pausch. Bull fertility and semen quality are not correlated with dairy and production traits in Brown Swiss cattle. *JDS Communications*, 3(2):120–125, March 2022. ISSN 2666-9102, 2666-9102. doi: 10.3168/jdsc.2021-0164.

[104] Claire P. Prowse-Wilkins, Thomas J. Lopdell, Ruidong Xiang, Christy J. Vander Jagt, Mathew D. Littlejohn, Amanda J. Chamberlain, and Michael E. Goddard. Genetic variation in histone modifications and gene expression identifies regulatory variants in the mammary gland of cattle. *BMC genomics*, 23(1):815, December 2022. ISSN 1471-2164. doi: 10.1186/s12864-022-09002-9.

[105] Yong Zeng, Rahi Jain, Musaddeque Ahmed, Haiyang Guo, Yuan Zhong, Wei Xu, and Housheng Hansen He. Memo-eQTL: DNA methylation modulated genetic variant effect on gene transcriptional regulation, May 2023.

[106] Can Yuan, Lijing Tang, Thomas Lopdell, Vyacheslav A. Petrov, Claire Oget-Ebrad, Gabriel Costa Monteiro Moreira, José Luis Gualdrón Duarte, Arnaud Sartelet, Zhangrui Cheng, Mazdak Salavati, Claire D. Wathes, Mark A. Crowe, GplusE Consortium, Wouter Coppieters, Mathew Littlejohn, Carole Charlier, Tom Druet, Michel Georges, and Haruko Takeda. An organism-wide ATAC-seq peak catalogue for the bovine and its use to identify regulatory variants. *Genome Research*, page gr.277947.123, September 2023. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.277947.123.

[107] C.p. Prowse-Wilkins, T.j. Lopdell, R. Xiang, C.j. Vander Jagt, M.d. Littlejohn, A.j. Chamberlain, and M.e. Goddard. 552. Regulatory QTL and exon expression QTL in the mammary gland of dairy cows. In *Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP)*, chapter 552, pages 2289–2292. Wageningen Academic Publishers, December 2022. doi: 10.3920/978-90-8686-940-4_552.

[108] Meritxell Oliva, Kathryn Demanelis, Yihao Lu, Meytal Chernoff, Farzana Jasmine, Habibul Ahsan, Muhammad G. Kibriya, Lin S. Chen, and Brandon L. Pierce. DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nature Genetics*, 55(1):112–122, January 2023. ISSN 1546-1718. doi: 10.1038/s41588-022-01248-z.

Chapter 2

Investigating the impact of reference assembly choice on genomic analyses in a cattle breed

Audald Lloret-Villas[1], Meenu Bhati[1], Naveen Kumar Kadri[1], Ruedi Fries[2] and Hubert Pausch[1]

[1] Animal Genomics, ETH Zürich, Zürich, Switzerland.
[2] Chair of Animal Breeding, TU München, Freising-Weihenstephan, Germany.

Published in *BMC Genomics (2021) 19;22(1):363*

Contribution: I participated in conceiving the study, designing the experiments, analysing the results and writing the manuscript. I also developed and compiled reproducible pipelines.

Abstract

Background: Reference-guided read alignment and variant genotyping are prone to reference allele bias, particularly for samples that are greatly divergent from the reference genome. A Hereford-based assembly is the widely accepted bovine reference genome. Haplotype-resolved genomes that exceed the current bovine reference genome in quality and continuity have been assembled for different breeds of cattle. Using whole genome sequencing data of 161 Brown Swiss cattle, we compared the accuracy of read mapping and sequence variant genotyping as well as downstream genomic analyses between the bovine reference genome (ARS-UCD1.2) and a highly continuous Angus-based assembly (UOA_Angus_1).

Results: Read mapping accuracy did not differ notably between the ARS-UCD1.2 and UOA_Angus_1 assemblies. We discovered 22,744,517 and 22,559,675 high-quality variants from ARS-UCD1.2 and UOA_Angus_1, respectively. The concordance between sequence- and array-called genotypes was high and the number of variants deviating from Hardy-Weinberg proportions was low at segregating sites for both assemblies. More artefactual INDELs were genotyped from UOA_Angus_1 than ARS-UCD1.2 alignments. Using the composite likelihood ratio test, we detected 40 and 33 signatures of selection from ARS-UCD1.2 and UOA_Angus_1, respectively, but the overlap between both assemblies was low. Using the 161 sequenced Brown Swiss cattle as a reference panel, we imputed sequence variant genotypes into a mapping cohort of 30,499 cattle that had microarray-derived genotypes using a two-step imputation approach. The accuracy of imputation (Beagle R^2) was very high (0.87) for both assemblies. Genome-wide association studies between imputed sequence variant genotypes and six dairy traits as well as stature produced almost identical results from both assemblies.

Conclusions: The ARS-UCD1.2 and UOA_Angus_1 assemblies are suitable for reference-guided genome analyses in Brown Swiss cattle. Although differences in read mapping and genotyping accuracy between both assemblies are negligible, the choice of the reference genome has a large impact on detecting signatures of selection that already reached fixation using the composite likelihood ratio test. We developed a workflow that can be adapted and reused to compare the impact of reference genomes on genome analyses in various breeds, populations and species.

Keywords: Reference genome comparison, Bovine, Alignment quality, Sequence variants, Functional annotation, Signatures of selection, Genome-wide association study

2.1 Background

Representative reference genomes are paramount for genome research. A reference genome is an assembly of digital nucleotides that are representative of a species' genetic constitution. Like the coordinate system of a two-dimensional map, the coordinates of the reference genome unambiguously point to nucleotides and annotated genomic features. Because the physical position and alleles of sequence variants are determined according to reference coordinates, the adoption of a universal reference genome is required to compare findings across studies. Otherwise, the conversion of genomic coordinates between assemblies is necessary [1]. Updates and amendments to the reference genome change the coordinate system.

Reference genomes of important farm animal species including cattle, pig and chicken were assembled more than a decade ago using bacterial artificial chromosome and whole-genome shotgun sequencing [2, 3, 4]. The initial reference genome of domestic cattle (Bos taurus taurus) was generated from a DNA sample of the inbred Hereford cow L1 Dominette 01449 [3, 5]. An annotated bovine reference genome enabled systematic assessment and characterization of sequence variation within and between cattle populations using reference-guided alignment and variant detection [3, 6]. A typical genome-wide alignment of DNA sequences from a *B. taurus taurus* individual differs at between 6 and 8 million single nucleotide polymorphisms (SNPs) and small ($<$ 50 bp) insertions and deletions (INDELs) from the reference genome [7, 8]. More variants are detected in cattle with greater genetic distance from the Hereford breed [9]. The bovine reference genome neither contains allelic variation nor nucleotides that are private to animals other than L1 Dominette 01449. As a result, read alignments may be erroneous particularly at genomic regions that differ substantially between the sequenced individual and the reference genome [10]. The use of consensus reference genomes or variation-aware reference graphs may mitigate this type of bias [11, 12, 13].

The quality of reference genomes improved spectacularly over the past 15 years. Decreasing error rates and increasing outputs of long-read ($>$ 10 Kb) sequencing technologies such as PacBio single molecule real-time (SMRT) [14] and Oxford Nanopore sequencing [15] revolutionised the assembly of reference genomes. Sophisticated genome assembly methods enable to assemble gigabase-sized and highly-repetitive genomes from long sequencing reads at high continuity and accuracy [16, 17, 18]. The application of "trio-binning" [19] facilitates the *de novo* assembly of haplotype-resolved genomes that exceed in quality and continuity all previously assembled reference genomes. This approach now offers an opportunity to obtain reference-quality genome assemblies and

identify hitherto undetected variants in non-reference sequences, thus making the full spectrum of sequence variation amenable to genetic analyses [17, 19].

Reference-quality assemblies are available for Hereford (ARS-UCD1.2) [20], Angus (UOA_Angus_1) [17] and Highland cattle [21]. In addition, reference-quality assemblies are available for yak (*Bos grunniens*) [21] and Brahman (*Bos taurus indicus*) [17] which are closely related to taurine cattle. Any of these resources may serve as a reference for reference-guided sequence read alignment, variant detection and annotation. Linear mapping and sequence variant genotyping accuracy may be affected by the choice of the reference genome and the divergence of the DNA sample from the reference genome [22, 23, 24, 25]. It remains an intriguing question, which reference genome enables optimum read mapping and variant detection accuracy for a particular animal [11, 12, 13].

Here, we assessed the accuracy of reference-guided read mapping and sequence variant detection in 161 Brown Swiss (BSW) cattle using two highly continuous bovine genome assemblies that were created from Hereford (ARS-UCD1.2) and Angus (UOA_-Angus_1) cattle. Moreover, we detect signatures of selection and perform sequence-based association studies to investigate the impact of the reference genome on down-stream genomic analyses.

2.2 Results

Short paired-end whole-genome sequencing reads of 161 BSW cattle (113 males, 48 females) were considered for our analysis. All raw sequencing data are publicly available at the Sequencing Read Archive of the NCBI [26] or the European Nucleotide Archive of the EMBL-EBI [27] (see accession IDs in Additional file 1).

Alignment quality and depth of coverage

Following the removal of adapter sequences, and reads and bases of low sequencing quality, between 173 and 1,411 million reads per sample (mean $\pm$ standard deviation: 360 $\pm$ 165 million reads) were aligned to expanded versions of the Hereford-based ARS-UCD1.2 and the Angus-based UOA_Angus_1 assemblies that included sex chromosomal sequences and unplaced scaffolds (see Methods) using a reference-guided alignment approach. The Hereford assembly is a primary assembly because it was created from a purebred animal [20]. The Angus assembly is haplotype-resolved because it was created from an Angus x Brahman cross using "trio-binning" [17]. The average num-

ber of reads per sample that aligned to sex chromosomes, the mitochondrial genome and unplaced contigs were slightly higher for UOA_Angus_1 (66 $\pm$ 39 million) than ARS-UCD1.2 (64 $\pm$ 38 million).

We considered the 29 autosomes to investigate alignment quality. The total length of the autosomes was 2,489,385,779 bp for ARS-UCD1.2 and 2,468,157,877 bp for UOA_Angus_1. An average number of 295 $\pm$ 131 and 293 $\pm$ 130 million reads per sample aligned to autosomal sequences of ARS-UCD1.2 and UOA_Angus_1, respectively. The slightly higher number of reads that mapped to ARS-UCD1.2 is likely due to its longer autosomal sequence. In order to ensure consistency across all analyses performed, we retained 263 $\pm$ 118 (89.28%) and 261 $\pm$ 117 (89.17%) uniquely mapped and properly paired reads (*i.e.*, all reads except those with a SAM-flag value of 1796) that had mapping quality higher than 10 (high-quality reads hereafter) per sample, as such reads qualify for sequence variant genotyping using the best practice guidelines of the Genome Analysis Toolkit (GATK) [28, 29] (Table 2.1). The number of reads that mapped to the autosomes but were discarded due to low mapping quality (either SAM-flag 1796 or MQ $<$ 10) were almost identical (32 $\pm$ 20 million) for both assemblies (Additional file 2). Most of the discarded reads (83.37% for ARS-UCD1.2 and 82.29% for UOA_Angus_1) were flagged as duplicates.

Table 2.1: **Mapping statistics for the 161 BSW samples**. Summary statistics extracted from the BAM files after aligning the samples to either the ARS-UCD1.2 or UOA_Angus_1 assembly. Uniquely mapped and properly paired reads with MQ $>$ 10 are considered as high-quality reads. The percentage of autosomal reads that are high-quality reads is calculated per sample and per chromosome. Coverage of high-quality reads is calculated per sample and per chromosome.

Parameter	Unit	ARS-UCD1.2	UOA_Angus_1
Autosomal reads	Million	47,502	47,128
	Million / sample	295 $\pm$ 131	293 $\pm$ 130
Autosomal high-quality reads	Million	42,418	42,029
	Million / sample	263 $\pm$ 118	261 $\pm$ 117
	% / sample	89.28 $\pm$ 5.06	89.17 $\pm$ 5.06
	% / chromosome	89.28 $\pm$ 0.34	89.17 $\pm$ 0.56
Coverage	fold / sample	14.13 $\pm$ 7.26	14.11 $\pm$ 7.25
	fold / chromosome	14.13 $\pm$ 0.14	14.11 $\pm$ 0.15

The mean percentage of high-quality reads was slightly higher (0.10 $\pm$ 0.63) for the ARS-UCD1.2 than UOA_Angus_1 autosomes but greater differences existed at some chromosomes. The proportion of high-quality reads was higher for the ARS-UCD1.2 as-

sembly than the UOA_Angus_1 assembly at 16 out of the 29 autosomes. The greatest difference was observed for chromosome 20, for which the proportion of high-quality reads was 2.03 percent points greater for the ARS-UCD1.2 assembly than the UOA_Angus_1 assembly ($P = 4.5 \times 10^{-4}$). Of 8.59 $\pm$ 3.81 and 8.69 $\pm$ 3.88 million reads that aligned to chromosome 20 of ARS-UCD1.2 and UOA_Angus_1, respectively, 7.66 $\pm$ 3.42 and 7.57 $\pm$ 3.38 million were high-quality reads. Among the 13 autosomes for which the percentage of high-quality reads was greater for the UOA_Angus_1 than ARS-UCD1.2 assembly, the greatest difference (0.75 percent points) was observed for chromosome 13.

Average genome coverage ranged from 8.8- to 62.4-fold per sample for both assemblies. The mean coverage of the BAM files was nearly identical for the ARS-UCD1.2 (14.13 $\pm$ 7.26) and UOA_Angus_1 (14.11 $\pm$ 7.25) assembly. Chromosome wise, no differences were detected ($P = 0.36$) across the two assemblies considered. The mean coverage was between 13.76 (chromosome 19) and 14.45 (chromosome 27) for ARS-UCD1.2 and between 13.76 (chromosome 19) and 14.52 (chromosome 14) for UOA_Angus_1.

Sequence variant genotyping and variant statistics

Single nucleotide polymorphisms (SNPs), insertions and deletions (INDELs) were discovered from the BAM files following the GATK best practice guidelines [28, 29]. Using the HaplotypeCaller and GenotypeGVCFs modules of GATK, we detected 24,760,861 and 24,557,291 autosomal variants from the ARS-UCD1.2 and UOA_Angus_1 alignments, respectively, of which 22,744,517 (91.86%) and 22,559,675 (91.87%) high-quality variants were retained after applying site-level hard filtration using the VariantFiltration module of GATK (Additional file 3). The mean transition/transversion ratio was 2.15 for the high-quality variants detected from either of the assemblies.

For 32.40 and 33.80% of the high-quality variants, the genotype of at least one out of 161 BSW samples was missing using the ARS-UCD1.2 and UOA_Angus_1 alignments, respectively. Across all chromosomes, the number of missing genotypes was slightly higher ($P = 0.087$) for variants called from UOA_Angus_1 than ARS-UCD1.2 alignments. The percentage of variants with missing genotypes was highest on chromosome 12 in both assemblies. At least one missing genotype was observed for 49.79 and 37.39% of the chromosome 12 variants for the UOA_Angus_1 and ARS-UCD1.2-called genotypes. Beagle [30] (version 4.1) phasing and imputation was applied to improve the genotype calls from GATK and impute the missing genotypes.

112 sequenced animals that had an average fold sequencing coverage of 13.47 $\pm$ 6.45 and 13.46 $\pm$ 6.44 when aligned to ARS-UCD1.2 and UOA_Angus_1, respectively, also

had Illumina BovineHD array-called genotypes at 530,372 autosomal SNPs. We considered the microarray-called genotypes as a truth set to calculate non-reference sensitivity, non-reference discrepancy and the concordance between array-called and sequence-called genotypes (Table 2.2). The average concordance between array- and sequence-called genotypes was greater than 98 and 99.5% before and after Beagle imputation, respectively, for variants called from both assemblies. We observed only slight differences in the concordance metrics between variants called from either ARS-UCD1.2 or UOA_Angus_1, indicating that the genotypes of the 112 BSW cattle were accurately called from both assemblies, and that Beagle phasing and imputation further increased the genotyping accuracy.

Table 2.2: **Comparisons between array-called and sequence variant genotypes**. Non-reference sensitivity (NRS), non-reference discrepancy (NRD) and the concordance (CONC) between array-called and sequence-called genotypes for 112 BSW cattle that had BovineHD and sequence-called genotypes at 530,372 autosomal SNPs.

	GATK hard filtering			GATK hard filtering + Beagle imputation		
	NRS	**NRD**	**CONC**	**NRS**	**NRD**	**CONC**
ARS-UCD1.2	99.14	2.75	98.13	99.77	0.60	99.59
UOA_Angus_1	99.37	2.45	98.09	99.88	0.47	99.64

Because Beagle phasing and imputation improved the genotype calls from GATK, the subsequent analyses are based on the imputed sequence variant genotypes. After imputation, 81,674 (0.36%, 72,121 SNPs, 9,553 INDELs) and 104,217 (0.46%, 75,342 SNPs, 28,875 INDELs) variants were fixed for the alternate allele in ARS-UCD1.2 and UOA_Angus_1, respectively (Additional file 3). Both the number and the percentage of variants fixed for the alternate allele was higher (0.10 percent points the latter, P = 0.027) for the UOA_Angus_1 than the ARS-UCD1.2 assembly. While the proportion and number of SNPs fixed for the alternate allele did not differ significantly (P = 0.65) between the assemblies, 0.61 percent points more INDELs (P = 1.45 x 10^{-9}) were fixed for the alternate allele in UOA_Angus_1 than ARS-UCD1.2. 22,488,261 and 22,289,905 variants were polymorphic (*i.e.*, minor allele count $\geq$ 1) among the 161 BSW animals in ARS-UCD1.2 and UOA_Angus_1, respectively (Table 2.3). The number of variants detected per sample ranged from 6.91 to 8.58 million (7.28 $\pm$ 0.15) in ARS-UCD1.2 and from 6.93 to 8.44 million (7.26 $\pm$ 0.15) in UOA_Angus_1. More SNPs and INDELs were discovered for the ARS-UCD1.2 than UOA_Angus_1 assembly.

Table 2.3: **Variants segregating among 161 BSW samples.** Number of high-quality non-fixed variants discovered after aligning the samples to ARS-UCD1.2 and UOA_Angus_1 assemblies. Numbers in parentheses reflect the variant density (number of variants per Kb) along the autosomes.

	ARS-UCD1.2	**UOA_Angus_1**
Non-fixed variants (per Kb)	22,488,261 (9.03)	22,289,905 (9.03)
Non-fixed SNPs (per Kb)	19,557,039 (7.86)	19,446,648 (7.88)
Non-fixed INDELs (per Kb)	2,931,222 (1.18)	2,843,257 (1.15)

To take the length of the autosomes into consideration, we calculated the number of variants per Kb. While the overall variant and INDEL density was slightly higher for the ARS-UCD1.2 assembly, the SNP density was slightly higher for the UOA_Angus_1 assembly (Table 2.3).

The number and density of high-quality variants segregating on the 29 autosomes was 2.04 (P = 0.51) and 0.45 (P = 0.39) percent points higher, respectively, for the ARS-UCD1.2 than the UOA_Angus_1 assembly (Fig. 2.1, Additional file 4). The difference in the number of variant sites detected from both assemblies was lower for SNPs (1.71 percent points) than INDELs (4.28 percent points). Chromosomes 9 and 12 were the only autosomes for which more variants were detected using the UOA_Angus_1 than ARS-UCD1.2 assembly. Differences in the number of variants detected were evident for chromosomes 12 and 28. While chromosome 12 has 29% more variants when aligned to UOA_Angus_1, chromosome 28 has 31% more variants when aligned to ARS-UCD1.2.

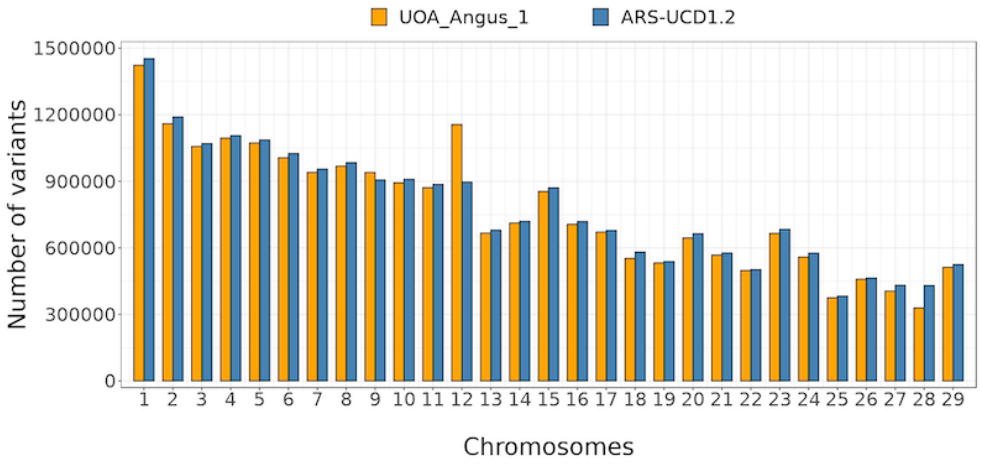


Figure 2.1: **Total number of variants of autosomes for both assemblies.** Number of variants detected on autosomes when the 161 BSW samples are aligned to the ARS-UCD1.2 (blue) and UOA_Angus_1 (orange) assembly.

The variant density of 26 out of the 29 autosomes (except for chromosomes 9, 12 and 26) was higher for the ARS-UCD1.2 assembly than the UOA_Angus_1 assembly. However, the density of INDELs was only higher for chromosome 12. Chromosome 23 had a higher variant density than all other chromosomes for both assemblies, with an average number of 13 variants detected per Kb. The high variant density at chromosome 23 primarily resulted from an excess of polymorphic sites within a ∼5 Mb segment (between 25 and 30 Mb in the ARS-UCD1.2 and between 22 and 27 Mb in UOA_Angus_1) encompassing the bovine major histocompatibility complex (BoLA) (Additional file 5). Other autosomes with density above 10 variants per Kb for both assemblies were chromosomes 12, 15 and 29. We observed the least variant density (∼8 variants per Kb) at chromosome 13. Chromosome 12 carries a segment with an excess of variants at ∼70 Mb in both assemblies. Visual inspection revealed that the segment with an excess of polymorphic sites was substantially larger in UOA_Angus_1 (7.6 Mb) than ARS-UCD1.2 (3.5 Mb) (Fig. 2.2). The variant-rich region at chromosome 12 coincides with a large segmental duplication that compromises reference-guided variant genotyping from short-read sequencing data and that has been described earlier [31, 32, 33]. Because of the greater number of variants and variant density in UOA_Angus_1, this extended region had a large impact on the cumulative genome-wide metrics presented in Table 2.3. When the same metrics were calculated without chromosome 12, the average density of both SNPs and INDELs was higher for ARS-UCD1.2 than UOA_Angus_1 (Additional file 6). Segments with an excess of polymorphic sites were also detected on the ARS-UCD1.2 chromosomes 4 (113-114 Mb), 5 (98-105 Mb), 10 (22-26 Mb), 18 (60-63 Mb), and 21 (20-21 Mb). The corresponding regions in the UOA_Angus_1 assembly showed the same excess of polymorphic sites. However, these regions were shorter, and their variant density was lower compared to the extended segment at chromosome 12. The strikingly higher number (+31%) of variants discovered at chromosome 28 for ARS-UCD1.2 than UOA_Angus_1 was due to an increased length of chromosome 28 in the ARS-UCD1.2 assembly (Fig. 2.2).

Of 22,488,261 and 22,289,905 high-quality non-fixed variants, 848,100 (3.78%) and 857,206 (3.83%) had more than two alleles in the ARS-UCD1.2 and UOA_Angus_1 alignments, respectively (Additional file 7). Most (69.75% for ARS-UCD1.2 and 69.09% for UOA_Angus_1) of the multi-allelic sites were INDELs. The difference in the percentage of multiallelic SNPs across assemblies was negligible. However, the difference in percentage of multiallelic INDELs was 0.69 percent points higher ($P = 2.55 \times 10^{-9}$) for UOA_Angus_1 than ARS-UCD1.2 autosomes.

In order to detect potential flaws in sequence variant genotyping, we investigated if the genotypes at the high-quality non-fixed variants agreed with Hardy-Weinberg pro-

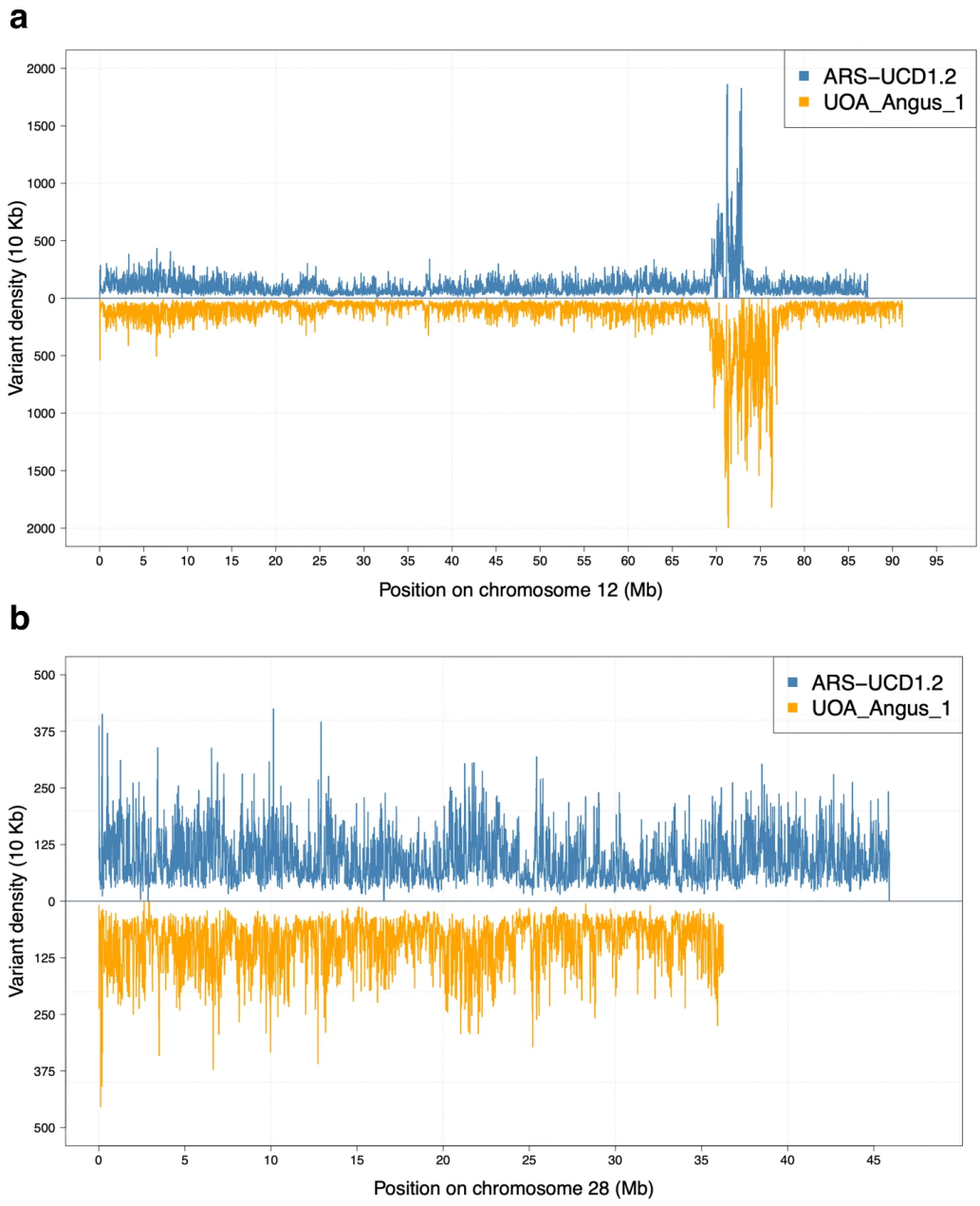


Figure 2.2: **Density of variants across chromosomes 12 and 28.** The number of variants within non-overlapping windows of 10 Kb for chromosome 12 (A) and 28 (B). The x-axis indicates the physical position along the chromosome (in Mb). The number of variants within each 10 Kb window is shown on the y-axis. Assembly ARS-UCD1.2 is displayed above the horizontal line (blue) and assembly UOA_Angus_1 is displayed below the horizontal line (orange).

portions. We observed 218,734 (0.97%) and 243,408 (1.09%) variants for ARS-UCD1.2 and UOA_Angus_1, respectively, for which the observed genotypes deviated significantly ($P < 10^{-8}$, Additional file 7) from expectations. The proportion of high-quality non-fixed variants for which the genotypes do not agree with Hardy-Weinberg proportions is 0.12 percent points higher for the UOA_Angus_1 than ARS-UCD1.2 assembly. At chromosome 12, 3.29 percent points more variants deviated from Hardy-Weinberg proportions for the UOA_Angus_1 than the ARS-UCD1.2 assembly (Additional file 8); more than twice the difference observed for any other autosome. When variants located on chromosome 12 were excluded from this comparison, we observed 199,304 (0.92%) and 180,264 (0.85%) variants for the ARS-UCD1.2 and UOA_Angus_1 assembly, respectively, for which the observed genotypes deviated significantly ($P < 10^{-8}$) from expectations.

Functional annotation of polymorphic sites

Using the VEP software, we predicted functional consequences based on the Ensembl genome annotation for 19,557,039 and 19,446,648 SNPs, and 2,931,222 and 2,843,257 INDELs, respectively, that were discovered from the ARS-UCD1.2 and UOA_Angus_1 alignments. Most SNPs were in either intergenic (66.30% and 56.56%) or intronic regions (32.55% and 42.09%) for ARS-UCD1.2 and UOA_Angus_1, respectively (Table 2.4, Additional file 9). Only 224,549 and 262,775 (1.15% and 1.35%) of the SNPs were in exons for ARS-UCD1.2 and UOA_Angus_1, respectively. The majority of INDELs was in either intergenic (65.76% and 55.95%) or intronic regions (33.84% and 43.47%) for ARS-UCD1.2 and UOA_Angus_1, respectively (Table 2.4, Additional file 9). Only 11,561 and 16,391 (0.40% and 0.58%) INDELs were in exonic sequences. While the number and proportion of variants in coding regions was similar for both assemblies, we observed marked differences in the number of variants annotated to intergenic and intronic regions. The percentage of SNPs and INDELs annotated to intergenic regions is 9.74 and 9.81 percent points higher, respectively, for the ARS-UCD1.2 than UOA_Angus_1 assembly. In contrast, the percentage of SNPs and INDELs annotated to intronic regions is 9.54 and 9.63 percent points higher, respectively, for the UOA_Angus_1 than the ARS-UCD1.2 assembly. According to the Ensembl annotation of the autosomal sequences, intergenic, intronic and exonic regions span respectively 61.53, 34.77 and 3.80% in ARS-UCD1.2 and 52.32, 42.32 and 5.36% in UOA_Angus_1.

Either moderate or high impacts on protein function were predicted for 89,812 and 103,576 SNPs, and 10,259 and 11,847 INDELs (0.46 and 0.53% of the total annotated SNPs and 0.35 and 0.41% of the total annotated INDELs), respectively, that were discovered from ARS-UCD1.2 and UOA_Angus_1 alignments (Tables 2.5 and 2.6). The number of

Table 2.4: **Number of SNPs and INDELs annotated using the VEP software per region and assembly.** Annotated SNPs and INDELs are classified by region were detected. The total number of annotated variants per assembly and region are displayed here. The table lists only the most severe annotation. The percentage of variants placed in each region per variant type and assembly is shown between parentheses.

	ARS-UCD1.2		**UOA_Angus_1**	
	SNPs	INDELs	SNPs	INDELs
Exonic regions (%)	224,549 (1.15)	11,561 (0.40)	262,775 (1.35)	16,391 (0.58)
Intronic regions (%)	6,365,765 (32.55)	992,015 (33.84)	8,185,503 (42.09)	1,236,006 (43.47)
Intergenic regions (%)	12,966,725 (66.30)	1,927,646 (65.76)	10,998,370 (56.56)	1,590860 (55.95)

variants with putatively high or moderate effects was higher for the UOA_Angus_1 than ARS-UCD1.2 assembly for 14 of 16 functional classes of annotations. Differences across all autosomes were observed for SNPs that potentially affect splice acceptor variants (345 for ARS-UCD1.2 and 395 for UOA_Angus_1, P = 0.032) and SNPs that potentially cause the loss of a stop codon (155 for ARS-UCD1.2 and 218 for UOA_Angus_1, P = 0.037). Differences across all autosomes also resulted for INDELs that potentially cause inframe deletions (1,761 for ARS-UCD1.2 and 1,972 for UOA_Angus_1, P = 0.0035), INDELs that potentially cause inframe insertions (850 for ARS-UCD1.2 and 985 for UOA_Angus_1, P = 0.0013) and INDELs that potentially cause the gain of a stop codon (218 for ARS-UCD1.2 and 288 for UOA_Angus_1, P = 0.016).

Table 2.5: **SNPs in high or moderate effect categories.** Number of SNPs in high and moderate (marked with an asterisk) effect categories per assembly.

	ARS-UCD1.2	**UOA_Angus_1**
Missense variant*	86,634	99,773
Stop gained	1,466	1,911
Splice donor variant	506	525
Splice acceptor variant	345	395
Start lost	271	319
Stop lost	155	218

Signatures of selection

Next, we investigated how the choice of the reference genome impacts the detection of putative signatures of selection in the 161 BSW cattle. We used the composite likelihood ratio (CLR) test to identify beneficial adaptive alleles that are either close to fixation

Table 2.6: **INDELs in high or moderate effect categories.** Number of INDELs in high and moderate (marked with an asterisk) effect categories per assembly.

	ARS-UCD1.2	**UOA_Angus_1**
Frameshift variant	6,289	7,435
Inframe deletion*	1,761	1,972
Inframe insertion*	850	985
Splice donor variant	291	298
Splice acceptor variant	292	292
Stop gained	218	288
Protein altering variant*	87	107
Start lost	20	14
Stop lost	11	15
Transcript ablation	5	6

or recently reached fixation [34]. As information on ancestral and derived alleles was not available, we considered 19,370,683 (ARS-UCD1.2) and 19,255,155 (UOA_Angus_1) sequence variants that were either polymorphic or fixed for the alternate allele in the 161 BSW cattle. The CLR test revealed 40 and 33 genomic regions (merged top 0.1% windows) encompassing ∼2.5 and ∼2.48 Mb, and 29 and 27 genes, respectively, from the ARS-UCD1.2 and the UOA_Angus_1 alignments (Fig. 2.3, Additional file 10, Additional file 11).

A putative signature of selection on chromosome 6 encompassing the *NCAPG* gene had high CLR values in both assemblies ($CLR_{ARS\text{-}UCD1.2}$ = 4064; $CLR_{UOA\_Angus\_1}$ = 3838). Another signature of selection was detected for both assemblies upstream of the *KITLG* gene on chromosome 5 (ARS-UCD1.2: 18.48 - 18.86 Mb, $CLR_{ARS\text{-}UCD1.2}$ = 655; UOA_Angus_1: 18.48 - 18.84, $CLR_{UOA\_Angus\_1}$ = 657). However, most of the signatures of selection were detected for only one assembly. A putative selective sweep on chromosome 13 was identified using the ARS-UCD1.2 but not the UOA_Angus_1 assembly. The putative selective sweep was between 11.5 and 12 Mb encompassing three protein coding (*CCDC3*, *CAMK1D* and ENSBTAG00000050894) and one non-coding gene (ENSBTAG00000045070). The top window (CLR=1373) was between 11,962,310 and 12,022,317 bp. In order to investigate why the CLR test revealed strong evidence for the presence of a signature of selection in ARS-UCD1.2 but not in UOA_Angus_1, we investigated the corresponding region in both assemblies using dot plots, variant density, alternate allele frequency and alignment coverage. The dot plot revealed that the orientation of bovine chromosome 13 is flipped in the UOA_Angus_1 assembly. The putative signature

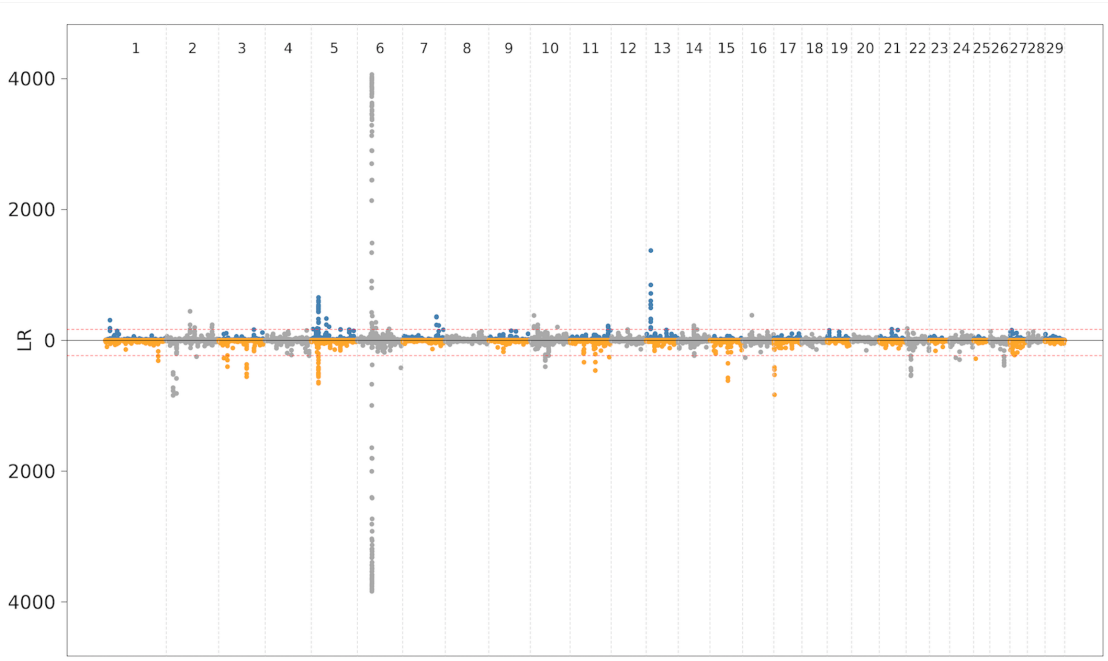


Figure 2.3: **Genome wide distribution of selection signals from CLR.** Selection signal distribution for both ARS-UCD1.2 (top panel) and UOA_Angus_1 assemblies (bottom panel). Red dotted line shows top 0.1% signal.

of selection is next to but clearly distinct from a region with a very high SNP density and sequence coverage in both assemblies (Additional file 12). We detected 350 SNP within the top window (5.87 SNP / Kb) of which 145 were fixed for the alternate allele. Within the corresponding region on UOA_Angus_1, we detected 209 SNP (3.48 SNP / Kb) of which 13 were fixed for the alternate allele. This pattern indicates that the 161 sequenced BSW cattle carry a segment in the homozygous state that is more similar to the UOA_Angus_1 than ARS-UCD1.2 reference genome. We observed the reciprocal pattern for a putative selective sweep on chromosome 22 that was detected using UOA_Angus_1 but not ARS-UCD1.2 (Additional file 13).

Genome-wide association testing

Next, we imputed genotypes for autosomal variants that were detected using the two assemblies for 30,499 cattle that had (partially imputed) Illumina BovineHD array-derived genotypes. The average imputation accuracy (Beagle R^2) was 0.87 $\pm$ 0.27 (median: 0.99) in the ARS-UCD1.2 and 0.87 $\pm$ 0.26 (median: 0.99) in the UOA_Angus_1 assembly. To prevent bias resulting from imputation errors, we removed variants that had low frequency (minor allele count < 3), low accuracy of imputation (Beagle $R^2 < 0.5$) or for which the observed genotypes deviated significantly ($P < 10^{-6}$) from Hardy-Weinberg proportions from the imputed data. Following quality control, 12,761,165 and

12,602,069 imputed variants were respectively retained (with imputation accuracy of 0.95 $\pm$ 0.11 and 0.95 $\pm$ 0.10) for genetic investigations in the ARS-UCD1.2 and UOA_Angus_1 dataset representing 56.75 and 56.54% of the 22,488,261 and 22,289,905 high-quality segregating variants. We then carried out genome-wide association studies (GWAS) between imputed sequence variant genotypes and six traits, including stature and five dairy traits (milk yield, fat yield, protein yield, protein and fat percentage), for which between 11,294 and 12,396 cattle had phenotypes in the form of de-regressed proofs. The resulting Manhattan plots appeared very similar for both datasets (Fig. 2.4, Additional file 14). Across the six traits analysed, the number of significantly associated variants was similar when the association analyses were performed using imputed sequence variants identified in the two builds. The difference in the number of significantly associated variants ($P < 10^{-8}$) between the two builds is mainly due to variants that had P-values that were slightly above the threshold of 10^{-8} in one but not the other build.

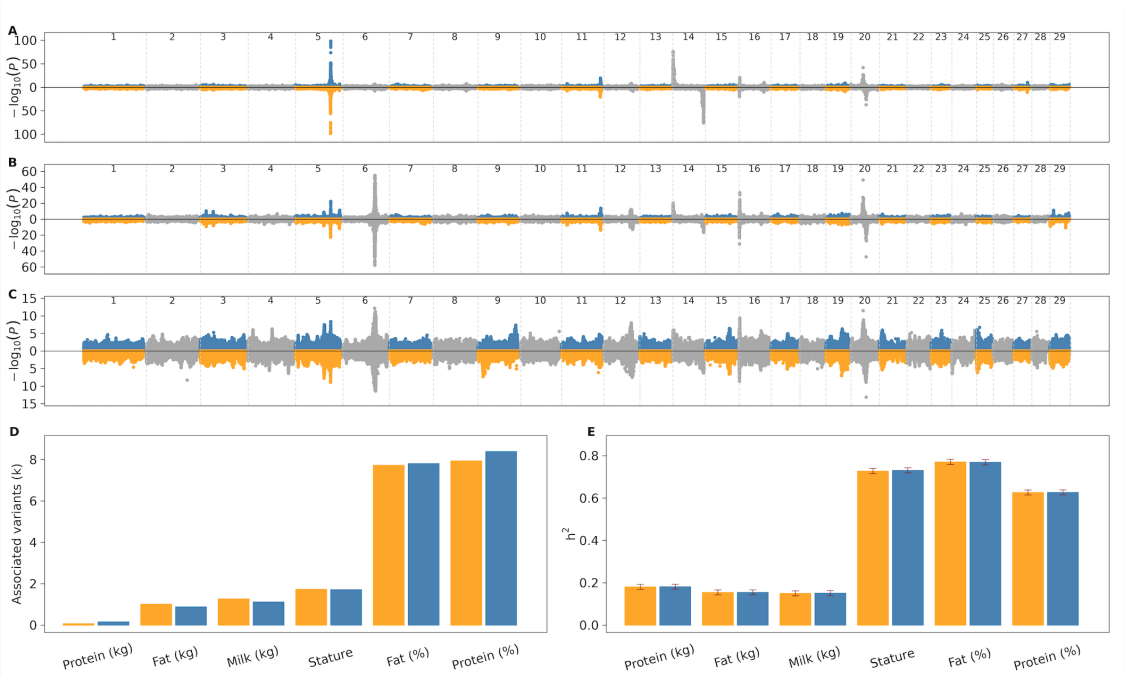


Figure 2.4: **Manhattan plots for fat percentage, protein percentage, and milk yield.** Number of significantly ($P < 10^{-8}$) associated variants in GWAS for seven traits. Estimated genomic heritability for stature and six dairy traits. Manhattan plots showing association of sequence variants - imputed using ARS-UCD1.2 (blue and grey) and UOA_Angus_1 (orange and grey) - with fat percentage (A), protein percentage (B) and milk yield (C). The orientation of some autosomes (*e.g.*, chromosome 14 & chromosome 20) is flipped between ARS-UCD1.2 and UOA_Angus_1. The number (in thousands) of variants - imputed using ARS-UCD1.2 (blue) and UOA_Angus_1 (orange) - significantly ($P < 10^{-8}$) associated with the seven traits considered for GWAS (D). Genomic heritability estimated using all autosomal variants imputed using ARS-UCD1.2 (blue) and UOA_Angus_1 assemblies (orange) (E). Standard errors of the estimates are indicated in read lines.

To investigate if causal variants can be readily identified from both assemblies, we inspected the QTL for dairy traits at chromosomes 14 and 20, respectively, for which p.Ala232Lys in *DGAT1* encoding Diacylglycerol O-Acyltransferase 1 and p.Phe279Tyr in *GHR* encoding Growth Hormone Receptor have been proposed as causal variants [35, 36]. The accuracy of imputation for the Phe279Tyr variant in the *GHR* gene was 0.92 and 0.88 for the ARS-UCD1.2 and UOA_Angus_1 assembly, respectively. In the association studies for milk yield, fat percentage and protein percentage, for which chromosome 20 QTL was detected, the p.Phe279Tyr variant was the most significantly associated variant in both assemblies. The SNP is located at 31,888,449 and 39,903,176 bp on the ARS-UCD1.2 and UOA_Angus_1 build (the orientation of chromosome 20 is flipped in UOA_Angus_1). The frequency of the milk yield-increasing and fat and protein content-decreasing tyrosine-encoding T allele was 12.90 and 13.02% in ARS-UCD1.2 and UOA_Angus_1, respectively, and the P-values for milk yield, fat percentage and protein percentage were 3.18×10^{-12}, 1.11×10^{-42}, 6.98×10^{-50} and 7.40×10^{-14}, 6.89×10^{-38}, 5.57×10^{-48}.

Two adjacent SNPs (ARS-UCD1.2: g.611019G>A & g.611020C>A; UOA_Angus_1: g.81672806C>T & g.81672805G>T; the orientation of chromosome 14 is flipped in UOA_Angus_1) in the coding sequence of *DGAT1* cause the p.Ala232Lys substitution that has a large effect on milk yield and composition. In 161 sequenced BSW cattle of our study, the alternate allele was detected in the heterozygous state in two and one animals using the ARS-UCD1.2 and UOA_Angus_1 datasets. When imputed into array-derived genotypes of the mapping cohort, the lysine variant had a frequency of 0.0082 (Beagle R^2: 0.98) and 0.0002 (Beagle R^2: 0.82) in the ARS-UCD1.2 and UOA_Angus_1 imputed genotypes. An association study between imputed sequence variant genotypes and fat percentage revealed strong association ($P = 1.46 \times 10^{-76}$) at the proximal region of chromosome 14 encompassing *DGAT1* in the ARS-UCD1.2 data (Fig. 2.4A). The top association signal resulted from a variant at position 420,486. The P-value of the p.Ala232Lys variant was only slightly higher ($P = 2.18 \times 10^{-76}$). Using the UOA_Angus_1 imputed data, we detected strong association at the corresponding region (Fig. 2.4A). The most significantly associated variant ($P = 1.80 \times 10^{-76}$) was at 81,673,955 bp. However, the p.Ala232Lys variant was not associated with fat percentage (P = 0.33). Also, the *DGAT1* gene was missing in the Ensembl annotation of the UOA_Angus_1 assembly.

Next, we estimated the genomic heritability (h^2) for stature and six dairy traits using a genomic restricted maximum likelihood estimation (GREML) approach. Therefore, we built a genomic relationship matrix separately for each assembly using the genotypes of all imputed autosomal variants that had minor allele count > 3 and imputation accuracy (Beagle R^2) > 0.5. The estimates for the genomic h^2 did not differ for all seven

traits (Fig. 2.4E). We then partitioned (genomic) h^2 by the 29 autosomes using the two imputed datasets. As seen for the total h^2, we found no difference in variance explained by individual autosomes between the two assemblies.

2.3 Discussion

We investigated whether the choice of the reference genome impacts genomic analyses in BSW cattle that have been sequenced with short paired-end reads. To the best of our knowledge, such an evaluation had not been performed so far in cattle. A Hereford-based genome assembly [20] is accepted by the bovine genomics community as reference genome for reference-guided alignment and variant detection in both taurine and indicine cattle [8, 9]. Recently, the application of sophisticated methods to assemble long sequencing reads provided reference-quality assemblies for cattle breeds other than Hereford [17, 21]. None of these novel reference-quality assemblies has been considered as a reference genome for sequence variant analysis so far. The genetic distance between the reference genome and the target sample and the properties (GC content, genome size, proportion of repeats) of the reference genome impact reference-guided mapping and variant genotyping [17, 24, 25, 37, 38]. To investigate reference-guided sequence analyses from different assemblies, we aligned short sequencing reads of 161 BSW cattle to the Hereford-based ARS-UCD1.2 and Angus-based UOA_Angus_1 assemblies. Widely used metrics (contig N50, scaffold N50, BUSCO completeness) suggest that both assemblies are of reference quality [17, 20]. The sequence read mapping and variant genotyping accuracy did not differ notably between the ARS-UCD1.2 and UOA_Angus_1 assemblies, indicating that both assemblies are suitable for reference-guided genome analyses in BSW cattle. The BSW, Angus and Hereford breeds are closely related as these breeds diverged relatively recently [39]. Greater genetic distance between the target breed and the reference genome might compromise mapping rate and alignment quality [24, 25, 40]. However, it is worth mentioning that the orientation of some chromosomes is flipped in UOA_Angus_1 (*i.e.*, the beginning of the chromosome corresponds to the end in the corresponding ARS-UCD1.2 entry). This does not affect sequence read mapping and variant genotyping but needs to be considered when comparing selection signatures and association signals across assemblies.

The number and density of INDELs that segregate in 161 BSW cattle was slightly lower when variants were called from the UOA_Angus_1 than ARS-UCD1.2 alignment. However, the proportion of multiallelic INDELs and INDELs fixed for the alternate allele was higher in the UOA_Angus_1 than ARS-UCD1.2 alignment. In fact, the absolute

number of INDELs fixed for the alternate allele was three times higher when the sequence data were aligned against the UOA_Angus_1 assembly. An excess of artefactual INDELs in long-read sequencing-based assemblies was noted by Watson and Warr [41]. Both the ARS-UCD1.2 and UOA_Angus_1 assembly were constructed from PacBio continuous long reads. While ARS-UCD1.2 was polished with short reads and manually curated, this step was not as extensively carried out for the UOA_Angus_1 assembly [17, 20]. Our results may indicate that UOA_Angus_1 contains somewhat more artefactual INDELs than ARS-UCD1.2. However, the absolute number of artefactual INDELs is low for both assemblies and their genotypes are likely to be discarded from most downstream analyses as most of them will be fixed for the alternate allele. Importantly, the concordance between sequence- and array-called genotypes was very high and the number of variants deviating from Hardy-Weinberg proportions was very low at segregating sites for both assemblies, indicating that reliable genotypes can be obtained from both ARS-UCD1.2 and UOA_Angus_1.

The length of chromosomes 12 and 28 differs considerably between the assemblies. A large segmental duplication affects chromosome 12 in both assemblies. This duplication compromises the mapping of sequencing reads, thereby causing misalignments and flaws in the resulting genotypes [31, 32, 33]. An excess of variants, including many for which the genotypes deviate from Hardy-Weinberg proportions, was detected for both assemblies within the segmental duplication. Because the segmental duplication is two times longer in UOA_Angus_1 than ARS-UCD1.2, the genome-wide number of variants, variant density, proportion of missing genotypes and number of variants deviating from Hardy-Weinberg proportions was higher using UOA_Angus_1. At chromosome 28, the variant density was similar for both assemblies, but the absolute number of variants detected was lower for UOA_Angus_1 because the chromosome was shorter. The UOA_Angus_1 assembly lacks approximately 9.5 million bases that likely correspond to the ARS-UCD1.2 chromosome 28 sequence from 36,496,661 bp onwards. According to the Ensembl (build 101) annotation of ARS-UCD1.2, this segment encompasses 67 genes that are consequently missing in the autosomal annotation of UOA_Angus_1.

Differences in the functional annotations predicted for variants obtained from ARS-UCD1.2 and UOA_Angus_1 were evident from the output of the VEP tool. The number of variants annotated to inter- and intragenic regions differed between the assemblies because the length of these features differed in the annotation files. The accuracy and quality of the annotation depend on whether a posterior manual validation of structures and functions is performed [42, 43]. An example for a striking difference in the coding sequence between both annotations is *DGAT1*, a gene that harbours a missense variant (p.Ala232Lys) with a large impact on dairy traits [36]. Our GWAS identified a QTL for

dairy traits at chromosome 14 in both assemblies. The QTL encompassed *DGAT1* using the ARS-UCD1.2 annotation. However, *DGAT1* was not annotated at the corresponding sequence of the UOA_Angus_1 assembly. Given the manual curation efforts of the ARS-UCD1.2 annotation in contrast to the mere computational-based inference of annotations for UOA_Angus_1 from the Ensembl database, we suspect that the latter produces more erroneous annotations [43]. In fact, the ARS-UCD1.2 assembly is currently the widely accepted and universally applied bovine reference genome [44, 45]. It is very unlikely that this will change soon because besides the completeness and continuity of the reference assembly, its functional annotation is crucial for downstream analyses. While tools exist to lift physical coordinates from one genomic context to another based on flanking sequences, this approach is cumbersome. Consequently, errors and gaps in the functional annotations of bovine reference-quality assemblies other than ARS-UCD1.2 are a major obstacle to switch references. The application of an augmented reference genome that contains ARS-UCD1.2 and its functional annotations as backbone as well as variants detected in other assemblies might solve such problems [12, 46].

We applied the composite likelihood ratio test to detect alleles that are either close to fixation or already reached fixation using genotypes obtained from both references. Supplying information about ancestral and derived alleles to the composite likelihood ratio test is required to determine which allele has been under selection and increases the statistical power to detect signatures of selection [34, 47]. Although we were unable to differentiate between ancestral and derived alleles, we identified strong signatures of selection from both assemblies at regions encompassing genes that were previously detected in different cattle breeds including BSW [48, 49, 50]. However, quantifying the overlap between the signatures of selection detected in our and previous studies is not readily possible. First, a resource like AnimalQTLdb [51] that would allow for a systematic assessment of signatures of selection across studies does not exist. Second, differences in marker density and parameter settings (*e.g.*, folded *versus* unfolded site frequency spectrum) may affect the mapping precision and preclude an immediate comparison between studies. Third, the use of different assemblies, as it was the case in our study, results in coordinates that need to be lifted from one to another assembly. By visually inspecting the genes encompassed by the signatures of selection and manually lifting coordinates from ARS-UCD1.2 to UOA_Angus_1, we were able to confirm that the signatures of selection at chromosome 6 encompassing the *NCAPG* and on chromosome 5 upstream *KITLG* were indeed identical between both assemblies and detected previously in BSW cattle [50, 52]. This finding suggests that plausible signatures of selection can be identified using folded site frequency spectrum. However, we also detected signatures of selection that did not overlap between both assemblies. For instance, a strong selective sweep on chromosome 13 was only detected using the ARS-

UCD1.2 assembly, while a putative sweep on chromosome 22 was only detected using the UOA_Angus_1 assembly. These differences were unexpected because the two assemblies were constructed from breeds that diverged relatively recent. In fact, Hereford and Angus are both taurine beef breeds that originate from Great Britain and phylogenetic analyses suggest that they are closely related [39]. The BSW cattle breed is also a taurine breed of European ancestry. When the BSW samples were aligned to the ARS-UCD1.2 assembly, the chromosome 13 region harbouring the signature of selection was depleted for variation, suggesting that the selected allele(s) already reached fixation. In fact, we observed many variants that were fixed for the alternate allele within the top windows at chromosome 13. These variants were absent when the sequencing data were aligned to UOA_Angus_1, because their alternate alleles in ARS-UCD1.2 correspond to reference alleles in UOA_Angus_1. Thus, our findings suggest that detecting selective sweeps that already reached fixation with the composite likelihood ratio test depends on the relationship between the study population and the reference genome if a folded site frequency spectrum is used. The CLR test would reveal the same regions from both assemblies if only segregating sites are considered for the analysis. However, restricting the analysis to segregating sites bears a risk of missing sweeps that already reached fixation.

To our knowledge, a quantitative assessment of differences arising from the use of different reference genomes had only been performed in humans at a single nucleotide variant (SNV) level [25, 38]. Recently, Low *et al.* [17] mapped 38 cattle samples from 7 breeds against the Brahman and Angus assemblies to detect larger structural variants that may be involved in the adaptability of indicine cattle to harsh environments. We considered 161 BSW cattle for a thorough characterization of reference-guided analyses from two assemblies. As such an evaluation may be regularly performed in the future for many species, we developed a workflow that can be adapted and reused for varies breeds, populations and species [53]. In fact, our evaluation is the first to compare sequence variant discovery from primary and haplotype-resolved assemblies. Therefore, our findings also show that haplotype-resolved reference-quality assemblies may readily serve as reference genomes for linear read mapping and variant genotyping.

2.4 Conclusions

Our results suggest that both the ARS-UCD1.2 and UOA_Angus_1 assembly are suitable for reference-guided genome analyses in BSW cattle. The choice of the reference may have a large impact on detecting signatures of selection that already reached fixa-

tion. Furthermore, curation of the reference genomes is required to improve the characterisation of functional elements. The workflow herein developed is a starting point for a comprehensive comparison of the impact of reference genomes on genomic analyses in various breeds, populations and species.

2.5 Methods

Data availability and code reproducibility

Short paired-end whole-genome sequencing reads of 161 BSW cattle were considered for our analyses. Accession numbers for all animals are available in Additional file 1.

In order to investigate the effect of different assemblies on downstream analyses, we considered the current bovine Hereford-based reference genome (ARS-UCD1-2) [20] and an Angus-based reference-quality assembly (UOA_Angus_1) [17] that was generated from a F1 Angus x Brahman cross. The assemblies were downloaded from the public repositories of the NCBI (GCA_002263795.2, GCA_003369685.2). The UOA_Angus_1 assembly does not contain the X chromosomal sequence because it represents the paternal haplotype of a male animal. The ARS-UCD1.2 assembly was created from a female cow, thus does not contain a Y chromosomal sequence. For the sake of completeness, we expanded the ARS-UCD1.2 assembly with the Y chromosomal sequence from Btau 5.0 and the UOA_Angus_1 assembly with the X chromosomal sequence from ARS-UCD1.2.

We compared the assemblies regarding mapping and variant calling, functional annotation, detection of signatures of selection, imputation and genome-wide association testing. Alignment, coverage, variant calling, imputation, annotation and analysis workflows were implemented as described below using Snakemake [54] (version 5.10.0). Python 3.7.4 has been used for running custom scripts as well as for submission and generation of Snakemake workflows.

Unless stated otherwise, the R (version 3.3.3) software environment and ggplot2 package (version 3.0.0) were used to create figures and perform statistical analyses. Paired t-test and Kruskal-Wallis rank sum test were applied to assess differences between assemblies for normal and not normal distributed values, respectively.

Alignment quality and depth of coverage

Quality assessment and control (removal of adapter sequences and reads and bases with low quality) of the raw sequencing data was carried out using the fastp software [55] (version 0.19.4) with default parameter settings. Reads were discarded when the phred-scaled quality was below 15 for more than 15% of the bases.

When necessary, the resulting FASTQ files were split into up to 13 read-group-specific FASTQ files to facilitate the read group aware processing of the data using gdc-fastq-splitter [56] (version 0.0.1). The filtered reads were subsequently aligned to both the ARS-UCD1.2 and UOA_Angus_1 assemblies (see above) using the MEM-algorithm of the Burrows-Wheeler Alignment (BWA) software [57, 58] (version 0.7.17) with option -M and –R to mark shorter split hits as secondary alignments and supply read group identifier and default values for all other parameters. Samblaster [59] (version 0.1.24) was used to mark duplicates in the SAM files, which were then converted into the binary format by using SAMtools [60] (version 1.6). Sambamba [61] (version 0.6.6) was used for coordinate-sorting (sort function) and to combine the read group-specific BAM files into sample-specific sorted BAM files. Duplicated reads and PCR duplicates of the merged and coordinate-sorted BAM files were marked using the MarkDuplicates module from Picard Tools [62] (version 2.18.17).

Uniquely mapped and properly paired reads that had mapping quality greater than 10 were obtained using SAMtools view -q 10 -F 1796. We considered a phred-scaled mapping quality threshold of 10 to retain only reads (referred to as high-quality reads) that qualify for variant genotyping according to best practice guidelines of the GATK [28, 29].

The mosdepth software [63] (version 0.2.2) was used to extract the number of reads that covered a genomic position in order to obtain the average coverage per sample and chromosome. We considered only high-quality reads (by excluding reads with mapping quality < 10 and SAM flag 1796).

Sequence variant genotyping and variant statistics

We used the BaseRecalibrator module of the Genome Analysis Toolkit (GATK - version 4.1.4.1) [64, 65] to adjust the base quality scores using 115,815,241 (ARS-UCD1.2) and 87,710,119 (UOA_Angus_1) unique positions from the Bovine dbSNP version 150, as known variants. To obtain the coordinates of known sites for the UOA_Angus_1 assembly, we used liftover coordinates obtained from the mapping of 120 bases flanking the known ARS-UCD1.2 positions to UOA_Angus_1 using the MEM-approach of

BWA (see above) with option –k 120 to consider only full-length matches. To discover and genotype variants from the recalibrated BAM files, we used the GATK according to the best practice guidelines [28, 29]. The GATK HaplotypeCaller module was run to produce gVCF (genomic Variant Call Format) files. The gVCF files were then consolidated using GenomicsDBImport and passed to the GenotypeGVCFs module to genotype polymorphic SNP and INDELs. We applied the VariantFiltration module for site-level filtration with the following recommended thresholds to retain high-quality SNP and INDELs: QualByDepth (QD) $>$ 2.0, Qual $>$ 30, Strand Odds Ratio (SOR) $<$ 3.0, FisherStrand (FS) $<$ 60.0, RMSMappingQuality (MQ) $>$ 40.0, MappingQualityRankSumTest (MQRankSum) $>$ 12.5, ReadPosRankSumTest (ReadPosRankSum) $>$ 8.0 for SNPs, and (QD) $>$ 2.0, Qual $>$ 30, Strand Odds Ratio (SOR) $<$ 10.0, FisherStrand (FS) $<$ 200.0, ReadPosRankSumTest (ReadPosRankSum) $>$ -20.0 for INDELs. Only variants with a genotyping rate of 50% or higher (this is, minimum of 161 alleles - AN) were considered. Variants not meeting all the criteria were discarded.

Beagle [30] (version 4.1) haplotype phasing and imputation was run to improve the raw genotype calls and impute missing genotypes. The genotype likelihood (gl) mode was applied in order to infer missing and adjust existing genotypes based on the phred-scaled likelihoods of all other non-missing genotypes.

Alternate allele frequency was calculated using the `–keep-allele-order –freq` flags with PLINK 1.9 [66] and non-segregating variants were subsequently filtered out from the imputed VCF file with the option `–mac 1 –remove-filtered-all` from VCFtools [67]. Biallelic variants have been retrieved by using the filter `–min-alleles 2 –max-alleles 2` with VCFtools. Index and stats for the relevant VCF files were generated through tabix [68], VCFtools and BCFtools [69], respectively. Per-sample stats were obtained by adding the `-v` flag when generating the stats with VCFtools. Observed genotypes were tested for deviation from Hardy-Weinberg proportions using the `–hwe 10e-8` and `–hardy –recode` flags with PLINK 1.9 [66]. Transition and transversion ratio of SNPs were calculated via VCFtools.

Functional annotation of polymorphic sites

Functional consequences of high-quality and non-fixed SNPs and INDELs were predicted according to the Ensembl (release 101) annotation of the bovine genome assembly ARS-UCD1.2 and UOA_Angus_1, respectively, using the Ensembl Variant Effect Predictor tool (VEP - version 91.3) [70] with default parameters and `–hgvs –symbol` nomenclature. The classification of variants according to sequence ontology terms and the prediction of putative impacts on protein function followed Ensembl guidelines. Basic

statistics of the annotation were calculated using AGAT [71] (version v0.5.1).

Signatures of selection

Signatures of recent selection were identified using the composite likelihood ratio (CLR) approach implemented in Sweepfinder2 [72]. We considered 19,370,683 (ARS-UCD1.2) and 19,255,155 (UOA_Angus_1) biallelic SNP (segregating sites and SNP that were fixed for the non-reference allele) to calculate the CLR in 20 Kb windows with pre-computed empirical alternate allele frequency. The top 0.1% windows were considered as putative selective sweeps. Adjacent top 0.1% windows were merged into regions. The gene content of the regions was determined according to the annotations from Ensembl (release 101) using BEDTools [73].

Dot plots

To identify sequence similarities and dissimilarities between the two assemblies, we inspected chromosome wise dot plots of pair-wise sequence alignments using LASTZ [74] (v1.04.03) with the options `–notransition –nogapped –step=20 –exact=50` using repeat-masked assemblies which we downloaded from Ensembl (release 101).

Imputation

Microarray-derived SNP genotypes were available for 30,499 BSW cattle typed on seven low-density (20k-150k) and one high-density chip (Illumina BovineHD; 777k). Coordinates of the SNP were originally determined according to the ARS-UCD1.2 build. To remap the SNP to the UOA_Angus_1 assembly, we used liftover coordinates obtained from the mapping of 120 bases flanking the BovineHD probes to the UOA_Angus_1 assembly using the MEM algorithm of BWA [57, 58] with option –k 120 to consider only full-length matches. Both the original and the remapped genotype data were imputed (separately) to the whole genome sequence level using a stepwise approach with reference panels aligned to the respective genome assemblies. First, genotypes for all animals typed at low density were imputed to higher density (N = 683,752 (ARS-UCD1.2) and 622,699 (UOA_Angus_1) SNP) using 1,166 reference animals with BovineHD-derived genotypes. In a second step, the partially imputed high-density genotypes were imputed to the sequence level using a reference panel of 161 sequenced animals. Both steps of imputation were carried out with Beagle 5.1 [75]. Variants with MAC > 3 (or) deviating significantly from Hardy-Weinberg proportions ($P < 10^{-6}$), (or) with imputation accuracy (Beagle R^2) less than 0.5 were filtered out. The imputed data with variants aligned to the ARS-UCD1.2 and UOA_Angus_1 assembly respectively, contained genotypes at 12,761,165 and 12,602,069 sequence variants.

Genome-wide association testing and estimation of genomic heritability

We tested the association between phenotypes in the form of de-regressed proofs for six traits and sequence variants in between 11,294 and 12,434 BSW cattle. We considered phenotypes for stature (N=11,294), milk yield (N=13,388), protein yield (N=12,392), fat yield (N=12,388), protein content (N=12,439), and fat content (N=12,434). The SNP-based association study was carried out using a linear mixed model implemented with the MLMA-approach of the GCTA software package [76]. The model included a genomic relationship matrix built from 560,777 autosomal SNPs that were typed on the BovineHD chip (positions mapped according to ARS-UCD1.2) and four principal components to account for relatedness and population stratification. The genomic heritability was estimated for the six traits using the genomic restricted maximum likelihood (GREML) approach implemented in GCTA [76]. Therefore, we used genomic relationship matrices (GRM) that were built from all imputed autosomal sequence variants. We also partitioned the genomic heritability onto individual autosomes using GRM built from variants of the respective autosomes.

References

[1] Hao Zhao, Zhifu Sun, Jing Wang, Haojie Huang, Jean Pierre Kocher, and Liguo Wang. CrossMap: A versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, 30(7):1006–1007, apr 2014. ISSN 14602059. doi: 10.1093/bioinformatics/btt730.

[2] Lawrence B. Schook, Jonathan E. Beever, Jane Rogers, Sean Humphray, Alan Archibald, Patrick Chardon, Denis Milan, Gary Rohrer, and Kellye Eversole. Swine Genome Sequencing Consortium (SGSC): A strategic roadmap for sequencing the pig genome. In *Comparative and Functional Genomics*, volume 6, pages 251–255, jun 2005. doi: 10.1002/cfg.479.

[3] The Bovine Genome Sequencing and Analysis Consortium, Christine G. Elsik, Ross L. Tellam, Kim C. Worley, Richard A. Gibbs, Donna M. Muzny, George M. Weinstock, David L. Adelson, Evan E. Eichler, Laura Elnitski, Roderic Guigó, et al. The genome sequence of taurine cattle: A window to ruminant biology and evolution. *Science*, 324(5926):522–528, apr 2009. ISSN 00368075. doi: 10.1126/science.1169588.

[4] LaDeana W International Chicken Genome Sequencing Consortium., Overall coordination:, Hillier. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, (432):695–716, 2004. doi: https://doi.org/10.1038/nature03154. URL www.nature.com/nature.

[5] Ross L. Tellam, Danielle G. Lemay, Curtis P. Van Tassell, Harris A. Lewin, Kim C. Worley, and Christine G. Elsik. Unlocking the bovine genome. *BMC Genomics*, 10:193, apr 2009. ISSN 14712164. doi: 10.1186/1471-2164-10-193.

[6] The Bovine HapMap Consortium, Evan E. Eichler, Roderic Guigó, Debora L. Hamernik, Steve M. Kappes, Harris A. Lewin, David J. Lynn, Frank W. Nicholas, Alexandre Reymond, Monique Rijnkels, Loren C. Skow, et al. Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science*, 324(5926):522–528, apr 2009. ISSN 00368075. doi: 10.1126/science.1169588.

[7] Sandra Jansen, Bernhard Aigner, Hubert Pausch, Michal Wysocki, Sebastian Eck, Anna Benet-Pagès, Elisabeth Graf, Thomas Wieland, Tim M. Strom, Thomas Meitinger, and Ruedi 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), jul 2013. ISSN 14712164. doi: 10.1186/1471-2164-14-446.

[8] Hans D. Daetwyler, Aurélien Capitan, Hubert Pausch, Paul Stothard, Rianne Van Binsbergen, Rasmus F. Brøndum, Xiaoping Liao, Anis Djari, Sabrina C. Rodriguez, Cécile Grohs, Diane Esquerré, Olivier Bouchez, Marie Noëlle Rossignol, Christophe Klopp, Dominique Rocha, Sébastien Fritz, André Eggen, Phil J. Bowman, David Coote, Amanda J. Chamberlain, Charlotte Anderson, Curt P. Vantassell, Ina Hulsegge, Mike E. Goddard, Bernt Guldbrandtsen, Mogens S. Lund, Roel F. Veerkamp, Didier A. Boichard, Ruedi Fries, and Ben J. Hayes. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genetics*, 46(8):858–865, 2014. ISSN 15461718. doi: 10.1038/ng.3034.

[9] L. Koufariotis, B. J. Hayes, M. Kelly, B. M. Burns, R. Lyons, P. Stothard, A. J. Chamberlain, and S. Moore. Sequencing the mosaic genome of Brahman cattle identifies historic and recent introgression including polled. *Scientific Reports*, 8(1), dec 2018. ISSN 20452322. doi: 10.1038/s41598-018-35698-5.

[10] Jacob Pritt, Nae Chyun Chen, and Ben Langmead. FORGe: Prioritizing variants for graph genomes. *Genome Biology*, 19(1):220, dec 2018. ISSN 1474760X. doi: 10.1186/s13059-018-1595-x.

[11] Danang Crysnanto, Christine Wurmser, and Hubert Pausch. Accurate sequence variant genotyping in cattle using variation-aware genome graphs. *Genetics Selection Evolution*, 51(1):21, may 2019. ISSN 12979686. doi: 10.1186/s12711-019-0462-x.

[12] Danang Crysnanto and Hubert Pausch. Bovine breed-specific augmented reference graphs facilitate accurate sequence read mapping and unbiased variant discovery. *Genome Biology*, 21(184), 2020. doi: https://doi.org/10.1186/s13059-020-02105-0. URL https://doi.org/10.1186/s13059-020-02105-0.

[13] Sara Ballouz, Alexander Dobin, and Jesse A. Gillis. Is it time to change the reference genome? *Genome Biology*, 20(1):1–9, 2019. ISSN 1474760X. doi: 10.1186/s13059-019-1774-4.

[14] John Eid. Real-Time DNA Sequencing fromSingle Polymerase Molecules. *Science*, 323(5910):130–133, jan 2009. ISSN 00368075. doi: 10.1126/science.1162986.

[15] Alexander S. Mikheyev and Mandy M.Y. Tin. A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources*, 14(6):1097–1102, nov 2014. ISSN 17550998. doi: 10.1111/1755-0998.12324.

[16] van Dijk Erwin L., Yan Jaszczyszyn, Delphine Naquin, and Claude Thermes. The Third Revolution in Sequencing Technology. *Trends Genet*, 34(9):666–681, sep 2018. ISSN 13624555. doi: 10.1016/j.tig.2018.05.008.

[17] Wai Yee Low, Rick Tearle, Ruijie Liu, Sergey Koren, Arang Rhie, Derek M. Bickhart, Benjamin D. Rosen, Zev N. Kronenberg, Sarah B. Kingan, Elizabeth Tseng, Françoise Thibaud-Nissen, Fergal J. Martin, Konstantinos Billis, Jay Ghurye, Alex R. Hastie, Joyce Lee, Andy W.C. Pang, Michael P. Heaton, Adam M. Phillippy, Stefan Hiendleder, Timothy P.L. Smith, and John L. Williams. Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. *Nature Communications*, 11(1), dec 2020. ISSN 20411723. doi: 10.1038/s41467-020-15848-y.

[18] Haoyu Cheng, Gregory T Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. Haplotype-resolved de novo assembly with phased assembly graphs. aug 2020. doi: http://arxiv.org/abs/2008.01237. URL http://arxiv.org/abs/2008.01237.

[19] Sergey Koren, Arang Rhie, Brian P. Walenz, Alexander T. Dilthey, Derek M. Bickhart, Sarah B. Kingan, Stefan Hiendleder, John L. Williams, Timothy P.L. Smith, and Adam M. Phillippy. De novo assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology*, 36(12):1174–1182, 2018. ISSN 15461696. doi: 10.1038/nbt.4277.

[20] Benjamin D. Rosen, Derek M. Bickhart, Robert D. Schnabel, Sergey Koren, Christine G. Elsik, Elizabeth Tseng, Troy N. Rowan, Wai Y. Low, Aleksey Zimin, Christine Couldrey, Richard Hall, Wenli Li, Arang Rhie, Jay Ghurye, Stephanie D. McKay, Françoise Thibaud-Nissen, Jinna Hoffman, Brenda M. Murdoch, Warren M. Snelling, Tara G. McDaneld, John A. Hammond, John C. Schwartz, Wilson Nandolo, Darren E. Hagen, Christian Dreischer, Sebastian J. Schultheiss, Steven G. Schroeder, Adam M. Phillippy, John B. Cole, Curtis P. Van Tassell, George Liu, Timothy P.L. Smith, and Juan F. Medrano. De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience*, 9(3), mar 2020. ISSN 2047217X. doi: 10.1093/gigascience/giaa021.

[21] Edward S. Rice, Sergey Koren, Arang Rhie, Michael P. Heaton, Theodore S. Kalbfleisch, Timothy Hardy, Peter H. Hackett, Derek M. Bickhart, Benjamin D. Rosen, Brian Vander Ley, Nicholas W. Maurer, Richard E. Green, Adam M. Phillippy, Jessica L. Petersen, and Timothy P.L. Smith. Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. *GigaScience*, 9(4):1–9, apr 2020. ISSN 2047217X. doi: 10.1093/gigascience/giaa029.

[22] Rachel M. Sherman, Juliet Forman, Valentin Antonescu, Daniela Puiu, Michelle Daya, Nicholas Rafaels, Meher Preethi Boorgula, Sameer Chavan, Candelaria Vergara, Victor E. Ortega, Albert M. Levin, Celeste Eng, Maria Yazdanbakhsh, James G. Wilson, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics*, 51(1):30–35, jan 2019. ISSN 15461718. doi: 10.1038/s41588-018-0273-y.

[23] Jacob F. Degner, John C. Marioni, Athma A. Pai, Joseph K. Pickrell, Everlyne Nkadori, Yoav Gilad, and Jonathan K. Pritchard. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–3212, oct 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp579.

[24] Torsten Günther and Carl Nettelblad. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genetics*, 15(7), jul 2019. ISSN 15537404. doi: 10.1371/journal.pgen.1008302.

[25] Yan Guo, Yulin Dai, Hui Yu, Shilin Zhao, David C. Samuels, and Yu Shyr. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, 109(2):83–90, mar 2017. ISSN 10898646. doi: 10.1016/j.ygeno.2017.01.005.

[26] Rasko Leinonen, Hideaki Sugawara, and Martin Shumway. The sequence read archive. *Nucleic Acids Research*, 39(SUPPL. 1), jan 2011. ISSN 03051048. doi: 10.1093/nar/gkq1019.

[27] Rasko Leinonen, Ruth Akhtar, Ewan Birney, Lawrence Bower, Ana Cerdeno-Tárraga, Ying Cheng, Iain Cleland, Nadeem Faruque, Neil Goodgame, Richard Gibson, Gemma Hoad, Mikyung Jang, Nima Pakseresht, Sheila Plaister, Rajesh Radhakrishnan, Kethi Reddy, Siamak Sobhany, Petra Ten Hoopen, Robert Vaughan, Vadim Zalunin, and Guy Cochrane. The European nucleotide archive. *Nucleic Acids Research*, 39(SUPPL. 1), jan 2011. ISSN 03051048. doi: 10.1093/nar/gkq967.

[28] Geraldine A. Van der Auwera, Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Kiran V. Garimella, David Altshuler, Stacey Gabriel, and Mark A. DePristo. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, (43):1110, 2013. ISSN 1934340X. doi: 10.1002/0471250953.bi1110s43.

[29] Broad_Institute. Germline short variant discovery (snps + indels), 2021. URL https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-.

[30] Brian L. Browning and Sharon R. Browning. Genotype Imputation with Millions of Reference Samples. *American Journal of Human Genetics*, 98(1):116–126, jan 2016. ISSN 15376605. doi: 10.1016/j.ajhg.2015.11.020.

[31] George E. Liu, Mario Ventura, Angelo Cellamare, Lin Chen, Ze Cheng, Bin Zhu, Congjun Li, Jiuzhou Song, and Evan E. Eichler. Analysis of recent segmental duplications in the bovine genome. *BMC Genomics*, 10, dec 2009. ISSN 14712164. doi: 10.1186/1471-2164-10-571.

[32] Derek M. Bickhart, Yali Hou, Steven G. Schroeder, Can Alkan, Maria Francesca Cardone, Lakshmi K. Matukumalli, Jiuzhou Song, Robert D. Schnabel, Mario Ventura, Jeremy F. Taylor, Jose Fernando Garcia, Curtis P. Van Tassell, Tad S. Sonstegard, Evan E. Eichler, and George E. Liu. Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Research*, 22(4):778–790, apr 2012. ISSN 10889051. doi: 10.1101/gr.133967.111.

[33] Hubert Pausch, Iona M. MacLeod, Ruedi Fries, Reiner Emmerling, Phil J. Bowman, Hans D. Daetwyler, and Michael E. Goddard. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. *Genetics Selection Evolution*, 49(1), feb 2017. ISSN 12979686. doi: 10.1186/s12711-017-0301-x.

[34] Rasmus Nielsen, Scott Williamson, Yuseob Kim, Melissa J. Hubisz, Andrew G. Clark, and Carlos Bustamante. Genomic scans for selective sweeps using SNP data. *Genome Research*, 15(11):1566–1575, nov 2005. ISSN 10889051. doi: 10.1101/gr.4252305.

[35] Sarah Blott, Jong-Joo Kim, Sirja Moisio, Anne Schmidt-Küntzel, Anne Cornet, Paulette Berzi, Nadine Cambisano, Christine Ford, Bernard Grisart, Dave Johnson, Latifa Karim, Patricia Simon, Russell Snell, Richard Spelman, Jerry Wong, Johanna Vilkki, Michel Georges, Frédéric Farnir, Wouter Coppieters, and Vialactia Biosciences. Molecular Dissection of a Quantitative Trait Locus: A Phenylalanine-to-Tyrosine Substitution in the Transmembrane Domain of the Bovine Growth Hormone Receptor Is Associated With a Major Effect on Milk Yield and Composition. *Genetics*, 163(1):253–66, jan 2003. ISSN 0016-6731.

[36] Bernard Grisart, Wouter Coppieters, Frédéric Farnir, Latifa Karim, Christine Ford, Paulette Berzi, Nadine Cambisano, Myriam Mni, Suzanne Reid, Patricia Simon, Richard Spelman, Michel Georges, and Russell Snell. Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Research*, 12 (2):222–231, 2002. ISSN 10889051. doi: 10.1101/gr.224202.

[37] Kathryn C. Asalone, Kara M. Ryan, Maryam Yamadi, Annastelle L. Cohen, William G. Farmer, Deborah J. George, Claudia Joppert, Kaitlyn Kim, Madeeha Froze Mughal, Rana Said, Metin Toksoz-Exley, Evgeny Bisk, and John R. Bracht. Regional sequence expansion or collapse in heterozygous genome assemblies. *PLoS Computational Biology*, 16(7), jul 2020. ISSN 15537358. doi: 10.1371/journal.pcbi.1008104.

[38] Bohu Pan, Rebecca Kusko, Wenming Xiao, Yuanting Zheng, Zhichao Liu, Chunlin Xiao, Sugunadevi Sakkiah, Wenjing Guo, Ping Gong, Chaoyang Zhang, Weigong Ge, Leming Shi, Weida Tong, and Huixiao Hong. Similarities and differences between variants called with human reference genome HG19 or HG38. *BMC Bioinformatics*, 20, mar 2019. ISSN 14712105. doi: 10.1186/s12859-019-2620-0.

[39] Jared E. Decker, Stephanie D. McKay, Megan M. Rolf, Jae Woo Kim, Antonio Molina Alcalá, Tad S. Sonstegard, Olivier Hanotte, Anders Götherström, Christopher M. Seabury, Lisa Praharani, Masroor Ellahi Babar, Luciana Correia de Almeida Regitano, Mehmet Ali Yildiz, Michael P. Heaton, Wan Sheng Liu, Chu Zhao Lei, James M. Reecy, Muhammad Saif-Ur-Rehman, Robert D. Schnabel, and Jeremy F. Taylor. Worldwide Patterns of Ancestry, Divergence, and Admixture in Domesticated Cattle. *PLoS Genetics*, 10(3), 2014. ISSN 15537404. doi: 10.1371/journal.pgen.1004254.

[40] Justin Bohling. Evaluating the effect of reference genome divergence on the analysis of empirical RADseq datasets. *Ecology and Evolution*, 10(14):7585–7601, jul 2020. ISSN 20457758. doi: 10.1002/ece3 .6483.

[41] Mick Watson and Amanda Warr. Errors in long-read assemblies can critically affect protein prediction. *Nat Biotechnol*, 37(2):124–126, feb 2019. ISSN 15461696. doi: 10.1038/s41587-018-0004-z.

[42] Sajeet Haridas, Asaf Salamov, and Igor V. Grigoriev. Fungal genome annotation. In *Methods in Molecular Biology*, volume 1775, pages 171–184. Humana Press Inc., School of Life and Medical Sciences University of Hertfordshire Hatfield, Hertfordshire, AL10 9AB, UK, 2018. doi: 10.1007/978-1-4939-7 804-5_15.

[43] Erin McDonnell, Kimchi Strasser, and Adrian Tsang. Manual gene curation and functional annotation. In *Methods in Molecular Biology*, volume 1775, pages 185–208. Humana Press Inc., School of Life and Medical Sciences University of Hertfordshire Hatfield, Hertfordshire, AL10 9AB, UK, 2018. doi: 10.1 007/978-1-4939-7804-5_16.

[44] Leif Andersson, Alan L. Archibald, Cynthia D. Bottema, Rudiger Brauning, Shane C. Burgess, Dave W. Burt, Eduardo Casas, Hans H. Cheng, Laura Clarke, Christine Couldrey, Brian P. Dalrymple, Christine G. Elsik, et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biology*, 16(1):57, mar 2015. ISSN 1474760X. doi: 10.1186/s13059-015-0622-4.

[45] Emily L. Clark, Alan L. Archibald, Hans D. Daetwyler, Martien A.M. Groenen, Peter W. Harrison, Ross D. Houston, Christa Kühn, Sigbjørn Lien, Daniel J. Macqueen, James M. Reecy, Diego Robledo, Mick Watson, Christopher K. Tuggle, and Elisabetta Giuffra. From FAANG to fork: application of highly annotated genomes to improve farmed animal production. *Genome Biology*, 21(1):285, dec 2020. ISSN 1474760X. doi: 10.1186/s13059-020-02197-8.

[46] Danang Crysnanto, Alexander S Leonard, Zih-Hua Fang, and Hubert Pausch. Novel functional sequences uncovered through a bovine multi-assembly graph. *BioRxiv*, jan 2021. doi: 10.1101/2021.01.08.425845. URL https://doi.org/10.1101/2021.01.08.425845.

[47] Christian D. Huber, Michael DeGiorgio, Ines Hellmann, and Rasmus Nielsen. Detecting recent selective sweeps while controlling for mutation rate and background selection. *Molecular Ecology*, 25(1): 142–156, jan 2016. ISSN 1365294X. doi: 10.1111/mec.13351.

[48] Sophie Rothammer, Doris Seichter, Martin Förster, and Ivica Medugorac. A genome-wide scan for signatures of differential artificial selection in ten cattle breeds. *BMC Genomics*, 14(908):1, 2013. doi: https://doi.org/10.1186/1471-2164-14-908. URL http://www.biomedcentral.com/1471-2164/14/908.

[49] Lingyang Xu, Derek M. Bickhart, John B. Cole, Steven G. Schroeder, Jiuzhou Song, Curtis P. Van Tassell, Tad S. Sonstegard, and George E. Liu. Genomic signatures reveal new evidences for selection of important traits in domestic cattle. *Molecular Biology and Evolution*, 32(3):711–725, 2015. ISSN 15371719. doi: 10.1093/molbev/msu333.

[50] Meenu Bhati, Naveen Kumar Kadri, Danang Crysnanto, and Hubert Pausch. Assessing genomic diversity and signatures of selection in Original Braunvieh cattle using whole-genome sequencing data. *BMC Genomics*, 21(1), jan 2020. ISSN 14712164. doi: 10.1186/s12864-020-6446-y.

[51] Z. Hu, C. Park, and J. Reecy. Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. *Nucleic Acids Research*, 47(1), 2019. ISSN 03051048. doi: 10.1093/nar/gky1084.

[52] Heidi Signer-Hasler, Alexander Burren, Markus Neuditschko, Mirjam Frischknecht, Dorian Garrick, Christian Stricker, Birgit Gredler, Beat Bapst, and Christine Flury. Population structure and genomic inbreeding in nine Swiss dairy cattle populations. *Genetics Selection Evolution*, 49(1), nov 2017. ISSN 12979686. doi: 10.1186/s12711-017-0358-6.

[53] ETH_Animal_Genomics. Github repository: Reference assembly choice, 2021. URL https://github.com/AnimalGenomicsETH/Reference_assembly_choice.

[54] Johannes Köster and Sven Rahmann. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012. ISSN 14602059. doi: 10.1093/bioinformatics/bts480.

[55] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890, sep 2018. ISSN 14602059. doi: https://doi.org/10.1093/bioinformatics/bty560.

[56] Kyle Hernandez. Cli for splitting a fastq that has multiple readgroups, 2020. URL https://github.com/kmhernan/gdc-fastq-splitter.

[57] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, jul 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp324.

[58] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. mar 2013. doi: http://arxiv.org/abs/1303.3997. URL http://arxiv.org/abs/1303.3997.

[59] Gregory G. Faust and Ira M. Hall. SAMBLASTER: Fast duplicate marking and structural variant read extraction. *Bioinformatics*, 30(17):2503–2505, sep 2014. ISSN 14602059. doi: https://doi.org/10.1093/bioinformatics/btu314.

[60] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, aug 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp352.

[61] Artem Tarasov, Albert J Vilella, Edwin Cuppen, Isaac J Nijman, and Pjotr Prins. Sambamba: fast processing of NGS alignment formats. 31(12):2032–4, 2015. doi: 10.1093/bioinformatics/btv098. URL 10.1093/bioinformatics/btv098.

[62] Broad_Institute. Picard tools, 2021. URL http://broadinstitute.github.io/picard/.

[63] Brent S. Pedersen and Aaron R. Quinlan. Mosdepth: Quick coverage calculation for genomes and exomes. *Bioinformatics*, 34(5):867–868, mar 2018. ISSN 14602059. doi: 10.1093/bioinformatics/btx699.

[64] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, sep 2010. ISSN 10889051. doi: 10.1101/gr.107524.110.

[65] Mark A. Depristo, Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, Guillermo Del Angel, Manuel A. Rivas, Matt Hanna, Aaron McKenna, Tim J. Fennell, Andrew M. Kernytsky, Andrey Y. Sivachenko, Kristian Cibulskis, Stacey B. Gabriel, David Altshuler, and Mark J. Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–501, may 2011. ISSN 10614036. doi: 10.1038/ng.806.

[66] Christopher C. Chang, Carson C. Chow, Laurent C.A.M. Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), feb 2015. ISSN 2047217X. doi: 10.1186/s13742-015-0047-8.

[67] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, aug 2011. ISSN 13674803. doi: 10.1093/bioinformatics/btr330.

[68] Heng Li. Tabix: Fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, 27 (5):718–719, mar 2011. ISSN 13674803. doi: 10.1093/bioinformatics/btq671.

[69] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, nov 2011. ISSN 13674803. doi: 10.1093/bioinformatics/btr509.

[70] William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R.S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The Ensembl Variant Effect Predictor. *Genome Biology*, 17 (1), jun 2016. ISSN 1474760X. doi: 10.1186/s13059-016-0974-4.

[71] J. Dainat. Agat: Another gff analysis toolkit to handle annotations in any gtf/gff format., 2021. URL https://www.doi.org/10.5281/zenodo.3552717. Version v0.5.1.

[72] Michael Degiorgio, Christian D. Huber, Melissa J. Hubisz, Ines Hellmann, and Rasmus Nielsen. SweepFinder2: Increased sensitivity, robustness and flexibility. *Bioinformatics*, 32(12):1895–1897, jun 2016. ISSN 14602059. doi: 10.1093/bioinformatics/btw051.

[73] Aaron R. Quinlan and Ira M. Hall. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, jan 2010. ISSN 13674803. doi: 10.1093/bioinformatics/btq033.

[74] Robert S Harris. *Improved Pairwise Alignment of Genomic DNA*. PhD thesis, Pennsylvania State University, USA, 2007. URL https://dl.acm.org/doi/book/10.5555/1414852.

[75] Brian L. Browning, Ying Zhou, and Sharon R. Browning. A One-Penny Imputed Genome from Next-Generation Reference Panels. *American Journal of Human Genetics*, 103(3):338–348, sep 2018. ISSN 15376605. doi: 10.1016/j.ajhg.2018.07.015.

[76] Jian Yang, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, 88(1):76–82, jan 2011. ISSN 00029297. doi: 10.1016/j.ajhg.2010.11.011.

Chapter 3

The size and composition of haplotype reference panels impact the accuracy of imputation from low-pass sequencing in cattle

Audald Lloret-Villas[1], Hubert Pausch[1] and Alexander S. Leonard[1]

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

Published in *Genomics Selection Evolution (2023) 55:33*

Contribution: I participated in conceiving the study, designing the experiments, analysing the results and writing the manuscript. I also developed and compiled reproducible pipelines.

Abstract

Background: Low-pass sequencing followed by sequence variant genotype imputation is an alternative to the routine microarray-based genotyping in cattle. However, the impact of haplotype reference panels and their interplay with the coverage of low-pass whole-genome sequencing data have not been sufficiently explored in typical livestock settings where only a small number of reference samples is available.

Methods: Sequence variant genotyping accuracy was compared between two variant callers, GATK and DeepVariant, in 50 Brown Swiss cattle with sequencing coverages ranging from 4 to 63-fold. Haplotype reference panels of varying sizes and composition were built with DeepVariant based on 501 individuals from nine breeds. High-coverage sequence data for 24 Brown Swiss cattle were downsampled to between 0.01- and 4-fold to mimic low-pass sequencing. GLIMPSE was used to infer sequence variant genotypes from the low-pass sequencing data using different haplotype reference panels. The accuracy of the sequence variant genotypes that were inferred from low-pass sequencing data was compared with sequence variant genotypes called from high-coverage data.

Results: DeepVariant was used to establish bovine haplotype reference panels because it outperformed GATK in all evaluations. Within-breed haplotype reference panels were more accurate and efficient to impute sequence variant genotypes from low-pass sequencing than equally-sized multibreed haplotype reference panels for all target sample coverages and allele frequencies. F1 scores greater than 0.9, which indicate high harmonic means of recall and precision of called genotypes, were achieved with 0.25-fold sequencing coverage when large breed-specific haplotype reference panels (n = 150) were used. In the absence of such large within-breed haplotype panels, variant genotyping accuracy from low-pass sequencing could be increased either by adding non-related samples to the haplotype reference panel or by increasing the coverage of the low-pass sequencing data. Sequence variant genotyping from low-pass sequencing was substantially less accurate when the reference panel lacked individuals from the target breed.

Conclusions: Variant genotyping is more accurate with DeepVariant than GATK. DeepVariant is therefore suitable to establish bovine haplotype reference panels. Medium-sized breed-specific haplotype reference panels and large multibreed haplotype reference panels enable accurate imputation of low-pass sequencing data in a typical cattle breed.

3.1 Background

More than one million cattle are genotyped every year using the microarray technology for the purpose of genomic prediction [1]. Access to whole-genome sequence variants can improve the accuracy of genomic predictions and facilitates the monitoring of trait associated alleles [2]. However, costs are still too high to sequence all individuals from a population to a sufficient coverage for calling variants.

Low-coverage whole-genome sequencing (lcWGS) followed by genotype imputation has emerged as an alternative with comparable costs to genotyping microarrays but with substantially higher marker density (tens of millions *versus* tens of thousands) to obtain genotypes for a target population [3, 4, 5, 6]. Sequencing coverage as low as 0.1-fold can be used to infer sequence variant genotypes that are as accurate as those obtained from genotyping microarrays, especially for rare variants, while sequencing coverage greater than 1-fold can have much higher accuracy [5]. For many imputation methods, reference panels that are representative for the target populations are a prerequisite for the accurate imputation of genotypes from lcWGS [7, 8, 9]. The 1000 Genomes Project (1KGP) and the Haplotype Reference Consortium (HRC) established such reference panels for several human ancestry populations [10, 11] and made them available through dedicated imputation servers [12]. A bovine imputation reference panel established by the 1000 Bull Genomes project is frequently used to infer sequence variant genotypes for large cohorts of genotyped taurine cattle, thus enabling powerful genome-wide analyses at the nucleotide level [13]. Sequenced reference panels are available for other animal species [14, 15], but they lack diversity as they were established mainly using data from mainstream breeds and thus are depleted for individuals from local or rare populations.

An exhaustive set of variants and accurate genotypes are crucial to compile informative haplotype reference panels. The Genome Analysis Toolkit (GATK) has been frequently applied to discover and genotype sequence variants in large reference populations of many livestock species [3, 14]. DeepVariant has recently emerged as an alternative machine learning-based variant caller [16]. Several studies suggest that DeepVariant has superior genotyping accuracy compared to GATK [17, 18, 19, 20]. However, DeepVariant has rarely been applied to call variants in species other than humans [21, 22].

In this study, we benchmark sequence variant genotyping of DeepVariant and GATK in a livestock population. Then, we build haplotype reference panels of varying sizes and composition with DeepVariant, and use GLIMPSE to impute sequence variant genotypes for cattle that had been sequenced at between 0.01- and 4-fold. We show that

within-breed haplotype reference panels outperform multibreed reference panels across all tested scenarios, provided that a sufficient number of sequenced samples is available.

3.2 Results

Variant calling with GATK and DeepVariant

We compared sequence variant calling between GATK and DeepVariant for 50 Brown Swiss (BSW) cattle for which the sequencing depth ranged from 4 to 63-fold (19.26 $\pm$ 11.09) along the autosomes. GATK and DeepVariant identified 18,654,649 and 18,748,114 variants, respectively, of which 7.79% and 8.38% were filtered out because of their low quality (Table 3.1). In total, 16,147,567 filtered variants were identified by both callers, but 1,053,716 and 1,292,671 variants were private to GATK and DeepVariant, respectively (Fig. 3.1a). Overall, DeepVariant had more private SNPs than GATK, but GATK had more private INDELs than DeepVariant (Additional file 15). 416,642 variants had the same coordinates but different alternative alleles. These discrepant sites were primarily INDELs (83%, as opposed to the 12% of INDELs in all shared variants). Multiallelic sites accounted for 3.44% and 3.31% of the variants (0.33% and 0.28% of the SNPs, and 23.22% and 23.94% of the INDELs) that passed the quality filters of GATK and DeepVariant, respectively. Multiallelic sites were enriched among the variants private to either GATK or DeepVariant (Additional file 16).

The biallelic variants called by GATK had a higher percentage of homozygous reference (HOMREF) and heterozygous (HET) genotypes whereas the biallelic variants called by DeepVariant had a higher percentage of homozygous alternative (HOMALT) genotypes (Fig. 3.1b and Additional file 17a). Missing genotypes were very rare ($<$0.01%) for GATK-called biallelic variants but accounted for 2.72% of the DeepVariant-called genotypes (Additional file 17b). Beagle phasing and imputation increased the number of HET genotypes for both GATK (mostly transitioning from HOMREF) and DeepVariant (mainly due to the refinement of missing genotypes) (Additional file 17c).

Table 3.1: **Summary of the variants called by GATK and DeepVariant (DV)**. Multiallelic sites are presented in parentheses. Ti:Tv ratios are restricted to biallelic SNPs. Functional consequences are predicted for biallelic SNPs / biallelic INDELs.

Variant caller	Sets	Variants	SNPs	INDELs	Ti:Tv ratio	High impact predicted SNPs / INDELs
GATK	Raw	18,654,649 (831,391)	16,135,130 (58,049)	2,617,546 (773,342)	2.16	2,680 / 4,493
GATK	Filtered-out	1,453,366 (239,008)	1,271,522 (8,577)	279,871 (230,431)	1.66	428 / 500
GATK	Filtered	17,201,283 (592,383)	14,863,608 (49,472)	2,337,675 (542,911)	2.20	2,252 / 3,993
DV	Raw	18,748,114 (702,173)	16,554,438 (54,438)	2,401,933 (647,735)	2.24	3,530 / 2,778
DV	Filtered-out	1,571,454 (270,963)	1,174,815 (11,834)	393,927 (259,108)	2.19	1,061 / 612
DV	Filtered	17,440,238 (577,997)	15,361,785 (42,899)	2,240,627 (535,098)	2.24	2,474 / 2,240

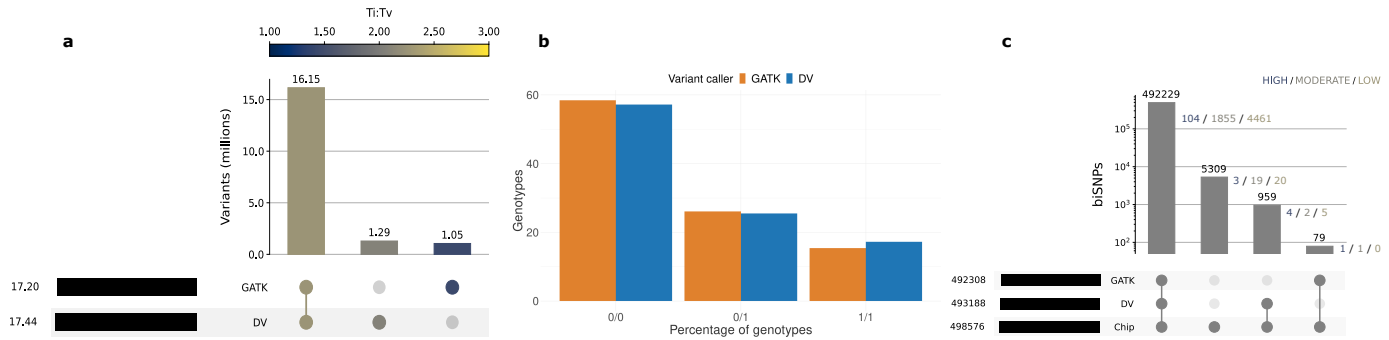


Figure 3.1: **Comparison of the variants called between DeepVariant (DV) and GATK.** a) Intersection of variants called with each variant caller (or both) and the Ti:Tv ratio of the biallelic SNPs of each set. b) Percentage of imputed genotypes called by each variant caller. c) Intersection of variant calls with truth genotyping arrays, where only variants at intersecting positions are retained. Variants with a low, moderate and high predicted impact from the intersecting sets are indicated.

Functional consequences on the protein sequence were predicted for all biallelic variants. DeepVariant identified 9% more SNPs that were predicted to have a high impact on protein function than GATK (Table 3.1 and Additional file 18). Around one fourth of the high impact SNPs detected by DeepVariant (24%) were not detected by GATK. GATK identified 78% more INDELs that were predicted to have a high impact on protein function than DeepVariant. More than half of the high impact INDELs detected by GATK (52%) were not detected by DeepVariant.

We investigated the ratio of transitions to transversions (Ti:Tv) to assess variant quality. Deviations from an expected genome-wide Ti:Tv ratio of $\sim$ 2.0-2.2 indicate random genotyping errors or sequencing artifacts [17, 20, 23, 24]. The Ti:Tv ratio was 2.16 and 2.24 for raw SNPs identified by GATK and DeepVariant, respectively (Table 3.1). While the Ti:Tv ratio was higher (2.20) for the GATK variants that met the quality filters, variant filtration had no impact on the Ti:Tv ratio for SNPs called by DeepVariant. The Ti:Tv ratio of the filtered-out SNPs was substantially lower for GATK (1.66) than for DeepVariant (2.19). SNPs private to GATK had lower Ti:Tv ratios than those private to DeepVariant (Fig. 3.1a). Substantial differences in the Ti:Tv ratio (0.81 points) were observed between overlapping and GATK-private SNPs but were smaller (0.18 points) between overlapping and DeepVariant-private SNPs.

Accuracy of variant calling

Thirty-three sequenced cattle also had between 17,575 and 490,174 SNPs genotyped with microarrays. The filtered biallelic SNPs called with GATK and DeepVariant (query sets) were compared to those genotyped with the microarrays (truth chip set). The vast majority (98.82%) of the SNPs present in the truth chip set was called by both tools (Fig. 3.1c). The number of overlapping SNPs present in the truth chip set was slightly larger for DeepVariant than for GATK. 1.06% (n = 5309) of the SNPs present in the truth chip set were not called by any of the software as biallelic SNPs. However, 3497 of these SNPs were present at the same position but had different alternative alleles (*e.g.*, multiallelic SNPs or INDELs) in DeepVariant *versus* GATK while the other 1812 positions were truly missing. Most of the biallelic SNPs private to the chip set (5265) were also missing in the raw calls from the variant callers. DeepVariant filtered out more variants present in the truth chip set than GATK.

The analysis of variant effect predictions for the filtered variants revealed that most low/moderate/high impact variants were called by both GATK and DeepVariant (99.4%, 98.8%, and 92.8%, respectively). However, DeepVariant additionally called 5/2/4 biallelic SNPs predicted as low/moderate/high impact respectively, while GATK only called

0/1/1 (Fig. 3.1c). Some of the low/moderate/high impact biallelic SNPs private to GATK (1 out of the 2) and DeepVariant (5 out of the 11) were called either as multiallelic SNPs or as INDELs by the other caller (Additional file 19). Only half (1 out of 2) of GATK's private variants have a MAF higher than 0.5, while most (9 out of 11) of the DeepVariant's private variants do, which suggests that GATK misses more variants that might have a larger impact in populations.

Genotyping accuracy of variant calls

GATK and DeepVariant called 492,265 and 493,145 variants from the truth chip set, respectively. GATK missed (8.13%) and miscalled (10.13%) more truth variants than DeepVariant. Around 90.6% of the discrepancies between the sequence variant genotypes and the truth chip set in both variant callers were due to missing genotypes in the sequence set. Of those, GATK missed proportionally more HOMALT than DeepVariant, and DeepVariant missed proportionally more HET variants. For the remaining ∼9.4% of mismatching genotypes (miscalled), GATK miscalled proportionally more HOM variants, and DeepVariant significantly miscalled proportionally more HET variants (Additional file 20). However, after imputation, the proportion of HET positions miscalled was higher in the GATK set and the proportion of HOMREF positions miscalled as HET was significantly higher in the DeepVariant set.

Recall, precision and F1 score of the filtered query sets were calculated to assess the genotyping accuracy for both variant callers. DeepVariant had strictly better F1 scores than GATK for the filtered data (mean of 0.9719 *versus* 0.9694, Fig. 3.2a and b). The difference was small but significant (Wilcoxon signed-rank test, p=2.3x10^{-10}). As expected, lower coverage (<20x) samples benefited from imputation, improving their F1 scores to values that were comparable to high-coverage samples. Imputation improved GATK genotypes more than DeepVariant genotypes at lower coverages, which could be due to better calibration of genotype likelihoods, but DeepVariant was still strictly better for coverage-folds higher than 7x. Overall, DeepVariant still had a significantly higher mean F1 score for the imputed data (0.9912 *versus* 0.9907, Wilcoxon signed-rank test p=4.2x10^{-05}, Fig. 3.2c).

We examined variant genotyping accuracy through Merfin [25]. Merfin filters out variants when the proportion of "reference" and "alternate" k-mers for that variant from the sample's short sequencing reads does not match the genotype and thus is likely incorrect. HET genotypes obtained with both GATK and DeepVariant had less support from the sequencing reads, as they are more difficult to genotype correctly than HOM genotypes. For both HET and HOMALT, more of the genotypes of DeepVariant

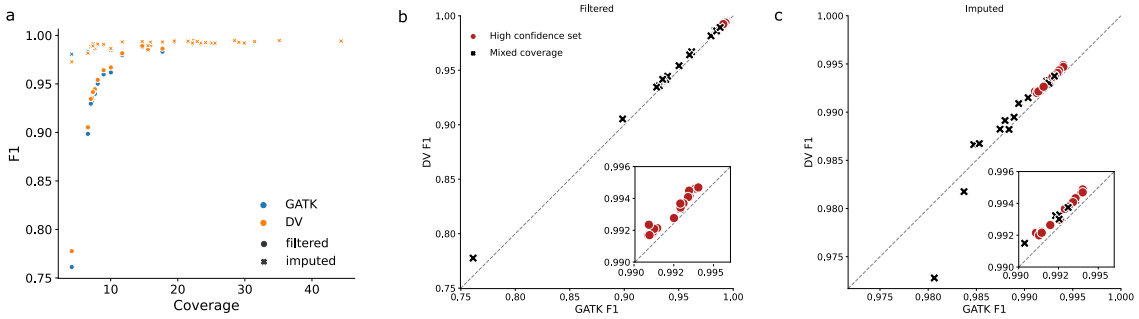


Figure 3.2: **Comparison of the F1 values obtained with hap.py from GATK and DeepVariant (DV) variant calls against the truth chip set for 33 samples.** a) Imputation improves genotype accuracy for sequence coverages lower than 20x but has little impact for sequence coverages higher than 20x. b) DV has a higher F1 score for every sample than GATK for post-filter variants. The high confidence set indicates the 17 microarray genotyped samples out of the 24 samples used later as a truth set for GLIMPSE imputation. c) Similar to (b) but for post-imputation variants.

than of GATK were supported (Fig. 3.3a). The difference between the tools was statistically significant for both genotypes (two-sided paired Wilcoxon test, p_{HET}=3.6x10^{-19}, p_{HOMALT}=1.8x10^{-19}).

In addition, we compared Mendelian concordance rate between the sequenced duos and trios across the two variant callers. There were only two family relationships in the previously examined 50 samples, and so we evaluated the concordance on a separate set of 206 samples (see Methods) forming seven trios (both parents available) and 142 duos (one parent available). DeepVariant had less genotypes that are in conflict with Mendelian inheritance compared to GATK (2.3% *versus* 3.8%, Fig. 3.3b, one-sided paired Wilcoxon signed-rank test p=1.3x10^{-24}). This was due to DeepVariant calling both more genotypes that were compatible as well as fewer that were incompatible with parent-offspring relationship.

Generation of a sequencing validation set for lcWGS imputation

We benchmarked the accuracy of low-pass sequence variant imputation in a target population consisting of 24 BSW samples with a mean autosomal coverage of 28.12 $\pm$ 9.07-fold. DeepVariant identified 15,948,663 variants (87.77% SNPs and 12.23% INDELs) in this 24-sample cohort of which we considered 13,854,932 biallelic SNPs as truth set.

The sequencing reads of these 24 samples were randomly downsampled to mimic medium (4x and 2x), low (1x, 0.5x, 0.25x, and 0.1x), and ultralow (0.01x) sequencing

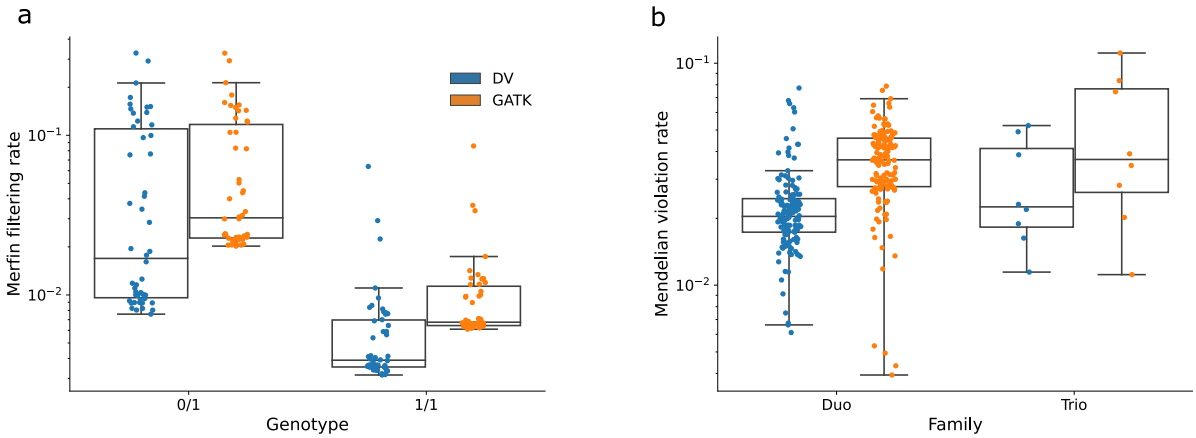


Figure 3.3: **Genotyping accuracy of variant calls validated with sequencing reads and Mendelian relationships.** a) Filtering rate of heterozygous (0/1) and homozygous alternate (1/1) variant calls post-imputation for GATK and DV. Higher filtering rate indicates the genotype/allele is not consistent with k-mers from the same-sample sequencing reads. b) Mendelian violation rate for 206 separate samples, with either 2 family members (Duo) or all 3 (Trio). Mendelian violations are defined as genotypes in the offspring that could not have been inherited from the parents. In the case of duos, only variants homozygous in the parent can be confirmed as violations of Mendelian inheritance.

coverage. We then aligned the reads to the reference sequence and produced genotype likelihoods from the pileup files. Subsequently, genotypes were imputed with GLIMPSE considering nine haplotype reference panels, and compared to the truth set to determine the accuracy of imputation.

The nine haplotype reference panels varied in size and composition. Five haplotype reference panels contained 150 cattle (full panels) of which either 0%, 10%, 25%, 50% or 100% were from the BSW breed (*i.e.*, the breed of the target samples). The other four panels contained either 75 or 30 cattle (reduced panels) that were either from the BSW breed or from breeds other than BSW (see Methods). DeepVariant identified between 17,035,514 and 28,755,400 sequence variants in the nine haplotype reference panels (Table 3.2). The full BSW panel contained 5,167,875 fewer biallelic SNPs than the full non-BSW panel. The 50% multibreed panel had the largest number of variants shared with the truth set and the smallest number of variants present in the truth set but missing in the reference panel, closely followed by the BSW panel. The reduced non-BSW panel (30 samples) had the smallest number of shared variants and the largest number of variants that were present in the truth but missing in the reference set.

Table 3.2: **General overview of the haplotype reference panels: number of samples, coverage and number of variants called.** Shared and private variants are considered through exact matching (position and alleles). Values are the mean of 3 replicas per haplotype panel.

Panel	Samples	Coverage	Variants	Biallelic SNPs	SNPs shared truth-query sets	Truth SNPs missing in haplotype panel	SNPs private to haplotype panel
BSW	150	9.40	22,493,568	19,682,362	13,537,126	317,806	6,145,236
BSW	75	9.65	19,883,488	17,345,201	13,373,462	481,470	3,971,739
BSW	30	9.42	17,035,514	14,839,600	12,810,541	1,044,391	2,029,059
Multibreed (50%)	150	10.48	27,710,504	24,325,185	13,568,744	286,188	10,756,441
Multibreed (25%)	150	10.86	28,755,400	25,266,484	13,531,721	323,211	11,734,763
Multibreed (10%)	150	11.44	28,608,506	25,126,433	13,427,451	427,481	11,698,982
Non-BSW	150	11.78	28,303,738	24,850,237	13,075,827	779,105	11,774,410
Non-BSW	75	11.78	25,059,239	21,968,792	12,868,909	986,023	9,099,883
Non-BSW	30	11.45	21,011,311	18,402,870	12,283,284	1,571,648	6,119,586

Assessment of lcWGS imputation with the different haplotype panels

Increasing the number of reference haplotypes enabled higher F1, recall and precision scores in all tested scenarios (Fig. 3.4a and Additional file 21). Imputation accuracy also improved with increasing lcWGS coverage, with the biggest change between 0.01x and 1x coverage, and continued to improve with diminishing returns between 1x and 4x coverage. The difference in accuracy between panels also decreased as coverage increased.

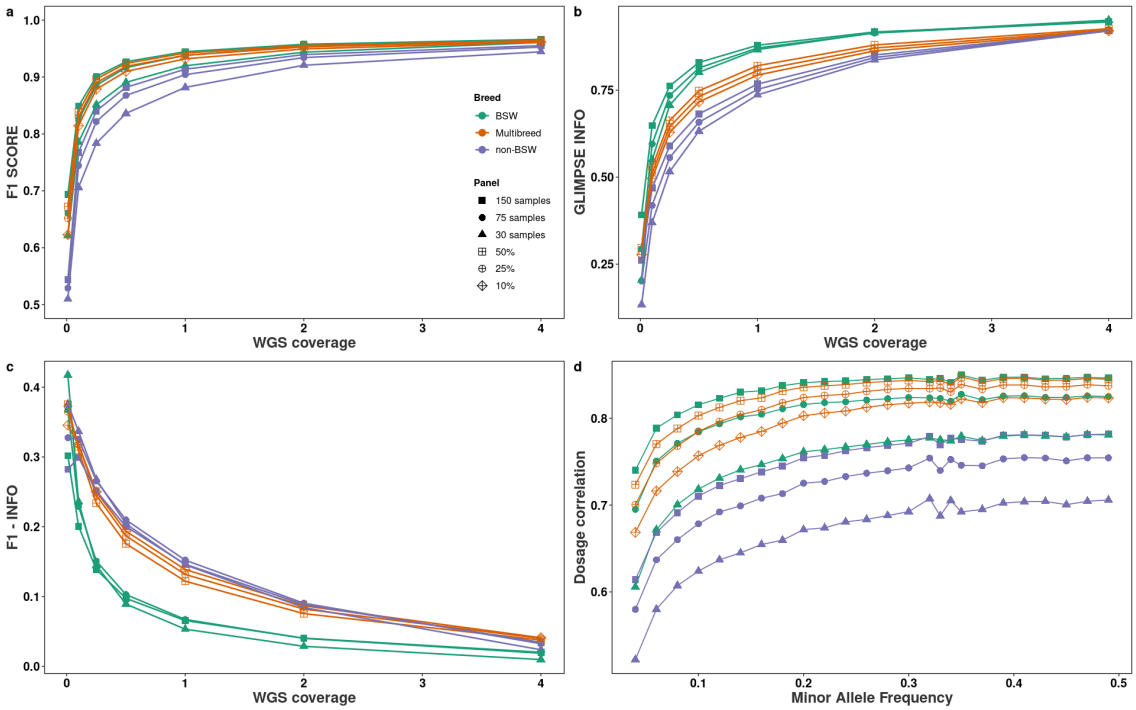


Figure 3.4: **Genotyping accuracy from low-pass whole-genome sequencing.** a) F1 score between truth and imputed variants. b) GLIMPSE INFO score achieved with different sequencing coverages and haplotype panels. c) Differences (subtraction) between F1 and GLIMPSE INFO average scores for different sequencing coverages and haplotype panels. d) Squared dosage correlation (r^2) between imputed data and truth set, stratified by MAF for lcWGS at 0.5x. Panels are indicated with colours and number/percentages of BSW samples are indicated with different shapes of points. Multibreed panels contain 150 samples. Points indicate the average of the results for all variants in three replicates.

The largest BSW haplotype reference panel (n = 150) performed better than any of the multibreed panels at all sequencing coverages. Multibreed panels outperformed BSW panels with a larger number of BSW samples, especially at low coverage. For instance, a large multibreed panel containing 10% BSW samples (n = 15) produced higher

F1 scores than a smaller breed-specific panel containing two times more BSW samples (n = 30). Similarly, a large multibreed panel containing 25% BSW samples (n = 37) provided higher F1 scores than a smaller breed-specific panel containing two times more BSW samples (n = 75) for lcWGS below 1-fold coverage. Accuracies were similar between large multibreed panels and smaller breed-specific panels when the coverage of the lcWGS was higher than 1-fold. All results were validated by three different conformations of the haplotype reference panels (replicas). Standard errors accounting for all the replicas did not overlap for any of the haplotype panels (Additional file 22a).

The INFO score [26] was higher for all BSW panels than for the multibreed panels across all coverages (Fig. 3.4b). A higher proportion of variants were imputed with an INFO score greater than 0.6 in the BSW than in non-BSW or multibreed panels (Additional file 22b). Therefore, panels for which the average INFO score was higher had also a major proportion of variants with high imputation quality, potentially selected for downstream analyses. The differences between BSW panels and the others were larger than those between multibreed and non-BSW panels. The average values of F1 and the average INFO scores were closer for the variants imputed with BSW panels (Fig. 3.4c). The differences between both metrics decreased as the coverage of the lcWGS increased (Additional file 22c and d).

The variants were then stratified by MAF, and the squared correlation of genotype dosages (r^2) was calculated (Fig. 3.4d). The correlations increased along with the MAF similarly for all the panels. The highest correlations were for BSW panel (150 samples) and multibreed panels (50% and 25%). The values increased substantially between 0-0.1 MAF and continued to increase slowly until the MAF reached 0.5 for all panels.

3.3 Discussion

Higher F1 scores against a microarray truth set, improved k-mer based variant filtering, and the fewer Mendelian errors suggest that DeepVariant is a superior variant caller to GATK for bovine short read sequencing. These results extend the evidence of the DeepVariant's greater accuracy that was established in multiple human studies [17, 18, 19, 20]. Ti:Tv ratios in the expected range of 2-2.2 [23, 24] suggest that variant calls private to DeepVariant contain genuine variants, whereas the lower Ti:Tv ratio in variants private to GATK indicate an excess of false positives. DeepVariant revealed more SNPs that have an impact based on their annotation, likely providing additional putative trait-associated candidates for downstream analyses. DeepVariant was approximately 3.5x

faster in end-to-end variant calling compared to GATK, due to greater multithreading potential and to the fact that it does not require pre-processing like GATK's base recalibration step (Additional file 23). The peak memory usage was approximately 65% higher for DeepVariant than for GATK (81 GB *versus* 49 GB). Although our work focused on CPU-only machines, DeepVariant also offers GPU acceleration (roughly 1.9x faster overall), while GATK has no official GPU support, although there are third-party developments (roughly 1.4x faster overall) [27].

To the best of our knowledge, our study is the first to establish bovine haplotype reference panels with DeepVariant. A within-breed panel consisting of 75 samples enabled us to genotype more than 13 million sequence variants in animals sequenced at a 0.5-fold sequencing coverage with F1 scores greater than 0.9. Larger haplotype reference panels (n = 150) from the same breed as the lcWGS data outperform multibreed panels across the whole low coverage spectrum (from 0.1- to 1-fold) and MAF, including rare variants. The development of such panels is a feasible alternative to using much larger multibreed panels, such as the 1000 Bull Genomes project imputation reference panel [13]. Such large panels, encompassing huge within- and across-breed diversity, may be regarded as the most complete and thus best genomic resources available in bovine genomics. However, using such large panels may be detrimental for breed-specific imputation (also described by Nawaz *et al.* [28]), as we observed many relevant sites were filtered out before imputation due to being multiallelic, resulting in a lower F1 score than the 75 sample BSW panel at 1-fold coverage and higher. The use of within-breed panels is also more computationally efficient and are 18% to 33% faster than that of multi- or different-breed panels of the same size (Additional file 24), and approximately 7 times faster than using the 1000 Bull Genomes Project panel.

In the absence of an adequately sized breed-specific panel (*e.g.*, less than 30 animals), F1 scores of 0.9 can also be reached either by increasing the coverage of the lcWGS or by adding distantly related samples from other breeds to the haplotype panels as even animals from seemingly unrelated breeds may share short common haplotypes. Both options will provide accurate sequence variant genotypes at affordable costs for samples from rare breeds, where large breed-specific haplotype reference panels cannot be easily established. For instance, F1 scores > 0.92 are observed at a 2-fold sequencing coverage for all tested haplotype panels with small differences among them. This is likely because higher coverages provide more information for imputation from the own sequencing reads, while lower coverages rely on the information from haplotypes in the panels. We also achieved F1 scores of 0.9 with large multibreed panels containing only 10% of within-breed samples (n = 15). However, reference panels that contain only few samples from the target breed are in general less informative as evidenced by the

lack of about 100K truth SNPs that were present in same-size breed-specific panels. Additionally, a threshold of non-related haplotypes from which only marginal gains to imputation accuracy are observed have been described [15, 28, 29]. Overall results are compatible with similar studies with haplotype panels of both larger and smaller sample sizes [15, 28, 30]. Genotypes imputed from lcWGS enable the prediction of genomic breeding values and facilitate powerful genome-wide association studies at nucleotide resolution [3, 31].

Although imputation accuracy (F1) and GLIMPSE's predicted imputation accuracy (INFO score) are respectively averaged over each sample and each variant, we note that F1 (truth) is strictly higher than INFO (estimation). The differences appear to be more pronounced for reference haplotype panels that are constituted from a different breed to the target sample and at lower coverages (*i.e.*, less than 0.25-fold coverage, where GLIMPSE's INFO scores are inaccurate [6]). While, for example, multibreed panels are nearly as equally accurate as the 150 sample BSW panel, the INFO scores are notably lower. Similarly, the INFO score drops more rapidly for lower coverages, suggesting that a fixed threshold may be unnecessarily conservative given the slower decay in F1. The GLIMPSE INFO score is also positively correlated with variant MAF, and thus filtering based on INFO predominantly removes low-frequency variants. While INFO and other imputation accuracy scores are still useful, additional care should be taken in determining a constant filtering threshold as more and different panels become available for use.

3.4 Conclusions

DeepVariant outperforms GATK for calling variants from bovine short sequencing reads and can be readily used to establish informative haplotype reference panels. Medium sized breed-specific haplotype reference panels enable accurate imputation of millions of sequence variant genotypes from low-pass (0.5-fold) sequence data. The same degree of accuracy of the imputed genotypes is achieved from larger multibreed reference panels that lack individuals from the target breed but contain individuals from distantly related breeds. Increasing the sequencing coverage compensated to a certain extent the lack of representative animals in the reference panels. Nevertheless, suboptimal haplotype reference panels lack variants private to the breed under study, especially rare variants.

3.5 Methods

Data availability and code reproducibility

Short paired-end whole-genome sequencing reads from 501 cattle from nine breeds were used: 327 Brown Swiss (BSW), 50 Fleckvieh, 13 Hereford, 57 Holstein, 2 Nordic Red, 14 Rätisches Grauvieh, 10 Simmental, 25 Tyrolean Grauvieh and 3 Wagyu cattle. Accession numbers for the raw data are available in the online version of the publication [32].

Computational workflows were implemented using Snakemake [33] (version 7.5.0 or newer). The R software environment (version 4.0.2) and ggplot2 package [34] (version 3.3.2) were used to create figures and perform statistical analyses. Scripts and workflows are available online [35].

Alignment, mapping quality and depth of coverage

Raw short sequencing reads were filtered with fastp [36] (version 0.23.1), and MultiQC [37] (version 1.11) was applied to collect the quality metrics across samples. Reads were split per read groups with gdc-fastq-splitter [38] (version 1.0.) and subsequently aligned with bwa-mem2 [39] using the *-M* and *-R* flags to a manually curated version of the current bovine Hereford-based reference genome (ARS-UCD1.2) [40] that included a Y chromosome as described in [41].

Samblaster [42] (version 0.1.26), Sambamba [43], samtools [44, 45] (version 1.12), and Picard tools [46] (version 2.25.7) were used to deduplicate and sort the BAM files.

We calculated average coverage with mosdepth [47] (version 0.3.2) considering all aligned reads that had a mapping quality (MQ) ≥ 10.

Comparison between variant callers

Testing set

Fifty BSW cattle with coverages ranging from 4 to 63-fold were selected as testing set for a comparison between GATK and DeepVariant.

GATK

We used the BaseRecalibrator module of GATK [23, 48] (version 4.2.2.0) to adjust the base quality scores of the deduplicated bam files using 115,815,224 unique positions from the Bovine dbSNP version 150 as known variants. Multi-sample variant calling was performed with the GATK HaplotypeCaller, GenomicsDBImport and GenotypeGVCFs modules according to the best practice guidelines [49, 50]. We applied the VariantFiltration module for site-level filtration using the thresholds indicated in [41] to retain high-quality single nucleotide polymorphisms (SNPs) and insertion/deletion variants (INDELs).

DeepVariant + GLnexus

DeepVariant [16] (version 1.2) was run on the deduplicated bam files using the *WGS* Illumina-trained model, producing a gVCF output per sample. The gVCF files were then merged and filtered using GLnexus [51] (version 1.4.1) with the *DeepVariantWGS* configuration but with the *revise_genotypes* flag set to false.

VCF imputation and statistics

We used Beagle 4.1 [52] (27Jan18.7e1) to improve genotype calls and impute sporadically missing genotypes from genotype likelihoods (*gl* mode). INDELs were left-normalised using bcftools [45] (version 1.12 or 1.15) *norm*. Variant and genotype counts, and Ti:Tv ratios were calculated with bcftools *stats* and bcftools *query*. VCF files were indexed with tabix [53, 54].

Variant annotation

Functional consequences of SNPs were predicted based on the Ensembl (release 104) annotation of the bovine reference assembly using the Variant Effect Predictor tool (VEP) [55] (version 106) with default parameter settings.

Evaluation of the accuracy of variant calling

Microarray-derived genotypes from 33 cattle that also had sequence-derived genotypes were our truth chip set. We intersected the truth (microarray) and query (WGS variants) VCF files using bcftools *isec* with both the *-c none* (exact - only matching REF:ALT alleles are allowed) and *-c all* (position - all coordinate matches are allowed) flags, and retained biallelic SNPs with bcftools *view* to compare the genotypes. Three-way intersection overlaps were counted with bedtools *multiinter* [56] and visualised with UpSetR [57, 58]. Since the microarray data contains fewer sites than WGS, we intersected the truth and query sets. Only positions where the truth genotypes were not homozygous for the reference allele (*i.e.*, the variants that segregate within the target samples) were retained. We calculated recall (percentage of true positives in the query set), precision (proportion of matching genotypes in both truth and query sets), and F1 scores (harmonic mean of precision and recall) using hap.py [59] (version 0.3.9) on a per-sample basis. Agreement between the imputed variant alleles/genotypes and raw sequencing reads was assessed with Merfin's k-mer-based filtering method [25] (commit fc4f89a). A k-mer database was prepared using Meryl (commit 51fad4b) with a k-mer size of 21 and minimum k-mer occurrence of 2 in the short sequencing reads. Variants that were poorly supported, *i.e.*, the alternate sequence (variant and flanking regions) appeared less often in k-mers than the reference sequence did in a genotype-aware proportion, were filtered out.

We assessed Mendelian consistency in filtered but not-imputed data from parent-offspring pairs and trios (accession IDs can be found in the online version of the publication [32]) using the bcftools *+mendelian* plugin [45]. We calculated discrepancy rate as the number of inconsistent sites divided by the total number of non-missing sites. For duos (dam-offspring or sire-offspring) only homozygous sites were considered. Assessing discrepancy was only possible when the parent genotype was homozygous (0/0 or 1/1).

Imputation of low-pass sequencing data

Generation of the haplotype panels

The BSW reference panels contained 150, 75 and 30 samples that were randomly selected from 303 BSW samples. The non-BSW panels contained 150, 75 and 30 samples that were randomly selected from 174 non-BSW samples. The multibreed panels were randomly selected from a combination of the above, and they contained 150 samples of which

50%, 25%, and 10% were BSW samples and the remaining were non-BSW. Three random replicates for each panel were created. Sequence variant genotypes were called for each panel with DeepVariant and sporadically missing genotypes were imputed with Beagle 4.1 [52] (27Jan18.7e1) as described above.

Truth sequencing set, truth variants and subsampling

Variants were called with DeepVariant and GLnexus as described previously for 24 BSW samples with a coverage higher than 20-fold to generate a truth set for assessing imputation accuracy. The raw whole-genome sequencing reads of the 24 BSW samples were then downsampled with seqtk [60] to mimic 4x, 2x, 1x, 0.5x, 0.25x, 0.1x, and 0.01x coverage, and subsequently aligned to ARS-UCD12 as described previously.

Genotype likelihoods for the variants that are present in the haplotype reference panel were estimated from the subsampled read alignments with bcftools *mpileup* and bcftools *call*. These were then imputed using the different haplotype panels and GLIMPSE [61] (version 1.1.1). We used 2-Mb windows and 200-kb buffer sizes during the chunk step followed by phasing and ligation to produce the final imputed variant calls.

Comparison of true and imputed variants

The accuracy of the imputed sequence variant genotypes was assessed with hap.py as described above. The minor allele frequency (MAF) of the imputed sequence variants was calculated with PLINK [62] (version 1.9). The estimated imputation quality was retrieved from the INFO flag from the VCF files produced by GLIMPSE with bcftools *query*. Pearson squared correlation between expected and actual dosages (r^2) was calculated with the bcftools *stats*.

References

[1] Michel Georges, Carole Charlier, and Ben Hayes. Harnessing genomic information for livestock improvement. *Nat Rev Genet*, 20:135–56, 2019. ISSN 1471-0064. doi: 10.1038/s41576-018-0082-2.

[2] Paul M. VanRaden, Melvin E. Tooker, Jeffrey R. O'Connell, John B. Cole, and Derek M. Bickhart. Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet Sel Evol*, 49:32, 2017. ISSN 1297-9686. doi: 10.1186/s12711-017-0307-4.

[3] Warren M. Snelling, Jesse L. Hoff, Jeremiah H. Li, Larry A. Kuehn, Brittney N. Keel, Amanda K. Lindholm-Perry, and Joseph K. Pickrell. Assessment of imputation from low-pass sequencing to predict merit of beef steers. *Genes (Basel)*, 11:1312, 2020. ISSN 2073-4425. doi: 10.3390/genes11111312.

[4] Roger Ros-Freixedes, Andrew Whalen, Gregor Gorjanc, Alan J. Mileham, and John M. Hickey. Evaluation of sequencing strategies for whole-genome imputation with hybrid peeling. *Genet Sel Evol*, 52: 18, 2020. ISSN 1297-9686. doi: 10.1186/s12711-020-00537-7.

[5] Robert W. Davies, Marek Kucka, Dingwen Su, Sinan Shi, Maeve Flanagan, Christopher M. Cunniff, Yingguang Frank Chan, and Simon Myers. Rapid genotype imputation from sequence with reference panels. *Nat Genet*, 53:1104–11, 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00877-0.

[6] Jun Teng, Changheng Zhao, Dan Wang, Zhi Chen, Hui Tang, Jianbin Li, Cheng Mei, Zhangping Yang, Chao Ning, and Qin Zhang. Assessment of the performance of different imputation methods for low-coverage sequencing in holstein cattle. *J Dairy Sci*, 105:3355–66, 2022. ISSN 0022-0302. doi: 10.3168/jds.2021-21360.

[7] Bogdan Pasaniuc, Nadin Rohland, Paul J. McLaren, Kiran Garimella, Noah Zaitlen, Heng Li, Namrata Gupta, Benjamin M. Neale, Mark J. Daly, Pamela Sklar, Patrick F. Sullivan, Sarah Bergen, Jennifer L. Moran, Christina M. Hultman, Paul Lichtenstein, Patrik Magnusson, Shaun M. Purcell, David W. Haas, Liming Liang, Shamil Sunyaev, Nick Patterson, Paul I. W. de Bakker, David Reich, and Alkes L. Price. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat Genet*, 44:631–5, 2012. ISSN 1546-1718. doi: 10.1038/ng.2283.

[8] Roger Ros-Freixedes, Andrew Whalen, Ching-Yi Chen, Gregor Gorjanc, William O. Herring, Alan J. Mileham, and John M. Hickey. Accuracy of whole-genome sequence imputation using hybrid peeling in large pedigreed livestock populations. *Genet Sel Evol*, 52:17, 2020. ISSN 1297-9686. doi: 10.1186/s12711-020-00536-8.

[9] Runyang Nicolas Lou, Arne Jacobs, Aryn P. Wilder, and Nina Overgaard Therkildsen. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Mol Ecol*, 30:5966–93, 2021. ISSN 1365-294X. doi: 10.1111/mec.16077.

[10] 1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, Jonathan L. Marchini, Shane McCarthy, Gil A. McVean, and Gonçalo R. Abecasis. A global reference for human genetic variation. *Nature*, 526:68–74, 2015. ISSN 1476-4687. doi: 10.1038/nature15393.

[11] Shane McCarthy, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R. Wood, Alexander Teumer, Hyun Min Kang, Christian Fuchsberger, Petr Danecek, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*, 48:1279–83, 2016. ISSN 1546-1718. doi: 10.1038/ng.3643.

[12] Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E. Locke, Alan Kwong, Scott I. Vrieze, Emily Y. Chew, Shawn Levy, Matt McGue, David Schlessinger, Dwight Stambolian, Po-Ru Loh, William G. Iacono, Anand Swaroop, Laura J. Scott, Francesco Cucca, Florian Kronenberg, Michael Boehnke, Gonçalo R. Abecasis, and Christian Fuchsberger. Next-generation genotype imputation service and methods. *Nature Genetics*, 48(10):1284–1287, October 2016. ISSN 1546-1718. doi: 10.1038/ng.3656.

[13] Hans D. Daetwyler, Aurélien Capitan, Hubert Pausch, Paul Stothard, Rianne van Binsbergen, Rasmus F. Brøndum, Xiaoping Liao, Anis Djari, Sabrina C. Rodriguez, Cécile Grohs, Diane Esquerré, Olivier Bouchez, Marie-Noëlle Rossignol, Christophe Klopp, Dominique Rocha, Sébastien Fritz, André Eggen, Phil J. Bowman, David Coote, Amanda J. Chamberlain, Charlotte Anderson, Curt P. VanTassell, Ina Hulsegge, Mike E. Goddard, Bernt Guldbrandtsen, Mogens S. Lund, Roel F. Veerkamp, Didier A. Boichard, Ruedi Fries, and Ben J. Hayes. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet*, 46:858–65, 2014. ISSN 1546-1718. doi: 10.1038/ng.3034.

[14] Wenqian Yang, Yanbo Yang, Cecheng Zhao, Kun Yang, Dongyang Wang, Jiajun Yang, Xiaohui Niu, and Jing Gong. Animal-ImputeDB: A comprehensive database with multiple animal reference panels for genotype imputation. *Nucleic Acids Res*, 48:D659–67, 2020. ISSN 1362-4962. doi: 10.1093/nar/gkz854.

[15] Zhen Wang, Zhenyang Zhang, Zitao Chen, Jiabao Sun, Caiyun Cao, Fen Wu, Zhong Xu, Wei Zhao, Hao Sun, Longyu Guo, Zhe Zhang, Qishan Wang, and Yuchun Pan. PHARP: A pig haplotype reference panel for genotype imputation. *Sci Rep*, 12:12645, 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-15851-x.

[16] Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T. Afshar, Sam S. Gross, Lizzie Dorfman, Cory Y. McLean, and Mark A. DePristo. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*, 36:983–7, 2018. ISSN 1546-1696. doi: 10.1038/nbt.4235.

[17] Taedong Yun, Helen Li, Pi-Chuan Chang, Michael F. Lin, Andrew Carroll, and Cory Y. McLean. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics*, 36:5582–9, 2021. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaa1081.

[18] Yi-Lin Lin, Pi-Chuan Chang, Ching Hsu, Miao-Zi Hung, Yin-Hsiu Chien, Wuh-Liang Hwu, FeiPei Lai, and Ni-Chung Lee. Comparison of GATK and DeepVariant by trio sequencing. *Sci Rep*, 12:1809, 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-05833-4.

[19] Jared O'Connell, Taedong Yun, Meghan Moreno, Helen Li, Nadia Litterman, Alexey Kolesnikov, Elizabeth Noblin, Pi-Chuan Chang, Anjali Shastri, Elizabeth H. Dorfman, Suyash Shringarpure, Adam Auton, Andrew Carroll, and Cory Y. McLean. A population-specific reference panel for improved genotype imputation in African Americans. *Commun Biol*, 4:1269, 2021. ISSN 2399-3642. doi: 10.1038/s42003-021-02777-9.

[20] Raphael O. Betschart, Alexandre Thiéry, Domingo Aguilera-Garcia, Martin Zoche, Holger Moch, Raphael Twerenbold, Tanja Zeller, Stefan Blankenberg, and Andreas Ziegler. Comparison of calling pipelines for whole genome sequencing: an empirical study demonstrating the importance of mapping and alignment. *Sci Rep*, 12:21502, 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-26181-3.

[21] Justin M. Zook, Jennifer McDaniel, Nathan D. Olson, Justin Wagner, Hemang Parikh, Haynes Heaton, Sean A. Irvine, Len Trigg, Rebecca Truty, Cory Y. McLean, Francisco M. De La Vega, Chunlin Xiao, Stephen Sherry, and Marc Salit. An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol*, 37:561–6, 2019. ISSN 1546-1696. doi: 10.1038/s41587-019-0074-6.

[22] T. Yun, C. McLean, PC. Chang, and A. Carroll. Improved non-human variant calling using species-specific deepvariant models. https://google.github.io/deepvariant/posts/2018-12-05-improved-non-human-variant-calling-using-species-specific-deepvariant-models/. [Accessed 26th April 2023].

[23] Mark A. DePristo, Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, Guillermo del Angel, Manuel A. Rivas, Matt Hanna, Aaron McKenna, Tim J. Fennell, Andrew M. Kernytsky, Andrey Y. Sivachenko, Kristian Cibulskis, Stacey B. Gabriel, David Altshuler, and Mark J. Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43:491–8, 2011. ISSN 1546-1718. doi: 10.1038/ng.806.

[24] Matthew N. Bainbridge, Min Wang, Yuanqing Wu, Irene Newsham, Donna M. Muzny, John L. Jefferies, Thomas J. Albert, Daniel L. Burgess, and Richard A. Gibbs. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome Biol*, 12:R68, 2011. ISSN 1474-760X. doi: 10.1186/gb-2011-12-7-r68.

[25] Giulio Formenti, Arang Rhie, Brian P. Walenz, Françoise Thibaud-Nissen, Kishwar Shafin, Sergey Koren, Eugene W. Myers, Erich D. Jarvis, and Adam M. Phillippy. Merfin: improved variant filtering, assembly evaluation and polishing via k-mer validation. *Nat Methods*, 19:696–704, 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01445-y.

[26] Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 11:499–511, 2010. ISSN 1471-0064. doi: 10.1038/nrg2796.

[27] Shanshan Ren, Nauman Ahmed, Koen Bertels, and Zaid Al-Ars. GPU accelerated sequence alignment with traceback for GATK HaplotypeCaller. *BMC Genomics*, 20:184, 2019. ISSN 1471-2164. doi: 10.1186/s12864-019-5468-9.

[28] Muhammad Yasir Nawaz, Priscila Arrigucci Bernardes, Rodrigo Pelicioni Savegnago, Dajeong Lim, Seung Hwan Lee, and Cedric Gondro. Evaluation of whole-genome sequence imputation strategies in Korean Hanwoo cattle. *Animals (Basel)*, 12:2265, 2022. ISSN 2076-2615. doi: 10.3390/ani12172265.

[29] Sanne van den Berg, Jérémie Vandenplas, Fred A. van Eeuwijk, Aniek C. Bouwman, Marcos S. Lopes, and Roel F. Veerkamp. Imputation to whole-genome sequence using multiple pig populations and its use in genome-wide association studies. *Genet Sel Evol*, 51:2, 2019. ISSN 0999-193X. doi: 10.1186/s12711-019-0445-y.

[30] Aine C. O'Brien, Michelle M. Judge, Sean Fair, and Donagh P. Berry. High imputation accuracy from informative low-to-medium density single nucleotide polymorphism genotypes is achievable in sheep. *J Anim Sci*, 97:1550–67, 2019. ISSN 1525-3163. doi: 10.1093/jas/skz043.

[31] Adéla Nosková, Meenu Bhati, Naveen Kumar Kadri, Danang Crysnanto, Stefan Neuenschwander, Andreas Hofer, and Hubert Pausch. Characterization of a haplotype-reference panel for genotyping by low-pass sequencing in Swiss Large White pigs. *BMC Genomics*, 22:290, 2021. ISSN 1471-2164. doi: 10.1186/s12864-021-07610-5.

[32] Audald Lloret-Villas, Hubert Pausch, and Alexander S. Leonard. The size and composition of haplotype reference panels impact the accuracy of imputation from low-pass sequencing in cattle. *Genetics Selection Evolution*, 55(1):33, May 2023. ISSN 1297-9686. doi: 10.1186/s12711-023-00809-y.

[33] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O. Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. Sustainable data analysis with Snakemake. *F1000Res*, 10:33, 2021. ISSN 2046-1402. doi: 10.12688/f1000research.29032.2.

[34] Hadley Wickham. *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

[35] Audald Lloret-Villas. Animal genomics eth github: Imputation of low-pass data, 2023. URL https://github.com/AnimalGenomicsETH/Low_pass_imputation.

[36] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34:i884–90, 2018. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty560.

[37] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32:3047–8, 2016. ISSN 1367-4811. doi: 10.1093/bioinformatics/btw354.

[38] K. Hernandez. Cli for splitting a fastq that has multiple readgroups. https://github.com/kmhernan/gdc-fastq-splitter. [Accessed 26th April 2023].

[39] Vasimuddin Md, Sanchit Misra, Heng Li, and Srinivas Aluru. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. preprint on arXiv. https://arxiv.org/abs/1907.12931v1, jul 2019.

[40] Benjamin D. Rosen, Derek M. Bickhart, Robert D. Schnabel, Sergey Koren, Christine G. Elsik, Elizabeth Tseng, Troy N. Rowan, Wai Y. Low, Aleksey Zimin, Christine Couldrey, Richard Hall, Wenli Li, Arang Rhie, Jay Ghurye, Stephanie D. McKay, Françoise Thibaud-Nissen, Jinna Hoffman, Brenda M. Murdoch, Warren M. Snelling, Tara G. McDaneld, John A. Hammond, John C. Schwartz, Wilson Nandolo, Darren E. Hagen, Christian Dreischer, Sebastian J. Schultheiss, Steven G. Schroeder, Adam M. Phillippy, John B. Cole, Curtis P. Van Tassell, George Liu, Timothy P. L. Smith, and Juan F. Medrano. De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience*, 9: giaa021, 2020. ISSN 2047-217X. doi: 10.1093/gigascience/giaa021.

[41] Audald Lloret-Villas, Meenu Bhati, Naveen Kumar Kadri, Ruedi Fries, and Hubert Pausch. Investigating the impact of reference assembly choice on genomic analyses in a cattle breed. *BMC Genomics*, 22:363, 2021. ISSN 1471-2164. doi: 10.1186/s12864-021-07554-w.

[42] Gregory G. Faust and Ira M. Hall. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*, 30:2503–5, 2014. ISSN 1367-4811. doi: 10.1093/bioinformatics/btu314.

[43] Artem Tarasov, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, 31:2032–4, 2015. ISSN 1367-4811. doi: 10.1093/bioinformatics/btv098.

[44] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25:2078–9, 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp352.

[45] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10:giab008, 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab008.

[46] Picard. Picard toolkit. https://broadinstitute.github.io/picard/. [Accessed 26th April 2023].

[47] Brent S. Pedersen and Aaron R. Quinlan. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*, 34:867–8, 2018. ISSN 1367-4811. doi: 10.1093/bioinformatics/btx699.

[48] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20:1297–303, 2010. ISSN 1549-5469. doi: 10.1101/gr.107524.110.

[49] Geraldine A. van der Auwera, Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo Del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Kiran V. Garimella, David Altshuler, Stacey Gabriel, and Mark A. DePristo. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*, 43:11.10.1–33, 2013. ISSN 1934-340X. doi: 10.1002/0471250953.bi1110s43.

[50] GATK. Gatk blog. https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels. [Accessed 26th April 2023].

[51] Michael F. Lin, Ohad Rodeh, John Penn, Xiaodong Bai, Jeffrey G. Reid, Olga Krasheninina, and William J. Salerno. GLnexus: joint variant calling for large cohort sequencing. preprint on bioRxiv. https://www.biorxiv.org/content/10.1101/343970v1, jun 2018.

[52] Brian L. Browning and Sharon R. Browning. Genotype imputation with millions of reference samples. *Am J Hum Genet*, 98:116–26, 2016. ISSN 1537-6605. doi: 10.1016/j.ajhg.2015.11.020.

[53] Heng Li. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, 27:718–9, 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq671.

[54] James K. Bonfield, John Marshall, Petr Danecek, Heng Li, Valeriu Ohan, Andrew Whitwham, Thomas Keane, and Robert M. Davies. HTSlib: C library for reading/writing high-throughput sequencing data. *GigaScience*, 10:giab007, 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab007.

[55] William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The Ensembl Variant Effect Predictor. *Genome Biol*, 17:122, 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0974-4.

[56] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26:841–2, 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq033.

[57] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. UpSet: visualization of intersecting sets. *IEEE Trans Vis Comput Graph*, 20:1983–92, 2014. ISSN 1941-0506. doi: 10.1109/TVCG.2014.2346248.

[58] Jake R Conway, Alexander Lex, and Nils Gehlenborg. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33:2938–40, 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx364.

[59] hap.py. Haplotype vcf comparison tools. https://github.com/Illumina/hap.py. [Accessed 26th April 2023].

[60] seqtk. seqtk github. https://github.com/lh3/seqtk. [Accessed 26th April 2023].

[61] Simone Rubinacci, Diogo M. Ribeiro, Robin J. Hofmeister, and Olivier Delaneau. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat Genet*, 53:120–6, 2021. ISSN 1546-1718. doi: 10.1038/s41588-020-00756-0.

[62] Christopher C. Chang, Carson C. Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4:7, 2015. ISSN 2047-217X. doi: 10.1186/s13742-015-0047-8.

Chapter 4

A preliminary analysis of methylation profiles obtained with HiFi data

Audald Lloret-Villas[1], Xena M. Mapel[1], Alexander S. Leonard[1] and Hubert Pausch[1]

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

Manuscript in preparation

Contribution: I participated in conceiving the study, designing the experiments, analysing the results and writing the draft. I also developed and compiled reproducible pipelines.

4.1 Background

Genetic variants are responsible for only part of trait variation [1]. Epigenetic marks are transmitted during meiosis and become critical for regulating gene expression and hence the phenotype of the cell [1, 2]. Methylation is an important aspect of the epigenome, with a central role in cell differentiation and development [2, 3]. DNA methylation induces epigenetic variations to the cytosine bases in the DNA sequence via the addition of a methyl group through the action of DNA methyltransferase [1]. Differential methylation may contribute substantially to variation in bovine traits, including production and fertility [1, 3]. The most common form of methylation occurs at the fifth carbon of the pyrimidine ring of cytosines (5-methylcytosine (5mC)) and is an essential modification for normal mammalian development and gene regulation [4, 5, 6]. This modification exists mostly at CpG dinucleotides and to a lesser degree in non-CpG contexts [1]. CpG sites occur with high frequency in genomic regions called CpG islands (CGIs) [7].

Commonly used methods for detecting 5mC rely on Illumina BeadChip platforms or involve sequencing with bisulfite treatment [2, 4, 8]. Chip-based sequencing is limited to a predefined subset of ∼850,000 CpG methylation sites (around a 3% of the 3.7 million total sites in mammalian genomes) [1, 8]. Bisulfite treatment, on the other hand, is a harsh process that converts unmethylated cytosines into uracils, which results in extensive DNA degradation and requires high amounts of input DNA [1, 2, 4]. Single molecule real time (SMRT) sequencing with ONT and PacBio overcomes these limitations by offering the possibility to obtain the canonical base sequencing and cytosine methylation status simultaneously with around half of the coverage required by bisulfite sequencing methods [2, 9, 10]. Long reads span more variants and thus can additionally reduce the high mapping uncertainty linked to short-read sequencing and be assigned to the relevant haplotypes (phased) [5, 6, 9]. Pacific Biosciences (PacBio) is a SMRT sequencing platform that uses circular consensus sequencing [11] to produce high fidelity (HiFi) reads as illustrated in Fig. 4.1. PacBio SMRT sequencing also reveals 5mC base modification based on signature changes in polymerase kinetics as the base is replicated by the polymerase [12]. Different copies of the same molecule can present considerable kinetics variability and thus computational tools are required to confidently identify the methylation status [6]. Sequencing advances have been accompanied by the rapid development of software tools to measure and call base modifications within these data [9, 12].

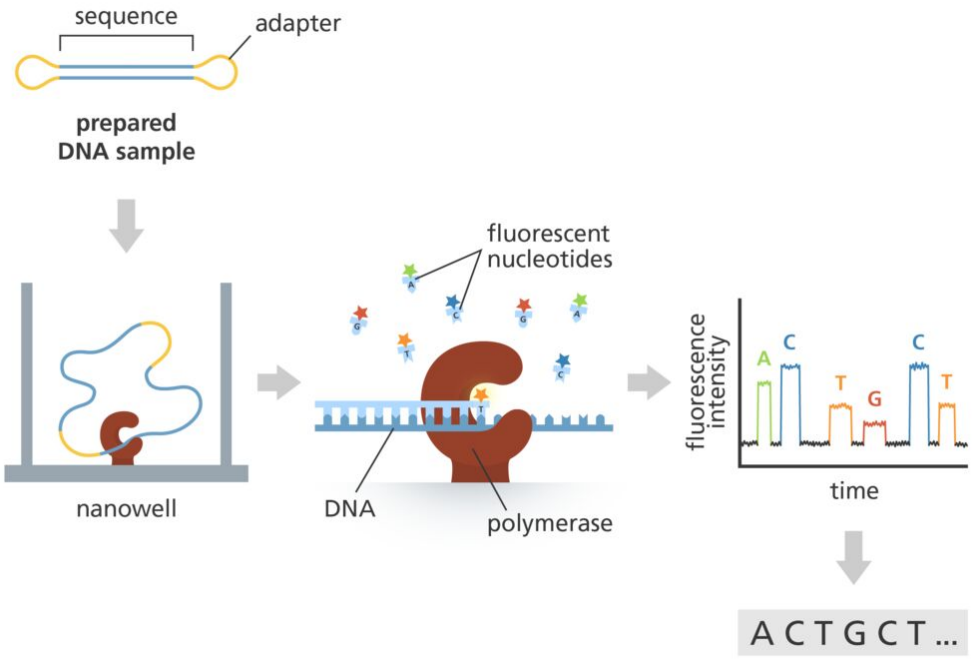


Figure 4.1: **PacBio SMRT sequencing.** Prepared DNA sample is bound to the polymerase (brown) via a linker (yellow) and placed onto the flow cell. A DNA-polymerase complex is connected to the bottom of each nanowell and a mix of fluorescently labelled nucleotides is added. The incorporation of each fluorescent nucleotides leads to a burst of light captured in the raw video data. A base calling algorithm translates the fluorescent intensity signal into its original DNA sequence. Credit: Laura Olivares Boldú / Wellcome Connecting Science: https://dam.wellcomegenomecampus.org/web/5410f7067f6803c7/yourgenome-illustrations/.

Single-molecule methylation data offer access to the variability of methylation across individual molecules (coordinated patterns of DNA methylation occurring in blocks) and reads [8, 9]. The DNA methylation patterns of different physiological conditions, development phases, and cell types are of enormous interest to interpret the mammalian gene regulation mechanism [3]. Here, 5mC modifications from HiFi data are used to perform a preliminary analysis of the methylation patterns along 39,376 autosomal CGIs in 105 testis samples and 15 epididymis samples from 120 different BSW bulls.

4.2 Results

120 sequenced male animals had an average fold autosomal coverage of 14.49 ± 7.88. 15 of the samples were sequenced from epididymis and 105 were sequenced from testes, yielding an average fold autosomal coverage of 12.70 ± 1.75 and 14.74 ± 8.37, respectively. Four of the testis samples were sequenced with both a Sequel IIe and a Revio machine and had an average fold autosomal coverage of 50.1 ± 5.2.

On average, 25.22M ± 0.67M of the autosomal cytosines were methylated across the 120 samples, ranging from 21.70M to 25.82M. 63.7% of the 5mC variance among samples was explained by the tissue type, as clusters of the two surveyed tissues were well defined (Fig. 4.2). The distribution of methylation scores across the CpG dinucleotides was clearly skewed towards methylated 5mC in both tissues (Fig. 4.3). The number of CpGs reached consistent peaks between 2 - 6 and 90 - 97 methylation scores for all samples. Considering only the 5mC nucleotides that had a minimum coverage of 6-fold, the number of methylated 5mC (methylation score > 80) was 6.88 and 7.32 times the number of unmethylated 5mC (methylation score < 20) for epididymis and testis samples, respectively.

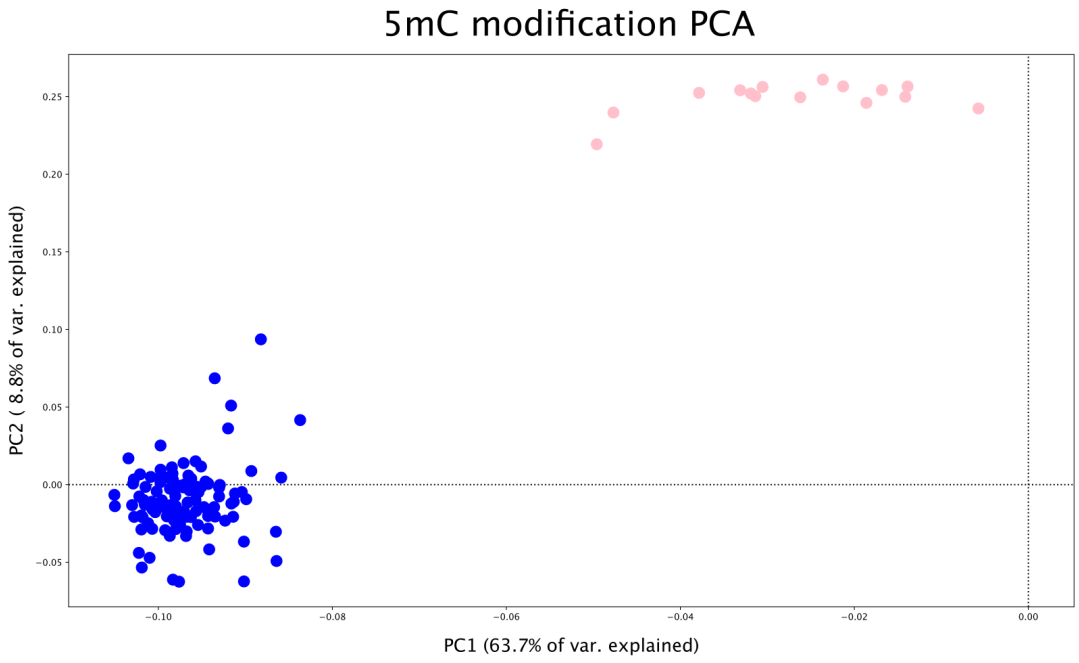


Figure 4.2: **Principal Component Analyses of the 5mC sites of the 120 analysed samples.** 105 testis samples (blue circles) cluster in the bottom left corner and 15 epididymis samples (pink circles) cluster in the upper right side.

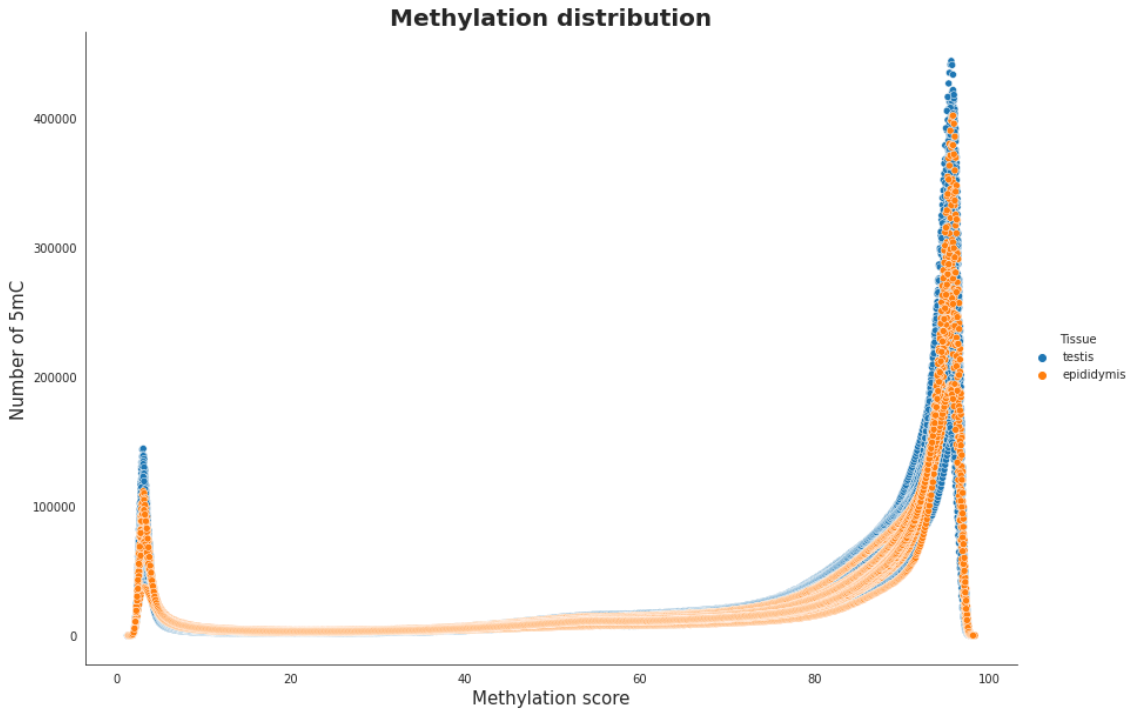


Figure 4.3: **Distribution of methylation score across the 120 samples (105 testis and 15 epididymis).**

39,376 annotated CpG islands (CGIs) were considered across the bovine autosomes (see Methods). The number of CGIs per chromosome ranged from 401 (chromosome 28) to 2,306 (chromosome 18). The average number of CpG dinucleotides within the CGIs along the genome was of 64.54 with a substantial variability (standard deviation of 72.57) ranging from 0 to 2,270, although 75% of the CGIs contained less than 90 CpGs and only 2% exceeded 250 CpGs per CGI. 4.0% and 1.2% of the CGIs did not have methylation information (0 CpGs) for at least 25% of the samples or had a low methylation variability (< than 1 standard deviation) for testes and epididymis, respectively. The resulting CGIs (38,904 for testis samples and 37,793 for epididymis samples) can be considered for further association studies.

The methylation patterns of several genes that were reported to be differentially expressed in testis and epididymis [13] were visually explored for a subset of samples. Gene TBX4 (ENSBTAG00000009968) is expressed in testis but not in epididymis whereas gene CACNG7 (ENSBTAG00000007506) is expressed in epididymis but not in testis. The 5mC methylation of these CpG dinucleotides was consistently higher for epididymis than for testis in both instances. The mean methylation score of the closest CGI (within the gene body) was 39.86 and 41.29 for testis, and 84.80 and 84.76 for epididymis for TBX4 and CACNG7, respectively. Fig. 4.4 illustrates visual differences in

CpG methylation in the body of both TBX4 and CACNG7 genes.

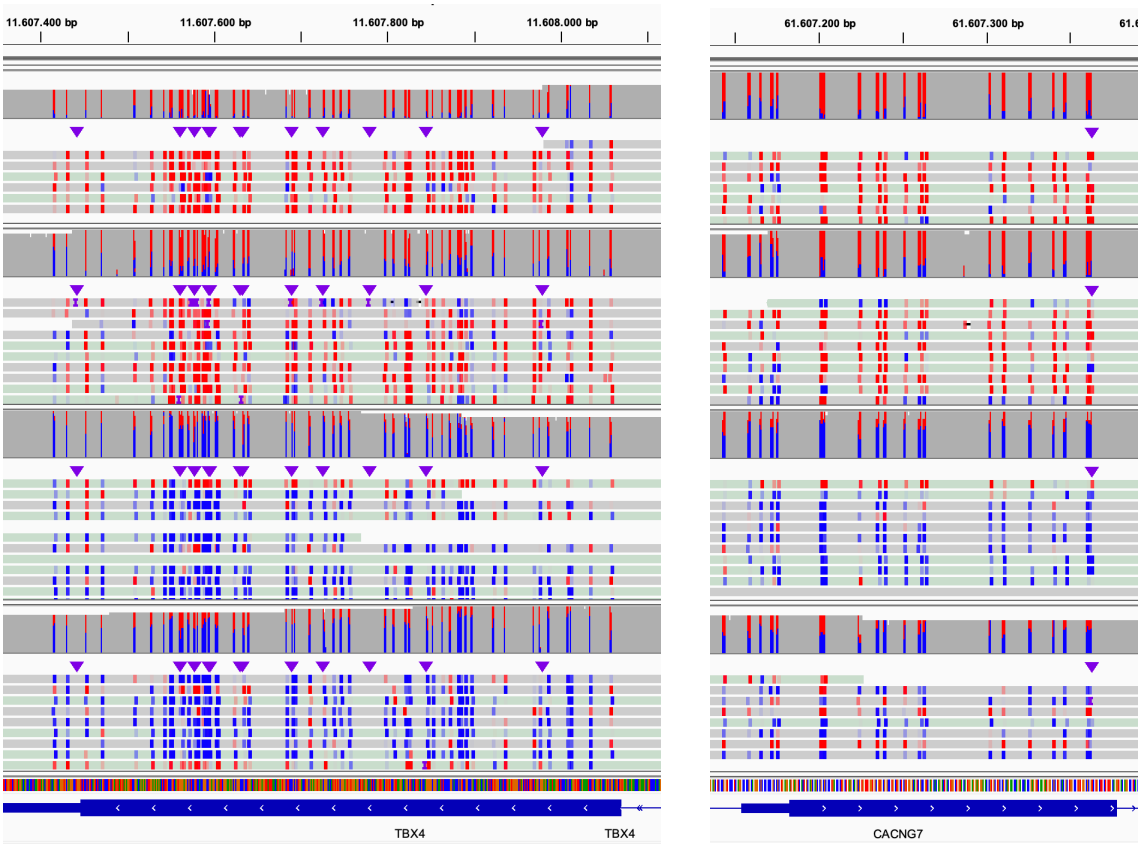


Figure 4.4: **5mC methylation patterns for 2 samples of epididymis (above) and 2 samples of testes (below).** Genes TBX4 (19:11607447-11633858) and CACNG7 (18:61607183-61622688) are surveyed.

4.3 Discussion

The 5mC methylation landscape is tissue-dependent and is tightly linked to genetic variation and gene expression [7, 14]. The combined development of sequencing technologies and base-callers offer an unprecedented opportunity to quantify formerly undetected epigenetic information [15]. This exploration serves as a proof of concept for direct retrieval and quantification of 5mC modifications from long-read data in bovine tissues. Distinguishable methylation patterns were observed across tissues, but variability also existed within tissue-specific samples. Asymmetrically methylated CpG dyads (hemimethylated CpGs) [16] were infrequent since the vast majority of the CpG dinucleotides were either fully methylated (preferentially) or fully unmethylated.

Hundreds of annotated CGIs had no methylation information. Even the highest coverage sample (56-fold) had 0 CpGs in 395 CGIs, mainly at the end of the chromosomes. Telomeres and centromeres are typically formed by repetitive regions that are challenging to assemble and annotate [17]. HiFi sequencing reads span long repeats and can better align to these regions. Hence, plausible reasons why no sequencing reads mapped these CGIs could be the flaws in the reference assembly or the alignments being considered as secondary. CGIs were filtered when not enough methylation information was available or when the variability was low, as these CGIs are not informative for downstream analyses. A substantial difference of 2.8 percent points was detected between filtered CGIs in epididymis and testes. Such differences are likely caused by the lower coverage (average of 2-fold) of epididymis samples in comparison to testes. The main factor that influenced filtering was the lack of CpGs in at least 25% of the samples in a given CGI: 3.5% and 1.1% of the CGIs in epididymis and testis, respectively. The remaining 0.5% and 0.1% CGIs were filtered due to similar methylation scores (less than 1 standard deviation) across all samples.

5mC methylation in the gene body showed different patterns in several genes that are expressed tissue-specifically. The biological context of methylation calls is complex and depends on the chromosomal location and the relationship with expression [7]. Generally, when methylation is found in CpGs that are part of genes, transcription is increased [1]. This was aligned with the observations for gene CACNG7 but not for gene TBX4, which held the reverse pattern. Fig. 4.4 offers only a snapshot of the methylation status within the gene body, but further context of methylation is required to fully understand the repression/enhancement of expression (for instance the methylation patterns in promoters or transcription start sites (TSSs)). Methylation or unmethylation of CGIs in such regions directly affect the accessibility to transcriptional activation and hence the enhancement of gene expression [1, 2, 4, 8, 18]. Large number of samples from different tissues and the inclusion of non-coding sequences in the analysis would enable the statistical exploration of differentially methylated regions and its relation to gene expression.

In heterozygous alleles and providing cis activity of the QTL (this is, alleles affecting proximal features), phenotypes such as methylation can be measured on the maternal chromosome in comparison to the parental chromosome [5]. Dozens of genes, including SNURF, PLAGL1, NAP1L5, ZIM2, IGF2, SLC38A4, TSSC4, SNRPN, MEST, IGF2R or PEG10 are imprinted in mammals [1, 4]. A visual exploration showed no clear differences between phased methylation profiles of these genes in high coverage samples (50.1-fold) that were sequenced with both Sequel and Revio machines. A further systematic determination of differentially methylated regions in imprinted genes as well

as other regions with parent of origin effect will enable allele-specific QTL analysis, similarly to what Prowse-Wilking *et al.* [18] performed with heights/peaks of histone modifications.

4.4 Methods

PacBio High Fidelity (HiFi) reads from 48 and 72 samples were sequenced on 158 SMRT cells with a Sequel IIe and with a Revio, respectively. Four of these samples were sequenced with both machines. Default CCS parameters were used, resulting in onboard 5mC methylation calls.

The reads were aligned with Minimap2 (v2.24) [19, 20] using the flags "-x map-hifi -y -Y" to a manually curated version of the current bovine Hereford-based reference genome (ARS-UCD1.2) [21] that included a Y chromosome as described in [22]. Coverage of the CRAM files was calculated with samtools coverage (minimum mapping quality of (-q) 20).

Single nucleotide variants (SNVs) were called with DeepVariant [23] (version 1.5. - PACBIO model). The gVCF files were then merged and filtered using GLnexus [24] (version 1.4.1 - DeepVariantWGS configuration).

Site methylation probabilities from the mapped HiFi reads were generated by the tool "aligned_bam_to_cpg_scores" (pb-CpG-tools [25]) using the "model" pileup mode. The aligned reads were phased with HiPhase [26] (version 0.9.0) with default parameters using the haplotype information from 20.80 million variants. A BigWig summary was created and a PCA was generated with the tools multiBigwigSummary and plotPCA of deepTools2 [27] (version 3.5.1), respectively.

The bovine coordinates for the bovine CpG islands (CGIs) were downloaded from the University of California Santa Cruz (UCSC) Genome Browser website [28]. Unmasked regions were included since repetitive regions may be differentially methylated [7]. The CpGs were mapped and binned into 39,376 unmasked CGIs and their methylation scores averaged with "bedtools map" [29]. A matrix with the methylation scores for all CGIs (rows) and samples (columns) was prepared. CGIs for which 25% of the samples or more had no methylation information (0 CpGs) were discarded, along with the CGIs where the methylation score variability was low among the samples (less than 1 standard deviation).

References

[1] Jana Halušková, Beáta Holečková, and Jana Staničová. DNA methylation studies in cattle. *Journal of Applied Genetics*, 62(1):121–136, February 2021. ISSN 2190-3883. doi: 10.1007/s13353-020-00604-1.

[2] Jared T. Simpson, Rachael E. Workman, P. C. Zuzarte, Matei David, L. J. Dursi, and Winston Timp. Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, 14(4):407–410, April 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4184.

[3] Rana Adnan Tahir, Da Zheng, Amina Nazir, and Hong Qing. A review of computational algorithms for CpG islands detection. *Journal of Biosciences*, 44(6):143, October 2019. ISSN 0973-7138. doi: 10.1007/s12038-019-9961-8.

[4] O. Y. Olivia Tse, Peiyong Jiang, Suk Hang Cheng, Wenlei Peng, Huimin Shang, John Wong, Stephen L. Chan, Liona C. Y. Poon, Tak Y. Leung, K. C. Allen Chan, Rossa W. K. Chiu, and Y. M. Dennis Lo. Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *Proceedings of the National Academy of Sciences*, 118(5):e2019768118, February 2021. doi: 10.1073/pnas.2019768118.

[5] Scott Gigante, Quentin Gouil, Alexis Lucattini, Andrew Keniry, Tamara Beck, Matthew Tinning, Lavinia Gordon, Chris Woodruff, Terence P. Speed, Marnie E. Blewitt, and Matthew E. Ritchie. Using long-read sequencing to detect imprinted DNA methylation. *Nucleic Acids Research*, 47(8):e46, May 2019. ISSN 1362-4962. doi: 10.1093/nar/gkz107.

[6] Zaka Wing-Sze Yuen, Akanksha Srivastava, Runa Daniel, Dennis McNevin, Cameron Jack, and Eduardo Eyras. Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. *Nature Communications*, 12(1):3438, June 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23778-6.

[7] Yang Zhou, Shuli Liu, Yan Hu, Lingzhao Fang, Yahui Gao, Han Xia, Steven G. Schroeder, Benjamin D. Rosen, Erin E. Connor, Cong-jun Li, Ransom L. Baldwin, John B. Cole, Curtis P. Van Tassell, Liguo Yang, Li Ma, and George E. Liu. Comparative whole genome DNA methylation profiling across cattle tissues reveals global and tissue-specific methylation patterns. *BMC Biology*, 18:85, July 2020. ISSN 1741-7007. doi: 10.1186/s12915-020-00793-5.

[8] Netanel Loyfer, Judith Magenheim, Ayelet Peretz, Gordon Cann, Joerg Bredno, Agnes Klochendler, Ilana Fox-Fisher, Sapir Shabi-Porat, Merav Hecht, Tsuria Pelet, Joshua Moss, Zeina Drawshy, Hamed Amini, Patriss Moradi, Sudharani Nagaraju, Dvora Bauman, David Shveiky, Shay Porat, Uri Dior, Gurion Rivkin, Omer Or, Nir Hirshoren, Einat Carmon, Alon Pikarsky, Abed Khalaileh, Gideon Zamir, Ronit Grinbaum, Machmud Abu Gazala, Ido Mizrahi, Noam Shussman, Amit Korach, Ori Wald, Uzi Izhar, Eldad Erez, Vladimir Yutkin, Yaacov Samet, Devorah Rotnemer Golinkin, Kirsty L. Spalding, Henrik Druid, Peter Arner, A. M. James Shapiro, Markus Grompe, Alex Aravanis, Oliver Venn, Arash Jamshidi, Ruth Shemer, Yuval Dor, Benjamin Glaser, and Tommy Kaplan. A DNA methylation atlas of normal human cell types. *Nature*, 613(7943):355–364, January 2023. ISSN 1476-4687. doi: 10.1038/s41586-022-05580-6.

[9] Roham Razaghi, Paul W. Hook, Shujun Ou, Michael C. Schatz, Kasper D. Hansen, Miten Jain, and Winston Timp. Modbamtools: Analysis of single-molecule epigenetic data for long-range profiling, heterogeneity, and clustering, July 2022.

[10] Mikhail Kolmogorov, Kimberley J. Billingsley, Mira Mastoras, Melissa Meredith, Jean Monlong, Ryan Lorig-Roach, Mobin Asri, Pilar Alvarez Jerez, Laksh Malik, Ramita Dewan, Xylena Reed, Rylee M. Genner, Kensuke Daida, Sairam Behera, Kishwar Shafin, Trevor Pesout, Jeshuwin Prabakaran, Paolo Carnevali, Jianzhi Yang, Arang Rhie, Sonja W. Scholz, Bryan J. Traynor, Karen H. Miga, Miten Jain, Winston Timp, Adam M. Phillippy, Mark Chaisson, Fritz J. Sedlazeck, Cornelis Blauwendraat, and Benedict Paten. Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nature Methods*, pages 1–10, September 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01993-x.

[11] Aaron M. Wenger, Paul Peluso, William J. Rowell, Pi-Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D. Olson, Armin Töpfer,

Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen-Shan Chin, Adam M. Phillippy, Michael C. Schatz, Gene Myers, Mark A. DePristo, Jue Ruan, Tobias Marschall, Fritz J. Sedlazeck, Justin M. Zook, Heng Li, Sergey Koren, Andrew Carroll, David R. Rank, and Michael W. Hunkapiller. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10):1155–1162, October 2019. ISSN 1546-1696. doi: 10.1038/s41587-019-0217-9.

[12] Anupama Jha, Stephanie C. Bohaczuk, Yizi Mao, Jane Ranchalis, Benjamin J. Mallory, Alan T. Min, Morgan O. Hamm, Elliott Swanson, Connor Finkbeiner, Tony Li, Dale Whittington, William Stafford Noble, Andrew B. Stergachis, and Mitchell R. Vollger. Fibertools: Fast and accurate DNA-m6A calling using single-molecule long-read sequencing, April 2023.

[13] Xena Marie Mapel, Naveen Kumar Kadri, Alexander S. Leonard, Qiongyu He, Audald Lloret-Villas, Meenu Bhati, Maya Hiltpold, and Hubert Pausch. Molecular quantitative trait loci in reproductive tissues impact male fertility in a large mammal, July 2023.

[14] Yong Zeng, Rahi Jain, Musaddeque Ahmed, Haiyang Guo, Yuan Zhong, Wei Xu, and Housheng Hansen He. Memo-eQTL: DNA methylation modulated genetic variant effect on gene transcriptional regulation, May 2023.

[15] Sam Kovaka, Shujun Ou, Katharine M. Jenike, and Michael C. Schatz. Approaching complete genomes, transcriptomes and epi-omes with accurate long-read sequencing. *Nature Methods*, 20(1): 12–16, January 2023. ISSN 1548-7105. doi: 10.1038/s41592-022-01716-8.

[16] Jafar Sharif and Haruhiko Koseki. Hemimethylation: DNA's lasting odd couple. *Science*, 359(6380): 1102–1103, March 2018. doi: 10.1126/science.aat0789.

[17] Ann M. Mc Cartney, Kishwar Shafin, Michael Alonge, Andrey V. Bzikadze, Giulio Formenti, Arkarachai Fungtammasan, Kerstin Howe, Chirag Jain, Sergey Koren, Glennis A. Logsdon, Karen H. Miga, Alla Mikheenko, Benedict Paten, Alaina Shumate, Daniela C. Soto, Ivan Sović, Jonathan M. D. Wood, Justin M. Zook, Adam M. Phillippy, and Arang Rhie. Chasing perfection: Validation and polishing strategies for telomere-to-telomere genome assemblies. *Nature Methods*, March 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01440-3.

[18] C.p. Prowse-Wilkins, T.j. Lopdell, R. Xiang, C.j. Vander Jagt, M.d. Littlejohn, A.j. Chamberlain, and M.e. Goddard. 552. Regulatory QTL and exon expression QTL in the mammary gland of dairy cows. In *Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP)*, chapter 552, pages 2289–2292. Wageningen Academic Publishers, December 2022. doi: 10.3920/978-90-8686-940-4_552.

[19] Heng Li. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, September 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty191.

[20] Heng Li. New strategies to improve minimap2 alignment accuracy. *Bioinformatics*, 37(23):4572–4574, December 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab705.

[21] Benjamin D. Rosen, Derek M. Bickhart, Robert D. Schnabel, Sergey Koren, Christine G. Elsik, Elizabeth Tseng, Troy N. Rowan, Wai Y. Low, Aleksey Zimin, Christine Couldrey, Richard Hall, Wenli Li, Arang Rhie, Jay Ghurye, Stephanie D. McKay, Françoise Thibaud-Nissen, Jinna Hoffman, Brenda M. Murdoch, Warren M. Snelling, Tara G. McDaneld, John A. Hammond, John C. Schwartz, Wilson Nandolo, Darren E. Hagen, Christian Dreischer, Sebastian J. Schultheiss, Steven G. Schroeder, Adam M. Phillippy, John B. Cole, Curtis P. Van Tassell, George Liu, Timothy P. L. Smith, and Juan F. Medrano. De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience*, 9(3): giaa021, March 2020. ISSN 2047-217X. doi: 10.1093/gigascience/giaa021.

[22] Audald Lloret-Villas, Meenu Bhati, Naveen Kumar Kadri, Ruedi Fries, and Hubert Pausch. Investigating the impact of reference assembly choice on genomic analyses in a cattle breed. *BMC Genomics*, 22(1):363, May 2021. ISSN 1471-2164. doi: 10.1186/s12864-021-07554-w.

[23] Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T. Afshar, Sam S. Gross, Lizzie Dorfman, Cory Y. McLean, and Mark A. DePristo. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10):983–987, November 2018. ISSN 1546-1696. doi: 10.1038/nbt.4235.

[24] Michael F. Lin, Ohad Rodeh, John Penn, Xiaodong Bai, Jeffrey G. Reid, Olga Krasheninina, and William J. Salerno. GLnexus: Joint variant calling for large cohort sequencing, June 2018.

[25] PacificBiosciences. pb-cpg-tools, 2023. URL https://github.com/PacificBiosciences/pb-CpG-tools.

[26] James M. Holt, Christopher T. Saunders, William J. Rowell, Zev Kronenberg, Aaron M. Wenger, and Michael Eberle. HiPhase: Jointly phasing small and structural variants from HiFi sequencing, May 2023.

[27] Fidel Ramírez, Devon P. Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S. Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 44(W1):W160–165, July 2016. ISSN 1362-4962. doi: 10.1093/nar/gkw257.

[28] Brian T Lee, Galt P Barber, Anna Benet-Pagès, Jonathan Casper, Hiram Clawson, Mark Diekhans, Clay Fischer, Jairo Navarro Gonzalez, Angie S Hinrichs, Christopher M Lee, Pranav Muthuraman, Luis R Nassar, Beagan Nguy, Tiana Pereira, Gerardo Perez, Brian J Raney, Kate R Rosenbloom, Daniel Schmelter, Matthew L Speir, Brittney D Wick, Ann S Zweig, David Haussler, Robert M Kuhn, Maximilian Haeussler, and W James Kent. The UCSC Genome Browser database: 2022 update. *Nucleic Acids Research*, 50(D1):D1115–D1122, October 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab959.

[29] Aaron R. Quinlan and Ira M. Hall. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq033.

Chapter 5

General discussion

Variation is paramount for genetic diversity, evolution, breeding, and any sort of genomic exploration. This thesis provides a comprehensive overview of the factors that impact the identification of DNA sequence variants in cattle and how variant genotyping can be optimised. It then shows how accurate variant discovery and refined genotyping are required for downstream analyses, allowing the identification of genomic regions involved in phenotypic traits and biological functions.

5.1 Impact of reference genome in genomic analyses

Given the importance of read alignment in variant detection, the extent to which the utilisation of different reference genomes affects both mapping and subsequent variant discovery and genotyping was investigated in Chapter 2. Overall, similar mapping performance, variant diversity (number of SNV, density of variants, variants fixed after imputation) and variant accuracy between the Hereford-based taurine reference and the haplotype-resolved Angus assemblies were observed. In addition to the analyses already presented, the accuracy of mapping short reads from BSW samples against further bovine reference genomes was also investigated (Table 5.1). Short-read WGS data from 19 animals were mapped to four different assemblies. Two of the assemblies were widely accepted bovine reference genomes: the previous version BosTau4.0 (Btau4.0) [1] and the current ARS-UCD1.2 [2]. The reads were also mapped to two haplotype-resolved reference-quality genomes that were created previously: the Angus assembly generated by *Low et al.* [3] and the BSW assembly generated by Leonard *et al.* [4]. Unsurprisingly, reference genomes from the target samples' population increased read mapping accuracy compared to assemblies from distinct breeds. There were two metrics for which the BSW reference genome was not the best assembly to map to and for which the curated ARS-UCD1.2 stood out: the proportion of autosomal high-quality reads and the percentage of multi-mapping autosomal reads.

The proportion and number of unplaced contigs (genomic regions normally formed by repetitive regions that cannot be integrated into the chromosomal structure), as well as the average number of reads spanning unplaced bases (reads not considered for downstream analyses) are also good indicators of assembly quality. Only 3% of the ARS-UCD1.2 assembly is formed by unplaced contigs, in contrast to 9% of the BSW haplotype-resolved assembly. However, the higher percentage of longer contigs (> 0.1 Mb) among these unplaced contigs in the BSW assembly, as well as a lower coverage of

Table 5.1: **Mapping stats for different bovine assemblies**. Short-reads WGS data of 19 BSW cattle were used for the evaluation. High-quality reads (HQR) are uniquely mapped and properly paired reads with quality greater than 10. The metrics considered best for each category are highlighted.

	Btau4.0	**ARS-UCD1.2**	**UOA_Angus_1**	**BSW**
Unmapped reads (%)	0.309	0.038	0.051	**0.021**
Reads mapping to autosomes (%)	78.77	82.46	81.78	**85.90**
Autosomal HQR (%)	83.81	**88.74**	88.62	85.91
Autosomal HQR per sample (M)	204.67	227.89	225.74	**229.24**
Multi-mapping autosomal reads (%)	0.57	**0.22**	0.25	0.23
Genomic % of unplaced contigs (> 0.1 Mb)	9.72 (4.71)	**3.17 (1.24)**	3.56 (1.77)	9.15 (8.39)
Unplaced contigs (> 0.1 Mb)	11,869 (494)	2,180 (112)	1,405 (127)	**949 (597)**
Mean coverage of unplaced contigs (> 0.1 Mb)	15.95x	17.76x	11.67x	**1.37x**

the sequencing reads in these regions, indicate that repetitive regions may be wrongly collapsed in the ARS-UCD1.2 assembly or wrongly expanded in the BSW assembly.

It is important to note that the breeds of all tested assemblies, along with the target sequencing data breed (BSW), belong to the taurine sub-species and diverged relatively recently. Long-reads sequencing approaches enable individual research groups to create better assemblies than huge consortia few years ago. It may be interesting to repeat our comparative study with haplotype-resolved genomes and target populations belonging to more distant breeds. It would be insightful to observe the performance of genomic analyses using farther reference genomes, especially in livestock species for which the quality of the available reference genomes is not as good as in cattle.

Applying sequencing and technological advances to the update of existing reference genomes is another conspicuous approach to improve genomic analyses. All mapping and assembly metrics explored in Table 5.1 were clearly better for ARS-UCD1.2 than the previous bovine version Btau4.0. The advent of telomere-to-telomere (T2T) assemblies [5] enables the most contiguous and comprehensive representation of the genomes to date thanks to long and accurate sequencing reads and software developments. These complete assemblies have been achieved so far for the human genome [5] and include less explored regions (often not included in the analyses for alignment limitations) such as centromeres, unplaced contigs and sex chromosomes [6, 7, 8, 9]. Similar efforts in cattle are underway. The availability of better assemblies will allow the detection of currently undetected variants, such as the ones located in alternative contigs [10] or in missing chromosomal sequences (almost 10 million bases encompassing dozens of genes are missing from the UOA_Angus_1 assembly - see Chapter 2). Overall, the combination

of T2T bovine assemblies and accurate long reads will help improve the mappability in challenging regions of the genome (such as the highly repetitive MHC complex in chromosome 23) and variant calling, both identifying novel variants as well as detecting false variants caused by the incomplete nature of previous assemblies. Given the higher quality of the newly generated reference genome, increasingly marginal gains are expected in the future. There are nevertheless still undetected variants and even such marginal gains can have implications in phenotypic traits and disease as described by Li *et al.* [10].

A primary problem to transition to an updated version of the reference genome is the strong connection to historical analyses. Combination and comparison of newly generated aligned reads with older alignments or microarrays based on previous versions require a convoluted step of remapping or translating (lifting over) the genomic coordinates, as illustrated in Fig. 5.1. More powerful tools for automatic genomic lift-over are required in order to pivot away from single reference genomes and apply population-specific approaches that leverage previous findings and resources [11].

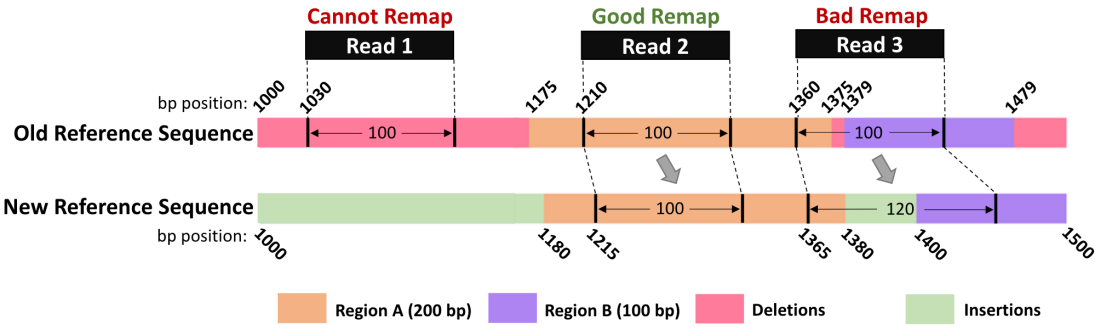


Figure 5.1: **Limitations for lifting over coordinates**. Remaping is challenged by insertions and deletions between the old and the new reference sequences - adapted from [11].

A major alternative to canonical assemblies exists for in population-specific analyses in cattle. Augmented genomes (also known as graph genomes or pangenomes) are non-linear reference sequences that contain ARS-UCD1.2 as a backbone as well as other highly contiguous bovine assemblies [12]. Bovine graph genomes aim at including the global breed diversity and thus integrate multiple haplotype assemblies that are being generated for different breeds [13]. The main application of pangenome references would be the improvement reference-based sequence mapping workflows. Not requiring the generation of breed-specific reference genomes for accurate alignment (see improvements when using the BSW assembly in Table 5.1), augmented genomes can reduce

mapping biases by including an ample range of reference genomes. Pangenomes have been described to improve unbiased variant detection, particularly in repetitive regions of the genome, and identification of novel functional sequences in cattle [4, 13, 14, 15, 16]. Pangenome approaches are novel compared to linear approaches and widespread adoption of graph-based approaches is a challenge. Continuous development and application of the relevant methods is foreseen for the coming years [17]. Provided the complete expansion of an ecosystem of tools for analyses of graph-based genome sequences and availability of diverse reference-quality haplotypes, T2T-based pangenomes can be an alternative to analyses using population-specific reference genomes.

5.2 Variant calling

The greater variant calling accuracy achieved with the combination of DeepVariant and GLnexus led to the identification of a larger set of variants in Chapter 3. These variants can be then included in comprehensive reference panels, which improve imputation, association testing and the identification of potentially causal variants [18]. Given this superior performance of the pipeline DV-GLnexus in bovine short-read data in comparison to GATK and the consistent development of the tools (*e.g.*, allowing mixed-ploidy calling), its implementation is encouraged for variant calling and genotyping efforts. But even when efficient pipelines like DV-GLnexus are used, scope for improvement is observed in variant discovery.

Limitations in variant discovery and validation

Variants have been jointly called for different cohorts through the analyses of Chapters 2, 3 and 4. The resulting callsets included between 15.95M and 28.76M autosomal SNPs and INDELs that helped quantifying the impact of the reference genome, enabled the comparison of two variant callers, formed haplotype reference panels, served as truth set for the assessment of imputation approaches, and provided the information for phasing methylated reads. And whilst most of the variants segregating in the population were captured (> 99% of the genetic variation is composed by SNPs and INDELs [19]), a large portion of genetic variation is missing from these analyses across all variant classes and genomic regions and regularly neglected in genomic analyses. Due to technical limitations, namely the prevailing use of short reads and software constrictions, genomic research in livestock has mainly focussed thus far on the discovery of autosomal SNPs. In fact, some analyses, such as lcWGS imputation with GLIMPSE [20], are further limited to only biallelic SNPs. INDELs and especially structural variants (SVs - more than 50-bp deletions, insertions, tandem duplications, inversions and translocations) are

also systematically overlooked in most of the studies analysing variant identification [21] and benchmarking genotype imputation efforts in cattle [22]. Increasingly contiguous assemblies make possible the assembly of highly repetitive and poorly assembled regions (*e.g.*, sex chromosomes), often excluded from genomic analyses [22, 23]. Accurate long reads facilitate the mapping in such regions and moving forward will enable the genotyping of undetected SNPs, INDELs and SVs, which can have phenotypically and functional importance [24, 25, 26].

The validation of the identified variants is also not free of burden. Variant validation is essential when variant calling pipelines are benchmarked or when the accuracy of variant imputation is estimated. Before sequencing technologies substantially reduced the error rates, microarray-derived genotypes were more accurate than those obtained by sequencing and have been consistently used in bovine studies to validate the variants. Indeed, microarray-called genotypes were used as a truth set in Chapter 2 to calculate non-reference sensitivity (NRS), non-reference discrepancy (NRD) and the concordance (CONC) between array-called and sequence-called genotypes. The same approach was used in Chapter 3 to evaluate the precision, recall and F1 scores for the genotypes identified with both variant callers. However, the use of microarray-derived genotypes as truth set has some limitations. They carry an overestimation of accuracy due to ascertainment bias, since the polymorphisms included in the arrays are particularly accessible [12]. The bovine genomics community would leverage a ground truth like the Genome in a Bottle (GiaB) [27] or similar datasets available in humans [28]. Such a high confidence truth-set would enable unbiased evaluation of the accuracy of genotype assignments in cattle. Analogous to Yun *et al.* [29] and Betschart *et al.* [30] in humans, the results from our studies could be corroborated on the light of such resource. Further tools for variant validation are available and can be complementary to existing approaches. Merfin [31], for instance, is a k-mer-based filtering approach that points out discrepancies between variant calls and the sequencing data. While recently implemented into the Vertebrate Genomes Project (VGP) pipeline [32], the studies from Leonard *et al.* [4] and the one described in Chapter 3 are the first to evaluate variant accuracy with Merfin using bovine sequencing data. Further studies in plant genomics applied such method thereafter [33]. Validating the variants through comparison across studies is also complicated given the different metrics (or different definitions of such metrics) used [34]. This challenge is also encountered for calculations of imputation accuracy and further discussion on such disparity of criteria is presented in the following section (5.3).

5.3 Remarks on imputation of lcWGS

Microarray chips *versus* lcWGS

Since the development of the bovine SNP genotyping assays around 15 years ago [35], the livestock breeding community has been using microarray chips for the identification of variants. Micoarrays fostered animal productivity, health, and the accuracy of selection within genetic improvement programs [35, 36]. The availability of array data for thousands of animals has allowed the imputation to sequence level and its use for genome-wide association testing. Nosková *et al.* [37] showed that methods for imputing lcWGS to sequence data in livestock are readily available with much higher variant density than microarray data (>350 times). The suitability of this approach is corroborated in cattle (Chapter 3). lcWGS thus allows the collection of orders of magnitude more data at a similar cost than microarray genotyping technologies and has been suggested as the primary data resource for imputation and downstream analyses [38]. The competitivity of low-pass data for imputation purposes has been also demonstrated commercially [39]. Even so, the implementation of lcWGS in routine livestock genotyping campaigns, genomic prediction and breeding strategies is not straightforward. Hundreds of thousands of samples are genotyped every year by breeding associations and fast turnarounds are required. Transitioning from established machinery and protocols to sequencing and computational imputation is very costly, can slow down the operational pipeline and the potential benefits may be insufficient for the breeders. High accuracy of genomic predictions is among the top interests of breeding strategies and WGS has limited potential to improve the accuracy of genomic predictions compared to marker arrays [40]. Academic goals, such as precise detection of QTL and understating of underlying biological mechanisms, require instead the maximum of variants for the maximum number of samples to be duly genotyped. Therefore, while imputation of lcWGS is a major advance for genomic studies in research environments, it might not be so beneficial for farmers and breeders.

In Chapter 3, the adequate composition of haplotype panels for an optimised imputation of low-pass data was investigated. For single-breed bovine populations it is argued that sequencing 150 animals at high coverage is sufficient to build haplotype reference panels that yield high imputation accuracies (F1 > 0.9) for lcWGS samples (0.1- to 1-fold). The generation of such a haplotype reference panel for breeds that are not mainstream might require a substantial initial investment when public data are not available. Once the haplotype reference panel is built, the sequencing of new samples at low coverage is feasible and affordable for large-scale analyses and enables powerful association studies with a lower number of animals.

Composition of haplotype reference panels

Representative haplotype reference panels are at the core of many of the methods for genotype imputation. The utilisation of multibreed haplotype reference panels for imputation purposes in livestock has been successful and encouraged over breed-specific haplotype reference panels (coinciding with the target population) for microarray data in a range of studies [22, 41, 42, 43, 44, 45]. However, as lcWGS data have been used for imputation, more accurate performances were reported for breed-specific haplotype reference panels (both in swine [45] and cattle (Chapter 3)). Despite having the larger number of variants, multi-breed haplotype panels include a reduced representation of haplotypes from the target breed and are thus less informative for the target population than same-size single-breed panels. This limited number of breed-specific haplotypes combined with the presence of haplotypes from other breeds affect threefold: the absence of breed-specific variants, the existence of out-breed variants and the presence of multiallelic variants population-wise (which are biallelic in the target breed). These conditions result into genotypes not being imputed (poorer imputation quality has been reported for underrepresented populations) [18], variants from different breeds being spuriously imputed (false positives) [44] and variants not being considered for imputation due to some imputation tools (*e.g.,* GLIMPSE) not allowing multiallelic variants. Additionally, imputation with breed-specific haplotype panels is computationally more effective. It is worth mentioning that haplotyping and the construction of haplotype reference panels will be greatly simplified by long-read data as these become increasingly available [46].

Lack of consensus in imputation metrics

Akin to the validation of variants (section 5.2), evaluation of variant accuracy after imputation is sensitive to the metrics and the ground callset used and can lead to different interpretations [18, 22]. The disparity of the parameters being used (some indicated in Fig. 1.4) causes discrepancies in how variant accuracy is evaluated and precludes an immediate comparison across similar publications. The number of metrics used in livestock studies broadly ranges from a) the correlation between imputed genotypes and observed in the truth set (r) [22, 42, 44, 47, 48], b) the correlation between true genotypes and imputed dosages [49, 50], c) the squared correlation between the true and imputed dosages (r^2) [45], d) the percentage of genotypes imputed correctly among the total imputed genotypes (concordance rate) [22, 42, 44, 45, 47], e) the concordance adjusted for chance agreement (imputation quality score, IQS) [44, 50], and f) allele concordance [47]. Rowan *et al.* [44], even described how some of the metrics (r and especially concordance rate), might be adequate for most of the variants but overestimates the accuracy of imputation for low-MAF variants. Accuracy at heterozygous sites is also suggested as a more sensitive measure than overall accuracy [51]. In Chapter 2, the variants were validated

by calculating three metrics between microarray-called (truth set) and sequence-called genotypes: non-reference sensitivity (NRS), non-reference discrepancy (NRD) and the concordance (CONC). This was aligned to previous bovine studies [12]. In Chapter 3, the metrics resulting from the haplotype VCF comparison tool "hap.py" (F1, recall and precision) were used to determine the accuracy of variants, in accordance with similar studies in humans [18, 30]. The hap.py metrics were complemented with the Pearson squared correlation between expected and actual dosages (r^2). As of today, cutting-edge studies validating genomic variation still rely on F1 scores [52] and r^2 seems to be one of the most popular metrics when reporting the accuracy in relation to the MAF [45].

The lack of a truth callset also hinders filtering from the imputed dataset. It is difficult to reach a balance between the removal of false positives (depending on quality and frequency) while keeping relevant variants important for further analyses (*e.g.*, GWAS). Some of the tools in such downstream analyses already account for the uncertainty of imputed variants by accommodating dosage scores, the probability associated to each variant of having one of the three genotypes [53]. Since this is not always the case and universal cut-offs do not exist [31], other metrics are required for an informed decision on setting a quality threshold. Imputation tools often provide themselves an internal confidence value for every imputed variant, but these are not always corresponding [54]. For instance, whilst BEAGLE provides the estimated squared correlation metric (r^2) between the estimated and the true allele dosage for the marker [55], GLIMPSE or IMPUTE2 generate an INFO score that is computed at each variant against the reference panel allele frequency and the estimated allele frequency and obtained from the genotype dosages [56]. In Chapter 3, the imputation accuracy against a truth set (F1) is related to the GLIMPSE INFO score to interpret the relationship between both. This connection highlights the careful usage of the estimated accuracy metrics provided by imputation tools to set reasonable cut-offs for downstream analyses without unwillingly remove low-frequency variants.

5.4 Relevance of functional annotation

The assignment of functional properties to each region of the genome is essential for the interpretation of biological implications of polymorphic loci. Provided a complete and correct annotation file, usually shared alongside the genome coordinates of high-quality assemblies, the impact of the variants can be anticipated and further functionally validated. However, confidently predicting the functional annotation to the different regions is challenging and laborious [10] and newly assembled reference genome usually lack

associated annotations. Despite the existence of more automated tools (such as protein-to-genome aligners, albeit limited to coding regions [57]), manual curation within fully-fledged annotation pipelines is still required to obtain reliable annotation sets [58, 59]. In fact, one of the most relevant differences detected when comparing the use of reference genomes in Chapter 2 were the predicted location of the variants. While the percentage in coding/exons regions was similar, the percentage of variants annotated to intergenic regions was ∼10 percent points higher in ARS-UCD1.2 and the percentage of variants annotated to intronic regions was ∼10 percent points higher in UOA_Angus_1. Since the manual annotation performed in ARS-UCD1.2 increases its reliability, the extra number of potentially high impact variants found with the UOA_Angus_1 assembly may be unreliable, and further validation is warranted. Even if new bovine reference-quality assemblies are equally or more contiguous and accurate than existing ones, the errors and gaps in the functional annotations are a major obstacle to switch references [60]. Annotations are also crucial for the benchmarking of variant callers. Thanks to the comprehensive annotation of the ARS-UCD1.2 reference genome, the detection of biologically important variants (often disease-causing or involved in production traits) can be duly compared across pipelines and tools. The findings from other studies [61] as GATK misses more functionally relevant variants that can have large impact in bovine populations were confirmed in Chapter 3. Also because the reliability of the existing bovine annotations, the coordinates of CpG islands could be retrieved for the study of methylation patterns in bull samples in Chapter 4.

5.5 The role of genetic variation in other omics data

Methylation variability among individuals is influenced by genetic variation [62]. To fully understand how genetic variation may affect phenotypes via the methylome, it is necessary to profile methylation states across the genome and see how and where they vary across animals and breeds [63]. Profiting from a uniquely large dataset of bovine long-reads with haplotype-phased methylation information from different tissues and development stages, Chapter 4 paves the way for a comprehensive exploration of methylation patterns in CpG islands. The identification of the loci that impact methylation (meQTLs) is a promising further exploration that can facilitate uncovering genetic regulatory mechanisms. Furthermore, the availability of long-read RNA-seq data as well as the new bio-computational frameworks to detect histone and base modifications other than 5mC (*e.g.*, m6A- and 2'-O-methylation) in long reads [64, 65, 66] shed new light on the intertwined multi-omic relationship between genomic variation, epigenetic modifications, histone accessibility and gene expression.

5.6 Good praxis in computational biology

Bioinformaticians and computational biologists have an important responsibility in the development and implementation of ethical and responsible research, in terms of data and code openness as well as environmental impact. I have implemented reproducible and optimised pipelines during my doctorate to comply with the FAIR principles (Findability, Accessibility, Interoperability, Reusability) [67] and reduce the electrical power implications of my analyses. For the sake of consistency and performance optimisation, given the multiple times read-alignment has been performed across the different analyses ($\sim$ 600 samples), a reproducible alignment pipeline was created [68]. While more customisable and containerised solutions may exist in the meantime, such an automatised pipeline was not openly available when this project started and has proven to be an important resource for a range of genomic analyses. Several publications leveraging this computational workflow have been since published [69, 70, 71, 72, 73].

Additionally, benchmarking of software and tools has been continuously performed in order not only to keep up with the rapid pace of bioinformatic development but also to optimise and update the pipelines. This complies with the eighth recommendation of Grealey *et al.* [74], according to which using the most up to date software is "the quickest, easiest, and potentially most impactful way to reduce one's carbon footprint". Newer versions of the tool BWA-MEM2 [75], for instance, heavily reduced the memory required for read alignment [76]. Utilisation of the workflow manager Snakemake [77] and the multi-thread features, available for most of the software, increased the parallelisation of analyses and allowed scaling these up to larger cohorts. Also, as indicated in Chapter 3, the utilisation of single-breed haplotype panels for the imputation of lcWGS is more computationally efficient than multi-breed panels. All software and versions that have been used and the pipelines that have been developed for the preparation of this thesis are openly accessible from the Animal Genomics Github repository [78].

References

[1] Christine G. Elsik, Ross L. Tellam, Kim C. Worley, and The Bovine. The Genome Sequence of Taurine Cattle: A window to ruminant biology and evolution. *Science (New York, N.Y.)*, 324(5926):522–528, April 2009. ISSN 0036-8075. doi: 10.1126/science.1169588.

[2] Benjamin D. Rosen, Derek M. Bickhart, Robert D. Schnabel, Sergey Koren, Christine G. Elsik, Elizabeth Tseng, Troy N. Rowan, Wai Y. Low, Aleksey Zimin, Christine Couldrey, Richard Hall, Wenli Li, Arang Rhie, Jay Ghurye, Stephanie D. McKay, Françoise Thibaud-Nissen, Jinna Hoffman, Brenda M. Murdoch, Warren M. Snelling, Tara G. McDaneld, John A. Hammond, John C. Schwartz, Wilson Nandolo, Darren E. Hagen, Christian Dreischer, Sebastian J. Schultheiss, Steven G. Schroeder, Adam M.

Phillippy, John B. Cole, Curtis P. Van Tassell, George Liu, Timothy P. L. Smith, and Juan F. Medrano. De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience*, 9(3): giaa021, March 2020. ISSN 2047-217X. doi: 10.1093/gigascience/giaa021.

[3] Wai Yee Low, Rick Tearle, Ruijie Liu, Sergey Koren, Arang Rhie, Derek M. Bickhart, Benjamin D. Rosen, Zev N. Kronenberg, Sarah B. Kingan, Elizabeth Tseng, Françoise Thibaud-Nissen, Fergal J. Martin, Konstantinos Billis, Jay Ghurye, Alex R. Hastie, Joyce Lee, Andy W. C. Pang, Michael P. Heaton, Adam M. Phillippy, Stefan Hiendleder, Timothy P. L. Smith, and John L. Williams. Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. *Nature Communications*, 11(1):2071, April 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-15848-y.

[4] Alexander S. Leonard, Danang Crysnanto, Zih-Hua Fang, Michael P. Heaton, Brian L. Vander Ley, Carolina Herrera, Heinrich Bollwein, Derek M. Bickhart, Kristen L. Kuhn, Timothy P. L. Smith, Benjamin D. Rosen, and Hubert Pausch. Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. *Nature Communications*, 13(1):3012, May 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-30680-2.

[5] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, et al. The complete sequence of a human genome. *Science (New York, N.Y.)*, 376(6588):44–53, April 2022. ISSN 1095-9203. doi: 10.1126/science.abj6987.

[6] Miten Jain, Hugh E. Olsen, Daniel J. Turner, David Stoddart, Kira V. Bulazel, Benedict Paten, David Haussler, Huntington F. Willard, Mark Akeson, and Karen H. Miga. Linear assembly of a human centromere on the Y chromosome. *Nature Biotechnology*, 36(4):321–323, April 2018. ISSN 1546-1696. doi: 10.1038/nbt.4109.

[7] Karen H. Miga, Sergey Koren, Arang Rhie, Mitchell R. Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, Edmund Howe, David Porubsky, Glennis A. Logsdon, Valerie A. Schneider, Tamara Potapova, Jonathan Wood, William Chow, Joel Armstrong, Jeanne Fredrickson, Evgenia Pak, et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, 585(7823):79–84, September 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2547-7.

[8] Pille Hallast, Peter Ebert, Mark Loftus, Feyza Yilmaz, Peter A. Audano, Glennis A. Logsdon, Marc Jan Bonder, Weichen Zhou, Wolfram Höps, Kwondo Kim, Chong Li, Savannah J. Hoyt, Philip C. Dishuck, David Porubsky, Fotios Tsetsos, Jee Young Kwon, Qihui Zhu, Katherine M. Munson, Patrick Hasenfeld, William T. Harvey, Alexandra P. Lewis, Jennifer Kordosky, Kendra Hoekzema, Rachel J. O'Neill, Jan O. Korbel, Chris Tyler-Smith, Evan E. Eichler, Xinghua Shi, Christine R. Beck, Tobias Marschall, Miriam K. Konkel, and Charles Lee. Assembly of 43 human Y chromosomes reveals extensive complexity and variation. *Nature*, pages 1–10, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06425-6.

[9] Arang Rhie, Sergey Nurk, Monika Cechova, Savannah J. Hoyt, Dylan J. Taylor, Nicolas Altemose, Paul W. Hook, Sergey Koren, Mikko Rautiainen, Ivan A. Alexandrov, Jamie Allen, Mobin Asri, Andrey V. Bzikadze, Nae-Chyun Chen, Chen-Shan Chin, Mark Diekhans, Paul Flicek, Giulio Formenti, et al. The complete sequence of a human Y chromosome. *Nature*, pages 1–11, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06457-y.

[10] He Li, Moez Dawood, Michael M. Khayat, Jesse R. Farek, Shalini N. Jhangiani, Ziad M. Khan, Tadahiro Mitani, Zeynep Coban-Akdemir, James R. Lupski, Eric Venner, Jennifer E. Posey, Aniko Sabo, and Richard A. Gibbs. Exome variant discrepancies due to reference-genome differences. *American Journal of Human Genetics*, 108(7):1239–1250, July 2021. ISSN 1537-6605. doi: 10.1016/j.ajhg.2021.05.011.

[11] Jeremie S. Kim, Can Firtina, Meryem Banu Cavlak, Damla Senol Cali, Nastaran Hajinazar, Mohammed Alser, Can Alkan, and Onur Mutlu. AirLift: A Fast and Comprehensive Technique for Remapping Alignments between Reference Genomes, February 2021.

[12] Danang Crysnanto, Christine Wurmser, and Hubert Pausch. Accurate sequence variant genotyping in cattle using variation-aware genome graphs. *Genetics, selection, evolution: GSE*, 51(1):21, May 2019. ISSN 1297-9686. doi: 10.1186/s12711-019-0462-x.

[13] A. Talenti, J. Powell, J. D. Hemmink, E. a. J. Cook, D. Wragg, S. Jayaraman, E. Paxton, C. Ezeasor, E. T. Obishakin, E. R. Agusi, A. Tijjani, W. Amanyire, D. Muhanguzi, K. Marshall, A. Fisch, B. R. Ferreira, A. Qasim, U. Chaudhry, P. Wiener, P. Toye, L. J. Morrison, T. Connelley, and J. G. D. Prendergast. A cattle graph genome incorporating global breed diversity. *Nature Communications*, 13(1):910, February 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-28605-0.

[14] Danang Crysnanto, Alexander S. Leonard, Zih-Hua Fang, and Hubert Pausch. Novel functional sequences uncovered through a bovine multiassembly graph. *Proceedings of the National Academy of Sciences of the United States of America*, 118(20):e2101056118, May 2021. ISSN 1091-6490. doi: 10.1073/pnas.2101056118.

[15] Alexander S. Leonard, Danang Crysnanto, Xena M. Mapel, Meenu Bhati, and Hubert Pausch. Graph construction method impacts variation representation and analyses in a bovine super-pangenome. *Genome Biology*, 24(1):124, May 2023. ISSN 1474-760X. doi: 10.1186/s13059-023-02969-y.

[16] Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K. Lucas, Jean Monlong, Haley J. Abel, Silvia Buonaiuto, Xian H. Chang, Haoyu Cheng, Justin Chu, Vincenza Colonna, Jordan M. Eizenga, Xiaowen Feng, Christian Fischer, Robert S. Fulton, Shilpa Garg, Cristian Groza, Andrea Guarracino, William T. Harvey, et al. A draft human pangenome reference. *Nature*, 617(7960):312–324, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-05896-x.

[17] Ting Wang, Lucinda Antonacci-Fulton, Kerstin Howe, Heather A. Lawson, Julian K. Lucas, Adam M. Phillippy, Alice B. Popejoy, Mobin Asri, Caryn Carson, Mark J. P. Chaisson, Xian Chang, Robert Cook-Deegan, Adam L. Felsenfeld, Robert S. Fulton, Erik P. Garrison, Nanibaa' A. Garrison, Tina A. Graves-Lindsay, Hanlee Ji, Eimear E. Kenny, Barbara A. Koenig, Daofeng Li, Tobias Marschall, Joshua F. McMichael, Adam M. Novak, Deepak Purushotham, Valerie A. Schneider, Baergen I. Schultz, Michael W. Smith, Heidi J. Sofia, Tsachy Weissman, Paul Flicek, Heng Li, Karen H. Miga, Benedict Paten, Erich D. Jarvis, Ira M. Hall, Evan E. Eichler, and David Haussler. The Human Pangenome Project: A global resource to map genomic diversity. *Nature*, 604(7906):437–446, April 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04601-8.

[18] Jared O'Connell, Taedong Yun, Meghan Moreno, Helen Li, Nadia Litterman, Alexey Kolesnikov, Elizabeth Noblin, Pi-Chuan Chang, Anjali Shastri, Elizabeth H. Dorfman, Suyash Shringarpure, Adam Auton, Andrew Carroll, and Cory Y. McLean. A population-specific reference panel for improved genotype imputation in African Americans. *Communications Biology*, 4:1269, November 2021. ISSN 2399-3642. doi: 10.1038/s42003-021-02777-9.

[19] 1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, Jonathan L. Marchini, Shane McCarthy, Gil A. McVean, and Gonçalo R. Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015. ISSN 1476-4687. doi: 10.1038/nature15393.

[20] Simone Rubinacci, Diogo M. Ribeiro, Robin J. Hofmeister, and Olivier Delaneau. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nature Genetics*, 53(1): 120–126, January 2021. ISSN 1546-1718. doi: 10.1038/s41588-020-00756-0.

[21] Bohu Pan, Rebecca Kusko, Wenming Xiao, Yuanting Zheng, Zhichao Liu, Chunlin Xiao, Sugunadevi Sakkiah, Wenjing Guo, Ping Gong, Chaoyang Zhang, Weigong Ge, Leming Shi, Weida Tong, and Huixiao Hong. Similarities and differences between variants called with human reference genome HG19 or HG38. *BMC Bioinformatics*, 20(Suppl 2):101, March 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-2620-0.

[22] Muhammad Yasir Nawaz, Priscila Arrigucci Bernardes, Rodrigo Pelicioni Savegnago, Dajeong Lim, Seung Hwan Lee, and Cedric Gondro. Evaluation of Whole-Genome Sequence Imputation Strategies in Korean Hanwoo Cattle. *Animals*, 12(17):2265, January 2022. ISSN 2076-2615. doi: 10.3390/ani12172265.

[23] Ruijie Liu, Wai Yee Low, Rick Tearle, Sergey Koren, Jay Ghurye, Arang Rhie, Adam M. Phillippy, Benjamin D. Rosen, Derek M. Bickhart, Timothy P. L. Smith, Stefan Hiendleder, and John L. Williams. New insights into mammalian sex chromosome structure and evolution using high-quality sequences from bovine X and Y chromosomes. *BMC Genomics*, 20(1):1000, December 2019. ISSN 1471-2164. doi: 10.1186/s12864-019-6364-z.

[24] Sen Zhao, Oleg Agafonov, Abdulrahman Azab, Tomasz Stokowy, and Eivind Hovig. Accuracy and efficiency of germline variant calling pipelines for human genome data. *Scientific Reports*, 10(1):20222, November 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-77218-4.

[25] Maya Hiltpold, Naveen Kumar Kadri, Fredi Janett, Ulrich Witschi, Fritz Schmitz-Hsu, and Hubert Pausch. Autosomal recessive loci contribute significantly to quantitative variation of male fertility in a dairy cattle population. *BMC genomics*, 22(1):225, March 2021. ISSN 1471-2164. doi: 10.1186/s12864-021-07523-3.

[26] Marie-Pierre Sanchez, Clémentine Escouflaire, Aurélia Baur, Fiona Bottin, Chris Hozé, Mekki Boussaha, Sébastien Fritz, Aurélien Capitan, and Didier Boichard. X-linked genes influence various complex traits in dairy cattle. *BMC Genomics*, 24(1):338, June 2023. ISSN 1471-2164. doi: 10.1186/s12864-023-09438-7.

[27] Justin M. Zook, Jennifer McDaniel, Nathan D. Olson, Justin Wagner, Hemang Parikh, Haynes Heaton, Sean A. Irvine, Len Trigg, Rebecca Truty, Cory Y. McLean, Francisco M. De La Vega, Chunlin Xiao, Stephen Sherry, and Marc Salit. An open resource for accurately benchmarking small variant and reference calls. *Nature Biotechnology*, 37(5):561–566, May 2019. ISSN 1546-1696. doi: 10.1038/s41587-019-0074-6.

[28] Sina Majidian, Daniel Paiva Agustinho, Chen-Shan Chin, Fritz J. Sedlazeck, and Medhat Mahmoud. Genomic variant benchmark: If you cannot measure it, you cannot improve it. *Genome Biology*, 24(1):221, October 2023. ISSN 1474-760X. doi: 10.1186/s13059-023-03061-1.

[29] Taedong Yun, Helen Li, Pi-Chuan Chang, Michael F. Lin, Andrew Carroll, and Cory Y. McLean. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics (Oxford, England)*, page btaa1081, January 2021. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaa1081.

[30] Raphael O. Betschart, Alexandre Thiéry, Domingo Aguilera-Garcia, Martin Zoche, Holger Moch, Raphael Twerenbold, Tanja Zeller, Stefan Blankenberg, and Andreas Ziegler. Comparison of calling pipelines for whole genome sequencing: An empirical study demonstrating the importance of mapping and alignment. *Scientific Reports*, 12(1):21502, December 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-26181-3.

[31] Giulio Formenti, Arang Rhie, Brian P. Walenz, Françoise Thibaud-Nissen, Kishwar Shafin, Sergey Koren, Eugene W. Myers, Erich D. Jarvis, and Adam M. Phillippy. Merfin: Improved variant filtering, assembly evaluation and polishing via k-mer validation. *Nature Methods*, pages 1–9, March 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01445-y.

[32] Arang Rhie, Shane A. McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, William Chow, Arkarachai Fungtammasan, Juwan Kim, Chul Lee, Byung June Ko, Mark Chaisson, Gregory L. Gedman, Lindsey J. Cantin, Francoise Thibaud-Nissen, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, 592(7856):737–746, April 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03451-0.

[33] Hiroyuki Tanaka, Tatsuki Hori, Shohei Yamamoto, Atsushi Toyoda, Kentaro Yano, Kyoko Yamane, and Takehiko Itoh. Haplotype-resolved chromosomal-level assembly of wasabi (Eutrema japonicum) genome. *Scientific Data*, 10(1):441, July 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02356-z.

[34] Peter Krusche, Len Trigg, Paul C. Boutros, Christopher E. Mason, Francisco M. De La Vega, Benjamin L. Moore, Mar Gonzalez-Porta, Michael A. Eberle, Zivana Tezak, Samir Lababidi, Rebecca Truty, George Asimenos, Birgit Funke, Mark Fleharty, Brad A. Chapman, Marc Salit, Justin M. Zook, and Global Alliance for Genomics and Health Benchmarking Team. Best practices for benchmarking germline small-variant calls in human genomes. *Nature Biotechnology*, 37(5):555–560, May 2019. ISSN 1546-1696. doi: 10.1038/s41587-019-0054-x.

[35] Lakshmi K. Matukumalli, Cynthia T. Lawley, Robert D. Schnabel, Jeremy F. Taylor, Mark F. Allan, Michael P. Heaton, Jeff O'Connell, Stephen S. Moore, Timothy P. L. Smith, Tad S. Sonstegard, and Curtis P. Van Tassell. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. *PLOS ONE*, 4(4):e5350, April 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0005350.

[36] Didier Boichard, Hoyoung Chung, Romain Dassonneville, Xavier David, André Eggen, Sébastien Fritz, Kimberly J. Gietzen, Ben J. Hayes, Cynthia T. Lawley, Tad S. Sonstegard, Curtis P. Van Tassell, Paul M. VanRaden, Karine A. Viaud-Martinez, George R. Wiggans, and for the Bovine LD Consortium. Design of a Bovine Low-Density SNP Array Optimized for Imputation. *PLOS ONE*, 7(3):e34130, March 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0034130.

[37] Adéla Nosková, Meenu Bhati, Naveen Kumar Kadri, Danang Crysnanto, Stefan Neuenschwander, Andreas Hofer, and Hubert Pausch. Characterization of a haplotype-reference panel for genotyping by low-pass sequencing in Swiss Large White pigs. *BMC genomics*, 22(1):290, April 2021. ISSN 1471-2164. doi: 10.1186/s12864-021-07610-5.

[38] Robert W. Davies, Marek Kucka, Dingwen Su, Sinan Shi, Maeve Flanagan, Christopher M. Cunniff, Yingguang Frank Chan, and Simon Myers. Rapid genotype imputation from sequence with reference panels. *Nature Genetics*, 53(7):1104–1111, July 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00877-0.

[39] Pickrell J. A platform for replacing genotyping arrays with sequencing, 2018. URL https://medium.com/the-gencove-blog/a-platform-for-replacing-genotyping-arrays-with-sequencing-cf570937968c.

[40] Roger Ros-Freixedes, Martin Johnsson, Andrew Whalen, Ching-Yi Chen, Bruno D. Valente, William O. Herring, Gregor Gorjanc, and John M. Hickey. Genomic prediction with whole-genome sequence data in intensely selected pig lines. *Genetics Selection Evolution*, 54(1):65, September 2022. ISSN 1297-9686. doi: 10.1186/s12711-022-00756-0.

[41] Rasmus Froberg Brøndum, Bernt Guldbrandtsen, Goutam Sahana, Mogens Sandø Lund, and Guosheng Su. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. *BMC genomics*, 15:728, August 2014. ISSN 1471-2164. doi: 10.1186/1471-2164-15-728.

[42] Marie-Pierre Sanchez, Armelle Govignon-Gion, Pascal Croiseau, Sébastien Fritz, Chris Hozé, Guy Miranda, Patrice Martin, Anne Barbat-Leterrier, Rabia Letaïef, Dominique Rocha, Mickaël Brochard, Mekki Boussaha, and Didier Boichard. Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. *Genetics, selection, evolution: GSE*, 49(1):68, September 2017. ISSN 1297-9686. doi: 10.1186/s12711-017-0344-z.

[43] Hubert Pausch, Iona M. MacLeod, Ruedi Fries, Reiner Emmerling, Phil J. Bowman, Hans D. Daetwyler, and Michael E. Goddard. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. *Genetics, selection, evolution: GSE*, 49(1):24, February 2017. ISSN 1297-9686. doi: 10.1186/s12711-017-0301-x.

[44] Troy N. Rowan, Jesse L. Hoff, Tamar E. Crum, Jeremy F. Taylor, Robert D. Schnabel, and Jared E. Decker. A multi-breed reference panel and additional rare variants maximize imputation accuracy in cattle. *Genetics Selection Evolution*, 51(1):77, December 2019. ISSN 1297-9686. doi: 10.1186/s12711-019-0519-x.

[45] Zhen Wang, Zhenyang Zhang, Zitao Chen, Jiabao Sun, Caiyun Cao, Fen Wu, Zhong Xu, Wei Zhao, Hao Sun, Longyu Guo, Zhe Zhang, Qishan Wang, and Yuchun Pan. PHARP: A pig haplotype reference panel for genotype imputation. *Scientific Reports*, 12(1):12645, July 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-15851-x.

[46] Runyang Nicolas Lou, Arne Jacobs, Aryn P. Wilder, and Nina Overgaard Therkildsen. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology*, 30(23):5966–5993, 2021. ISSN 1365-294X. doi: 10.1111/mec.16077.

[47] Aine C. O'Brien, Michelle M. Judge, Sean Fair, and Donagh P. Berry. High imputation accuracy from informative low-to-medium density single nucleotide polymorphism genotypes is achievable in sheep. *Journal of Animal Science*, 97(4):1550–1567, April 2019. ISSN 1525-3163. doi: 10.1093/jas/skz043.

[48] Warren M. Snelling, Jesse L. Hoff, Jeremiah H. Li, Larry A. Kuehn, Brittney N. Keel, Amanda K. Lindholm-Perry, and Joseph K. Pickrell. Assessment of Imputation from Low-Pass Sequencing to Predict Merit of Beef Steers. *Genes*, 11(11):E1312, November 2020. ISSN 2073-4425. doi: 10.3390/genes11111312.

[49] Roger Ros-Freixedes, Andrew Whalen, Ching-Yi Chen, Gregor Gorjanc, William O. Herring, Alan J. Mileham, and John M. Hickey. Accuracy of whole-genome sequence imputation using hybrid peeling in large pedigreed livestock populations. *Genetics, selection, evolution: GSE*, 52(1):17, April 2020. ISSN 1297-9686. doi: 10.1186/s12711-020-00536-8.

[50] Jun Teng, Changheng Zhao, Dan Wang, Zhi Chen, Hui Tang, Jianbin Li, Cheng Mei, Zhangping Yang, Chao Ning, and Qin Zhang. Assessment of the performance of different imputation methods for low-coverage sequencing in Holstein cattle. *Journal of Dairy Science*, 105(4):3355–3366, April 2022. ISSN 0022-0302. doi: 10.3168/jds.2021-21360.

[51] Richard M. Durbin, David Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Francis S. Collins, Francisco M. De La Vega, Peter Donnelly, Michael Egholm, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Bartha M. Knoppers, Eric S. Lander, Hans Lehrach, Elaine R. Mardis, Gil A. McVean, et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, October 2010. ISSN 1476-4687. doi: 10.1038/nature09534.

[52] Alexey Kolesnikov, Daniel E. Cook, Maria Nattestad, Euan A. Ashley, John Gorzynski, Sneha D. Goenka, Miten Jain, Brandy McNulty, Karen H. Miga, Benedict Paten, Pi-Chuan Chang, Andrew Carroll, and Kishwar Shafin. Local read haplotagging enables accurate long-read small variant calling, September 2023.

[53] Sanne van den Berg, Jérémie Vandenplas, Fred A. van Eeuwijk, Aniek C. Bouwman, Marcos S. Lopes, and Roel F. Veerkamp. Imputation to whole-genome sequence using multiple pig populations and its use in genome-wide association studies. *Genetics, Selection, Evolution : GSE*, 51:2, January 2019. ISSN 0999-193X. doi: 10.1186/s12711-019-0445-y.

[54] Gennady V. Khvorykh and Andrey V. Khrunin. Imputeqc: An R package for assessing imputation quality of genotypes and optimizing imputation parameters. *BMC Bioinformatics*, 21(12):304, July 2020. ISSN 1471-2105. doi: 10.1186/s12859-020-03589-0.

[55] Brian Browning. Beagle 5.4 specifications, 2022. URL https://faculty.washington.edu/browning/beagle/beagle_5.4_18Mar22.pdf.

[56] Oliver Delaneau. Glimpse: A tool for low-coverage whole-genome sequencing imputation, 2023. URL https://odelaneau.github.io/GLIMPSE/docs/tutorials/getting_started/#7-imputation-accuracy.

[57] Heng Li. Protein-to-genome alignment with miniprot. *Bioinformatics*, page btad014, January 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad014.

[58] Tomáš Brůna, Heng Li, Joseph Guhlin, Daniel Honsel, Steffen Herbold, Mario Stanke, Natalia Nenasheva, Matthis Ebel, Lars Gabriel, and Katharina J. Hoff. Galba: Genome annotation with miniprot and AUGUSTUS. *BMC Bioinformatics*, 24(1):327, August 2023. ISSN 1471-2105. doi: 10.1186/s12859-023-05449-z.

[59] Lars Gabriel, Tomas Bruna, Katharina Jasmin Hoff, Matthis Ebel, Alexandre Lomsadze, Mark Borodovsky, and Mario Stanke. BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA, September 2023.

[60] Yan Guo, Yulin Dai, Hui Yu, Shilin Zhao, David C. Samuels, and Yu Shyr. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, 109(2):83–90, March 2017. ISSN 0888-7543. doi: 10.1016/j.ygeno.2017.01.005.

[61] Yi-Lin Lin, Pi-Chuan Chang, Ching Hsu, Miao-Zi Hung, Yin-Hsiu Chien, Wuh-Liang Hwu, FeiPei Lai, and Ni-Chung Lee. Comparison of GATK and DeepVariant by trio sequencing. *Scientific Reports*, 12:1809, February 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-05833-4.

[62] Meritxell Oliva, Kathryn Demanelis, Yihao Lu, Meytal Chernoff, Farzana Jasmine, Habibul Ahsan, Muhammad G. Kibriya, Lin S. Chen, and Brandon L. Pierce. DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nature Genetics*, 55(1):112–122, January 2023. ISSN 1546-1718. doi: 10.1038/s41588-022-01248-z.

[63] Jessica Powell, Andrea Talenti, Andressa Fisch, Johanneke D. Hemmink, Edith Paxton, Philip Toye, Isabel Santos, Beatriz R. Ferreira, Tim K. Connelley, Liam J. Morrison, and James G. D. Prendergast. Profiling the immune epigenome across global cattle breeds. *Genome Biology*, 24(1):127, May 2023. ISSN 1474-760X. doi: 10.1186/s13059-023-02964-3.

[64] Anupama Jha, Stephanie C. Bohaczuk, Yizi Mao, Jane Ranchalis, Benjamin J. Mallory, Alan T. Min, Morgan O. Hamm, Elliott Swanson, Connor Finkbeiner, Tony Li, Dale Whittington, William Stafford Noble, Andrew B. Stergachis, and Mitchell R. Vollger. Fibertools: Fast and accurate DNA-m6A calling using single-molecule long-read sequencing, April 2023.

[65] Doaa Hassan Salem, Aditya Ariyur, Swapna Vidhur Daulatabad, Quoseena Mir, and Sarath Chandra Janga. Transcriptome-wide single molecule mapping of 2´-O-Methylation (Nm) sites in Nanopore direct RNA sequencing datasets using the Nm-nano framework, February 2023.

[66] Mitchell R. Vollger, Jonas Korlach, Kiara C. Eldred, Elliott Swanson, Jason G. Underwood, Katherine M. Munson, Yong-Han H. Cheng, Jane Ranchalis, Yizi Mao, Elizabeth E. Blue, Ulrike Schwarze, Christopher T. Saunders, Aaron M. Wenger, Aimee Allworth, Sirisak Chanprasert, Brittney L. Duerden, Ian Glass, Martha Horike-Pyne, Michelle Kim, Kathleen A. Leppig, Ian J. McLaughlin, Jessica Ogawa, Elisabeth A. Rosenthal, Sam Sheppeard, Stephanie M. Sherman, Samuel Strohbehn, Amy L. Yuen, Thomas A. Reh, Peter H. Byers, Michael J. Bamshad, Fuki M. Hisama, Gail P. Jarvik, Yasemin Sancak, Katrina M. Dipple, and Andrew B. Stergachis. Synchronized long-read genome, methylome, epigenome, and transcriptome for resolving a Mendelian condition, September 2023.

[67] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, March 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.18.

[68] Audald Lloret-Villas. Animal genomics eth github: Alignment and mapping statistics, 2021. URL https://github.com/AnimalGenomicsETH/Reference_assembly_choice/tree/master/Alignment.

[69] Meenu Bhati, Xena Marie Mapel, Audald Lloret-Villas, and Hubert Pausch. Structural variants and short tandem repeats impact gene expression and splicing in bovine testis tissue. *Genetics*, page iyad161, September 2023. ISSN 1943-2631. doi: 10.1093/genetics/iyad161.

[70] Xena Marie Mapel, Naveen Kumar Kadri, Alexander S. Leonard, Qiongyu He, Audald Lloret-Villas, Meenu Bhati, Maya Hiltpold, and Hubert Pausch. Molecular quantitative trait loci in reproductive tissues impact male fertility in a large mammal, July 2023.

[71] A. Nosková, A. Mehrotra, N. K. Kadri, A. Lloret-Villas, S. Neuenschwander, A. Hofer, and H. Pausch. Comparison of two multi-trait association testing methods and sequence-based fine mapping of six additive QTL in Swiss Large White pigs. *BMC Genomics*, 24(1):192, April 2023. ISSN 1471-2164. doi: 10.1186/s12864-023-09295-4.

[72] Elena O'Callaghan, Paula Navarrete-Lopez, Miriama Štiavnická, José M. Sánchez, Maria Maroto, Eva Pericuesta, Raul Fernández-González, Ciara O'Meara, Bernard Eivers, Margaret M. Kelleher, Ross D. Evans, Xena M. Mapel, Audald Lloret-Villas, Hubert Pausch, Miriam Balastegui-Alarcón, Manuel Avilés, Ana Sanchez-Rodriguez, Eduardo R. S. Roldan, Michael McDonald, David A. Kenny, Sean Fair, Alfonso Gutiérrez-Adán, and Patrick Lonergan. Adenylate kinase 9 is essential for sperm function and male fertility in mammals. *Proceedings of the National Academy of Sciences*, 120(42):e2305712120, October 2023. doi: 10.1073/pnas.2305712120.

[73] Esther Ewaoluwagbemiga, Audald Lloret-Villas, Adéla Nosková, Hubert Pausch, and Claudia Kasper. Genome-wide association study and regional heritability mapping of protein efficiency and performance traits in swiss large white pigs, in prep.

[74] Jason Grealey, Loïc Lannelongue, Woei-Yuh Saw, Jonathan Marten, Guillaume Méric, Sergio Ruiz-Carmona, and Michael Inouye. The Carbon Footprint of Bioinformatics. *Molecular Biology and Evolution*, 39(3):msac034, March 2022. ISSN 1537-1719. doi: 10.1093/molbev/msac034.

[75] Vasimuddin Md, Sanchit Misra, Heng Li, and Srinivas Aluru. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. *arXiv:1907.12931 [cs, q-bio]*, July 2019.

[76] bwa mem2. bwa-mem2 github, 2020. URL https://github.com/bwa-mem2/bwa-mem2.

[77] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O. Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. Sustainable data analysis with Snakemake. *F1000Research*, 10:33, 2021. ISSN 2046-1402. doi: 10.12688/f1000research.29032.2.

[78] Animal Genomics ETH. Eth animal genomics group github, 2023. URL https://github.com/AnimalGenomicsETH.

Supplementary Material
Chapter 2

Additional file 1
Sample IDs

BSW cattle IDs. Accession IDs of the 161 bovine samples used for our study.

SAMEA5159792	SAMEA5159791	SAMEA5159788	SAMEA5159783	SAMEA5159785	SAMEA5159799	SAMEA5159787
SAMEA5159761	SAMEA5159782	SAMEA5159775	SAMEA5159786	SAMEA5159784	SAMEA5159798	SAMEA5159781
SAMEA5159780	SAMEA5159797	SAMEA5159769	SAMEA5159778	SAMEA5159771	SAMEA5159779	SAMEA5159772
SAMEA5159773	SAMEA5159793	SAMEA5159770	SAMEA4644727	SAMEA4644728	SAMEA4644765	SAMEA4644766
SAMEA4644769	SAMEA19314418	SAMEA19315168	SAMEA4644754	SAMEA4644755	SAMEA4644756	SAMEA4644735
SAMEA4644757	SAMEA4644739	SAMEA4644741	SAMEA4644742	SAMEA4644758	SAMEA4644743	SAMEA4644762
SAMEA4644763	SAMEA6163199	SAMEA6272095	SAMEA6272096	SAMEA6272098	SAMEA6272099	SAMEA6272104
SAMEA6272105	SAMEA6272106	SAMEA6272108	SAMEA6272111	SAMEA6272116	SAMEA5159861	SAMEA5159863
SAMEA5159864	SAMEA5159865	SAMEA5159866	SAMEA5159867	SAMEA5159868	SAMEA5159869	SAMEA5159870
SAMEA5159871	SAMEA5159872	SAMEA5159873	SAMEA5159874	SAMEA5159875	SAMEA5159885	SAMEA6163175
SAMEA6163176	SAMEA6163177	SAMEA6163180	SAMEA6163182	SAMEA6163185	SAMEA6163186	SAMEA6163187
SAMEA6163188	SAMEA6163189	SAMEA6163190	SAMEA6163191	SAMEA6163192	SAMEA6163193	SAMEA6163194
SAMEA6163195	SAMEA19847668	SAMEA32997418	SAMEA32981668	SAMEA5714976	SAMEA5415485	SAMEA5159847
SAMEA32982418	SAMEA5415486	SAMEA5564716	SAMEA5415489	SAMEA5714979	SAMEA5159853	SAMEA5415488
SAMEA5714971	SAMEA32980918	SAMEA7573623	SAMEA7573624	SAMEA7573625	SAMEA7573627	SAMEA7573628
SAMEA7573631	SAMEA7573632	SAMEA7573633	SAMEA7573635	SAMEA7573636	SAMEA7573637	SAMEA7573639
SAMEA7573640	SAMEA7573642	SAMEA7573645	SAMEA7573646	SAMEA7573647	SAMEA7573649	SAMEA7573549
SAMEA7573551	SAMEA7573553	SAMEA7573554	SAMEA7573555	SAMEA7573556	SAMEA7573559	SAMEA7573560
SAMEA7573561	SAMEA7573562	SAMEA7573564	SAMEA7573565	SAMEA7573567	SAMEA7573570	SAMEA7573571
SAMEA7573573	SAMEA7573578	SAMEA7573583	SAMEA7573587	SAMEA7573588	SAMEA7573589	SAMEA7573591
SAMEA7573592	SAMEA7573593	SAMEA7573597	SAMEA7573599	SAMEA7573604	SAMEA7573607	SAMEA7573608
SAMEA7573610	SAMEA7573614	SAMEA7573616	SAMEA7573617	SAMEA7573539	SAMEA7573540	SAMEA7573541
SAMEA7573542	SAMEA7573543	SAMEA7573544	SAMEA7573545	SAMEA7573546	SAMEA7573547	SAMEA7573548

Additional file 2
Discarded reads

Number of mapped reads contained in the original files but not considered for our study. Number of reads mapped to sexual chromosomes and to unplaced contigs for both assemblies. Low quality mapping includes the number of reads filtered out when considering only uniquely mapped properly paired reads with a mapping quality threshold of 10. Sample-wise mean and standard deviation can be found between parentheses. The length of the sexual chromosomes and unplaced contigs is also included.

	ARS-UCD1.2		UOA_Angus_1	
	Million reads	Length (Mb)	Million reads	Length (Mb)
Chromosome X	1,805	139	1,649	139
Chromosome Y	374	43	432	16
Unplaced contigs	8,127	87	8,603	97
Sexual chromosomes and unplaced contigs	10,321 (64 $\pm$ 38)		10,684 (66 $\pm$ 39)	
Low quality mapping	5,084 (32 $\pm$ 20)		5,099 (32 $\pm$ 20)	

Additional file 3

Number of variants during the different filtering steps

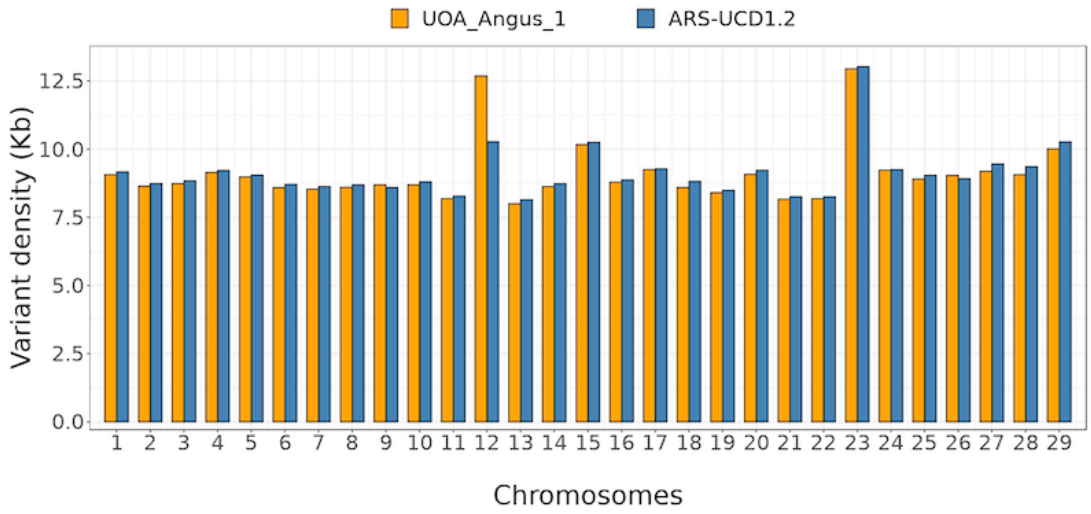
Number of variants during the different filtering steps: from original variants to high-quality and non-fixed variants. Original variants are considered as the raw variants retrieved from GATK. Low quality variants are discarded during hard-filtering and fixed variants are identified when the minor allele count (MAC) is set to 1 in VCFtools. The percentage of variants to the original variants before hard-filtering are in parentheses.

	ARS-UCD1.2	**UOA_Angus_1**
Variants before hard-filtering	24,760,861	24,557,291
SNPs before hard-filtering	21,529,068	21,385,798
INDELs before hard-filtering	3,231,793	3,171,493
Variants discarded by hard-filtering	2,016,344 (8.14)	1,997,616 (8.13)
SNPs discarded by hard-filtering	1,789,081 (8.31)	1,762,677 (8.24)
INDELs discarded by hard-filtering	227,263 (7.03)	234,939 (7.41)
Total number of fixed variants after hard-filtering and imputation	256,256 (1.03)	269,770 (1.10)
Total number of fixed SNPs after hard-filtering and imputation	182,948 (0.85)	176,473 (0.82)
Total number of fixed INDELs after hard-filtering and imputation	73,308 (2.27)	93,297 (2.94)
Fixed variants for the alternate after hard-filtering and imputation	81,674 (0.33)	104,217 (0.42)
Fixed SNPs for the alternate after hard-filtering and imputation	72,121 (0.33)	75,342 (0.35)
Fixed INDELs for the alternate after hard-filtering and imputation	9,553 (0.30)	28,875 (0.91)
High-quality and non-fixed variants	22,488,261 (90.82)	22,289,905 (90.77)
High-quality and non-fixed SNPs	19,557,039 (90.84)	19,446,648 (90.93)
High-quality and non-fixed INDELs	2,931,222 (90.70)	2,843,257 (89.65)

Additional file 4

Variant density of the autosomes for both assemblies.

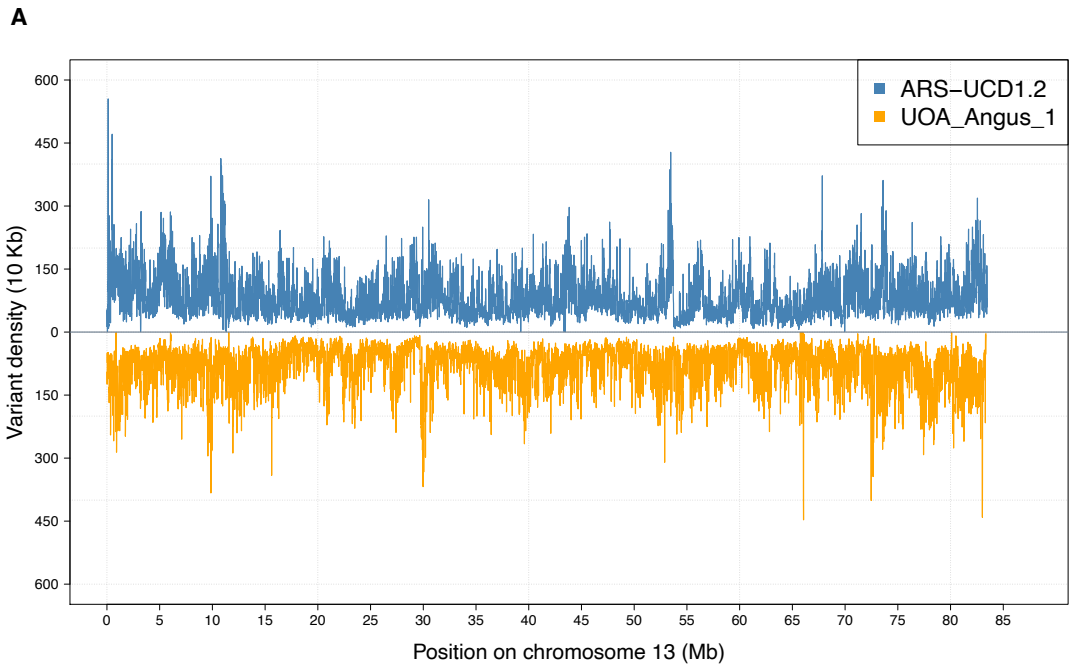
Number of variants detected per kilo base pair (Kb) along autosomal sequences of 161 BSW samples when aligned to the ARS-UCD1.2 (blue) and UOA_Angus_1 (orange) assembly.



Additional file 5

Density of variants across chromosomes 13 and 23.

The number of variants is shown within non-overlapping windows of 10 Kb for chromosome 13 (A) and 23 (B). The x-axis indicates the length of the chromosome (in Mb). The number of variants within each 10 Kb window is shown on the y-axis. Assembly ARS-UCD1.2 is displayed in the top panel (blue) and assembly UOA_Angus_1 is displayed as a mirror image in the bottom panel (orange).



Additional file 6

Density of high-quality and non-fixed variants per Kb along the autosomal genome

Unlike Table 3 in the main text, densities are calculated here when chromosome 12 is not considered.

	ARS-UCD1.2	UOA_Angus_1
High-quality and non-fixed variants per Kb	8.99	8.89
High-quality and non-fixed SNPs per Kb	7.82	7.76
High-quality and non-fixed INDELs per Kb	1.17	1.13

Additional file 7

Number and percentage of multiallelic variants.

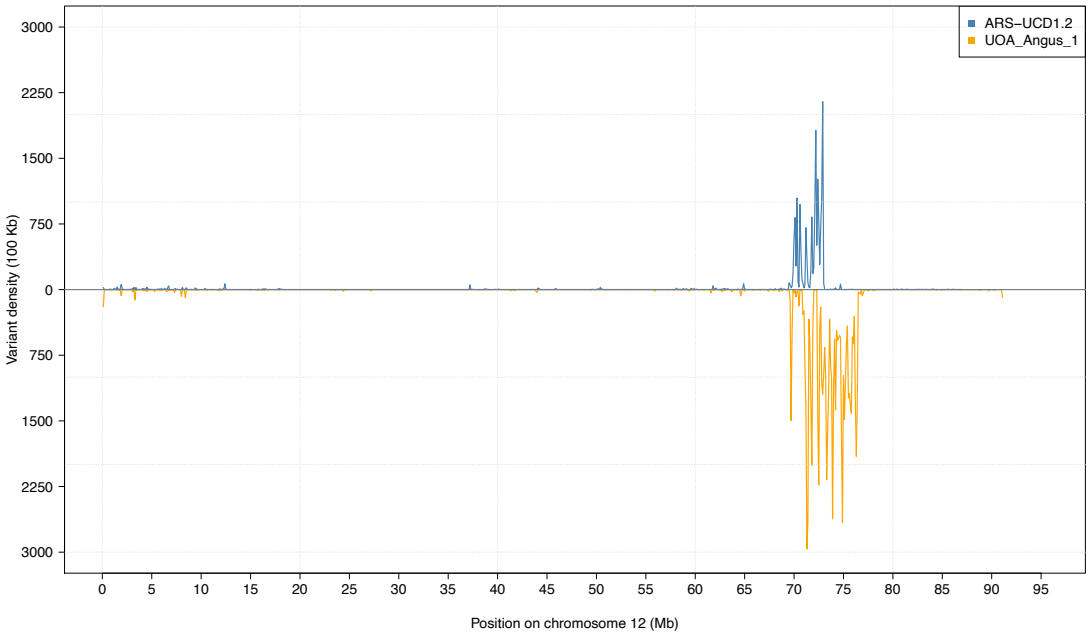
Percentage of multiallelic variants is obtained from the division of multiallelic variants to non-fixed high-quality variants. Multiallelic variants are identified when the '-min-alleles 2 -max-alleles 2' flag is set in VCFtools. Alleles not in Hardy-Weinberg proportions are the number of variants with P-value below the threshold of 10-8 when testing for Hardy-Weinberg proportions with PLINK. Percentages are between parentheses.

	ARS-UCD1.2	**UOA_Angus_1**
Multiallelic variants (%)	848,100 (3.78)	857,206 (3.83)
Multiallelic SNPs (%)	256,545 (1.32)	264,980 (1.34)
Multiallelic INDELs (%)	591,555 (20.21)	592,226 (20.90)
Alleles not in HWE (%)	218,734 (0.97)	243,408 (1.09)

Additional file 8

Density of variants deviating from Hardy-Weinberg proportion for chromosome 12.

The number of variants differing from Hardy-Weinberg proportion are plotted as non-overlapping windows of 10 Kb along the autosomal sequence. The y-axis relates the variant density, number of variants per 100 Kb, for each 10-Kb-windows.



Additional file 9

Summary of the annotated sequence ontology classes of SNPs and INDELs

SO terms are described by Ensembl. Total number of high-quality and non-fixed annotated SNPs and INDELs for both assemblies that were annotated using the release 101 annotation files with VEP tool.

	ARS-UCD1.2		UOA_Angus_1	
	SNPs	INDELs	SNPs	INDELs
intergenic_variant	11,638,948	1,711,642	9,711,869	1,389,027
intron_variant	6,255,771	970,043	8,019,897	1,203,207
upstream_gene_variant	716,481	116,312	693,951	109,258
downstream_gene_variant	611,296	99,692	592,550	92,575
synonymous_variant	117,628	0	122,781	0
missense_variant	86,634	0	99,773	0
3_prime_UTR_variant	69,582	14,052	121,959	24,291
5_prime_UTR_variant	25,688	4,641	27,838	4,999
non_coding_transcript_exon_variant	18,579	2,279	37,976	5,540
splice_region_variant	13,598	2,667	14,553	2,884
stop_gained	1,466	218	1,911	288
splice_donor_variant	506	291	525	298
splice_acceptor_variant	345	292	395	292
start_lost	271	20	319	14
stop_lost	155	11	218	15
stop_retained_variant	86	3	90	7
non_coding_transcript_variant	4	4	17	12
coding_sequence_variant	1	63	26	42
frameshift_variant	0	6,289	0	7,435
inframe_deletion	0	1,761	0	1,972
inframe_insertion	0	850	0	985
protein_altering_variant	0	87	0	107
transcript_ablation	0	5	0	6
start_retained_variant	0	0	0	3
Total	19,557,039	2,931,222	19,446,648	2,843,257

Additional file 10

Candidate selection signatures detected using ARS-UCD1.2 as reference.

Genomic coordinates, CLR values, P-values and encompassed genes for 40 candidate selection signatures.

Chr	Start_bin	End_bin	LR	p_value	Gene_id	Gene_name
1	11909358	11949363	307.87	0.0006832907281	.	.
2	61732510	61752514	443.46	0.0005627100114	ENSBTAG00000013302	R3HDM1
2	61812528	61832531	236.53	0.0007636778726	ENSBTAG00000026842, ENSBTAG00000043334	U6, ZRANB3
2	73795136	73815140	195.91	0.0008922973038	.	.
2	118664905	118684908	200.92	0.000860142446	ENSBTAG00000017855	ITM2C
2	119305044	119325048	237.48	0.0007476004437	.	.
5	4422441	4442444	172.15	0.0009807231628	.	.
5	15464355	15484358	181.43	0.0009405295905	ENSBTAG00000026966	RASSF9
5	18484879	18864943	654.65	0.0004421292947	.	.
5	18884948	18904950	174.97	0.0009646457339	.	.
5	39248477	39288483	332.4	0.0006511358704	ENSBTAG00000012624	PDZRN4
5	46269694	46289697	207.54	0.0008279875882	.	.
5	76054857	76074859	167.74	0.001012878021	.	.
6	36965109	37025116	801.44	0.0004260518658	.	.
6	37045120	38025251	4063.6	1.60774289e-05	ENSBTAG00000005932, ENSBTAG00000005932	DCAF16, NCAPG
6	38045255	38065256	243.55	0.0007315230148	.	.
6	38665338	38685340	197.61	0.0008762198749	.	.
6	38705344	38725346	372.06	0.000627019727	.	.
6	47946592	47966594	274.62	0.0007074068715	.	.
6	48086611	48106613	187.2	0.0009244521616	.	.
6	82531263	82571267	175.2	0.0009566070194	.	.
7	87913975	87973987	361.87	0.0006350584415	ENSBTAG00000050706	lncRNA
7	93715212	93735215	227.64	0.0007797553015	ENSBTAG00000003144	KIAA0825
10	9412209	9492226	379.34	0.0006189810126	ENSBTAG00000007371	SCAMP1
10	19254433	19274437	189.97	0.0009164134472	ENSBTAG00000012981	HEXA
10	19294442	19314446	236.89	0.0007556391582	ENSBTAG00000012981, ENSBTAG00000025814	HEXA, TMEM202
10	70426000	70446003	202.99	0.0008440650171	ENSBTAG00000020880, ENSBTAG00000053337	ARMH4, pseudogene
11	98935806	98995813	220.33	0.0007958327304	ENSBTAG00000015436, ENSBTAG00000015436	SLC27A4, SLC27A4
11	99075826	99095828	205.11	0.0008360263027	ENSBTAG00000030566	GLE1
13	11522254	11542255	546.23	0.0004742841525	ENSBTAG00000040490	CCDC3
13	11622267	11642268	173.93	0.0009726844483	.	.
13	11662272	11682273	282.1	0.0006913294426	ENSBTAG00000008650	CAMK1D
13	11762285	11822291	602.97	0.0004582067236	ENSBTAG00000008650	CAMK1D
13	11862298	11922304	846.99	0.0004180131513	ENSBTAG00000008650, ENSBTAG00000008650	CAMK1D, CAMK1D
13	11962310	12022317	1372.63	0.000393897008	ENSBTAG00000008650, ENSBTAG00000008650	CAMK1D, CAMK1D
14	40641384	40681396	224.91	0.000787794016	.	.
14	40701404	40721410	196.43	0.0008842585894	.	.
16	24027957	24047962	382.96	0.0006109422981	ENSBTAG00000010460	MARK1
21	32563431	32583436	170.14	0.0009968005916	ENSBTAG00000020441	HMG20A
22	3010076	3030083	180.02	0.000948568305	.	.

Additional file 11

Candidate selection signatures detected using UOA_Angus_1 as reference.

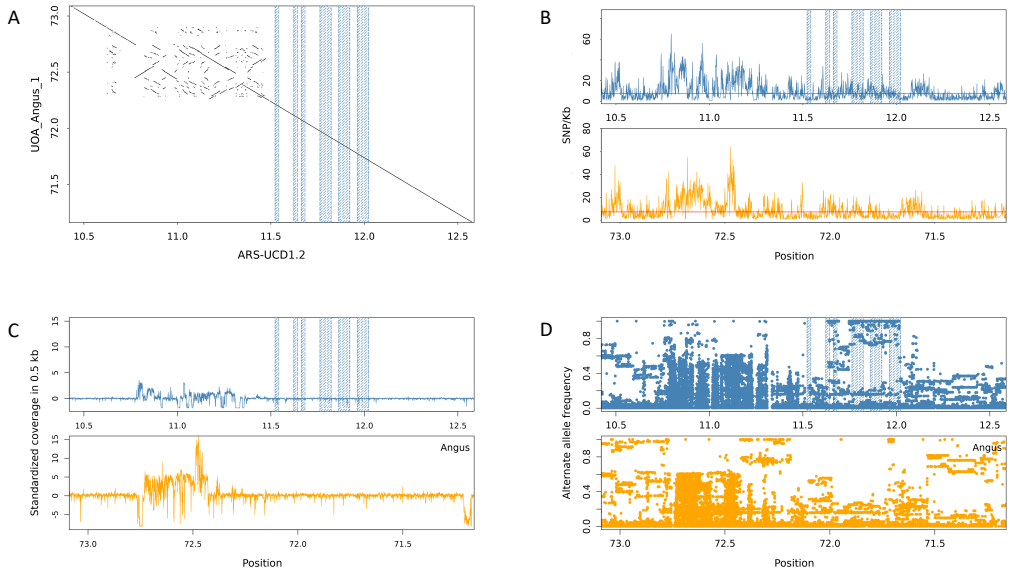
Genomic coordinates, CLR values, P-values and encompassed genes for 33 candidate selection signatures.

Chr	Start_bin	End_bin	LR	p_value	Gene_id	Gene_name
1	137438317	137478322	311.92	0.0008184897648	.	.
2	17663600	17743615	841.63	0.000421400671	ENSBIXG00000019970, ENSBIXG00000019990	protein_coding
2	17763620	17783623	522.31	0.000567270134	ENSBIXG00000019970, ENSBIXG00000019990	protein_coding
2	26365305	26405312	812.12	0.0004376083891	.	.
2	78575535	78595538	250.21	0.0009562553688	.	.
3	13562358	13582360	272.74	0.0008995283554	ENSBIXG00000026575	lncRNA
3	22823888	22863893	403.23	0.0007455550333	ENSBIXG00000024473, ENSBIXG00000024473	CHD1L, CHD1L
3	72712126	72812141	556.98	0.0005186469797	.	.
4	114410784	114430787	235.77	0.001004878523	ENSBIXG00000005066	DDC
5	18475816	18835845	657.1	0.0004700238253	.	.
6	37268147	37288149	671.8	0.0004619199663	.	.
6	37308153	37328155	996.08	0.0004132968119	.	.
6	37348158	38348296	3837.77	1.620771812e-05	ENSBIXG00000001446, ENSBIXG00000001446	DCAF16, FAM184B
6	113158688	113178689	420.87	0.000696931879	ENSBIXG00000023411	SORCS2
10	38786770	38826776	403.43	0.0007374511742	.	.
10	38846780	38906790	308.93	0.0008265936239	.	.
11	36244408	36264409	335.29	0.0008022820467	ENSBIXG00000017054	PSME4
11	65707622	65727623	461.66	0.0006321010065	.	.
11	65927646	65947647	334.11	0.0008103859058	ENSBIXG00000006581	lncRNA
11	102031584	102051585	257.41	0.0009238399326	ENSBIXG00000010765	CFAP77
14	41571833	41591836	234.04	0.001012982382	ENSBIXG00000018543	lncRNA
15	46674925	46694930	616.91	0.0004862315435	.	.
15	46714937	46754949	573.34	0.0005024392616	.	.
16	6821274	6841276	264.23	0.0009157360735	.	.
17	1820474	1840478	414.32	0.0007131395971	ENSBIXG00000012569	TLL1
17	1920496	1980508	832.49	0.0004295045301	ENSBIXG00000005990, ENSBIXG00000012551	lncRNA, protein_coding
22	12289822	12369857	541.3	0.0005267508387	ENSBIXG00000009047	SCN11A
22	12830060	12890085	526.75	0.000559166275	ENSBIXG00000008950	protein_coding
24	16032916	16052921	266.1	0.0009076322145	ENSBIXG00000007455	protein_coding
24	25735873	25755878	295.47	0.0008671129192	.	.
25	5462749	5482758	280.81	0.0008914244963	ENSBIXG00000030306	RBFOX1
26	36882914	36922929	246.56	0.0009724630869	.	.
26	37002963	37082996	384.01	0.0007536588924	ENSBIXG00000010951	lncRNA

Additional file 12

Selective sweeps on chromosome 13.

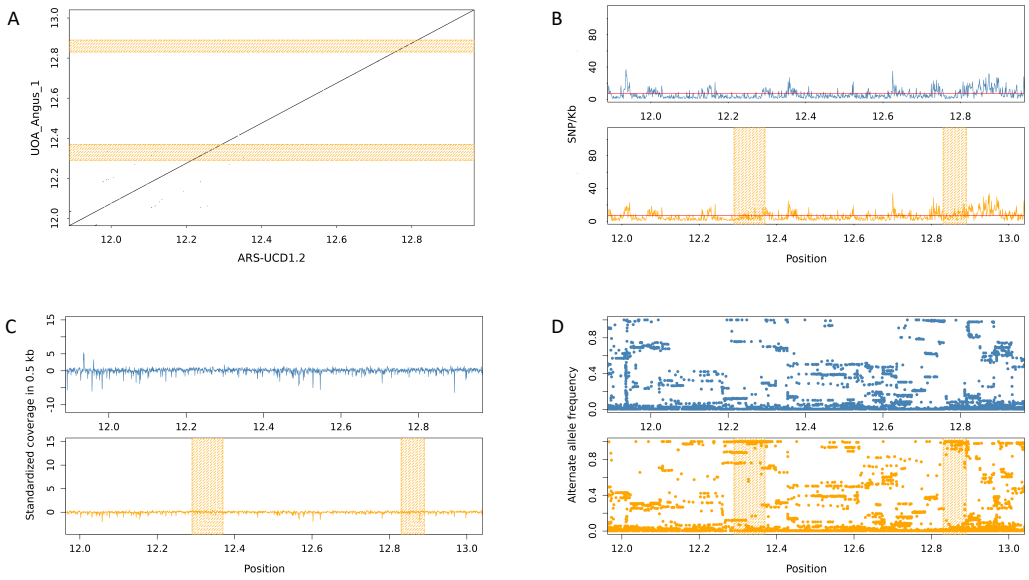
Chromosome 13 region in ARS-UCD1.2 from 10,501,688 - 12,506,844 Mb and corresponding region on UOA_Angus_1 between 71,231,671 73,018,009 Mb with highlighted six selective sweep region from 11.5 Mb to 12 Mb. (A) Dot plot between the two assemblies, (B) SNP density per Kb (red line represents the average SNP density/chromosome), (C) Standardized coverage per 0.5 Kb, (D) Alternate allele frequency of each SNP (each dot is per SNP).



Additional file 13

Selective sweeps on chromosome 22.

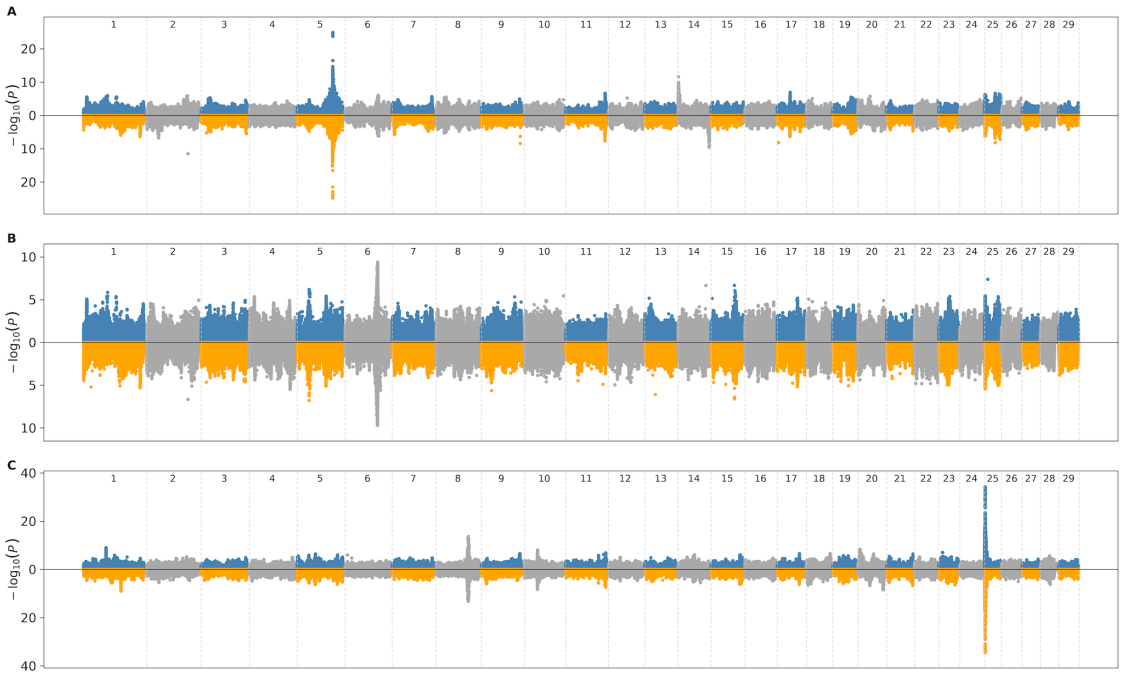
Chromosome 22 region in ARS-UCD1.2 from 11,928,425 - 12,925,926 Mb and corresponding region on UOA_Angus_1 between 12,003,259 13,000,720 Mb with highlighted two selective sweep region. (A) Dot plot between the two assemblies, (B) SNP density per Kb (red line represents the average SNP density/chromosome), (C) Standardized coverage per 0.5 Kb, (D) Alternate allele frequency of each SNP (where each dot is per SNP).



Additional file 14

Genome Wide Association Study (GWAS).

Manhattan plots showing association of sequence variants - imputed using ARS-UCD1.2 (blue and grey) and UOA_Angus_1 (orange and grey) - with fat yield (A), protein yield (B) and stature (C).



Supplementary Material
Chapter 3

Additional file 15

Percentage of overlapping variants across the different GATK and DV sets.

Two different intersection modes were used: exact match (same coordinates, and REF and ALT alleles) and position match (only coordinates were queried - in parentheses).

		Overlapping variants (%)	Overlapping SNPs (%)	Overlapping INDELs (%)
Biallelic	GATK	93.39 (93.55)	95.37 (95.38)	77.97 (79.35)
	DV	91.98 (92.15)	92.61 (92.71)	86.47 (87.18)
Multiallelic	GATK	37.17 (60.45)	24.73 (25.82)	41.26 (71.84)
	DV	38.10 (61.95)	17.52 (27.94)	34.41 (60.02)

Additional file 16

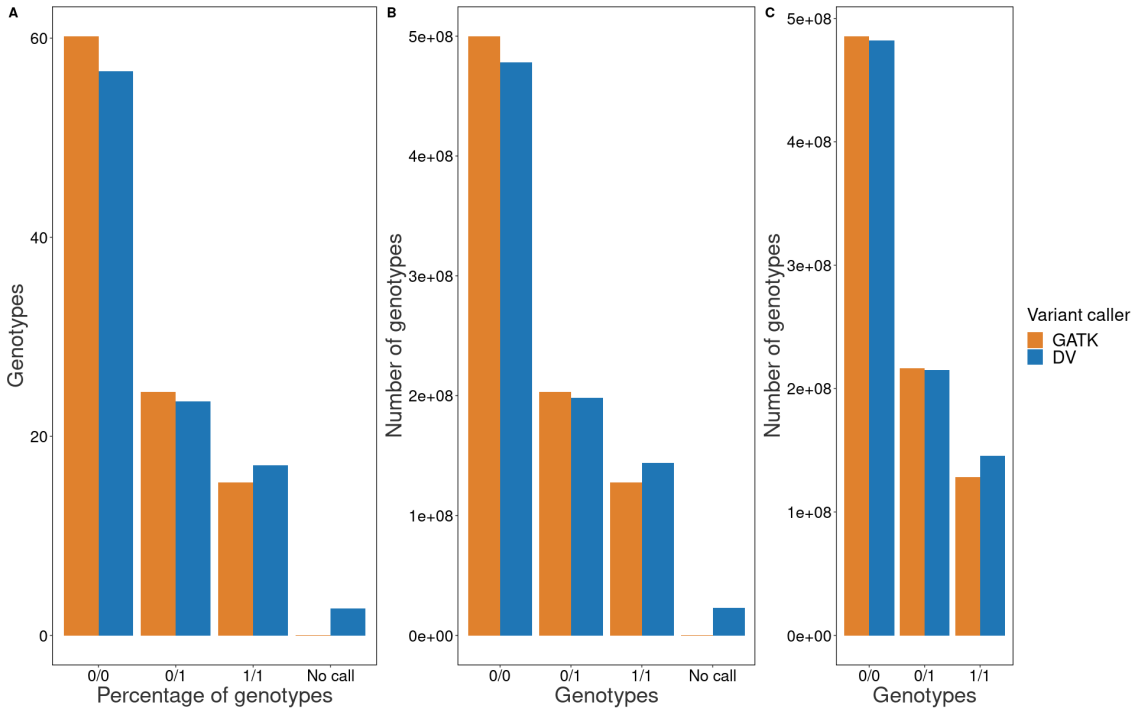
Number of total and multiallelic SNPs shared and private for the different GATK and DV sets.

Two different intersection modes were used: exact match (same REF and ALT alleles) and position match (only coordinates were queried - in parentheses).

	Overlapping variants			Private variants		
	SNPs	**Multiallelic SNPs**	**Multiallelic SNPs (%)**	**SNPs**	**Multiallelic SNPs**	**Multiallelic SNPs (%)**
GATK	14,071,834 (14,134,549)	36,257 (45,367)	0.26 (0.32)	791,774 (729,059)	13,215 (4,105)	1.67 (0.56)
DV	14,071,834 (14,244,141)	36,257 (39,134)	0.26 (0.27)	1,289,951 (1,117,644)	6,642 (3,765)	0.51 (0.34)

Additional file 17
Summary of genotypes.

a) Percentage of filtered genotypes called by each variant caller. b) Number of filtered genotypes called by each variant caller. d) Number of imputed genotypes called by each variant caller.



Additional file 18
Biallelic variants (SNPs / INDELs) annotated with VEP.

Biallelic variants (SNPs / INDELs) annotated with VEP and classified depending on the likely functional effects: high, moderate, low and modifier. Variants were considered before and after filtering (exact match for filtered out). Variants were also divided depending on whether were called by both variant callers (shared) or only one (private) - position matches were accepted.

	Set	High	Moderate	Low	Modifier
GATK	Raw	2,680 / 4,493	67,180 / 1,657	90,955 / 3,403	15,667,553 / 1,985,337
	Raw private	546 / 2,283	10,390 / 674	11,327 / 990	926,538 / 452,647
	Raw shared	2,134 / 2,210	56,790 / 983	79,628 / 2,413	14,741,015 / 1,532,690
	Filtered	2,252 / 3,993	57,675 / 1,522	82,618 / 3,126	14,574,453 / 1,883,261
	Filtered private	362 / 2,089	7,503 / 598	8,714 / 820	664,190 / 401,763
	Filtered shared	1,890 / 1,904	50,172 / 924	73,904 / 2,306	13,910,263 / 1,481,498
	Filtered out	428 / 500	9,505 / 135	8,337 / 277	1,093,100 / 102,076
DV	Raw	3,530 / 2,778	69,525 / 1,162	88,839 / 2,833	16,125,916 / 1,748,769
	Raw private	1,396 / 572	12,733 / 175	9,213 / 421	1,384,901 / 216,078
	Raw shared	2,134 / 2,206	56,792 / 987	79,628 / 2,412	14,741,015 / 1,532,691
	Filtered	2,474 / 2,240	58,457 / 1,068	80,765 / 2,730	15,013,146 / 1,700,025
	Filtered private	584 / 338	8,283 / 142	6,863 / 424	1,102,883 / 218,527
	Filtered shared	1,890 / 1,902	50,174 / 926	73,902 / 2,306	13,910,263 / 1,481,498
	Filtered out	1,061 / 612	11,214 / 135	8,226 / 348	1,142,970 / 134,643

Additional file 19

VEP annotation of GATK and DV private variants.

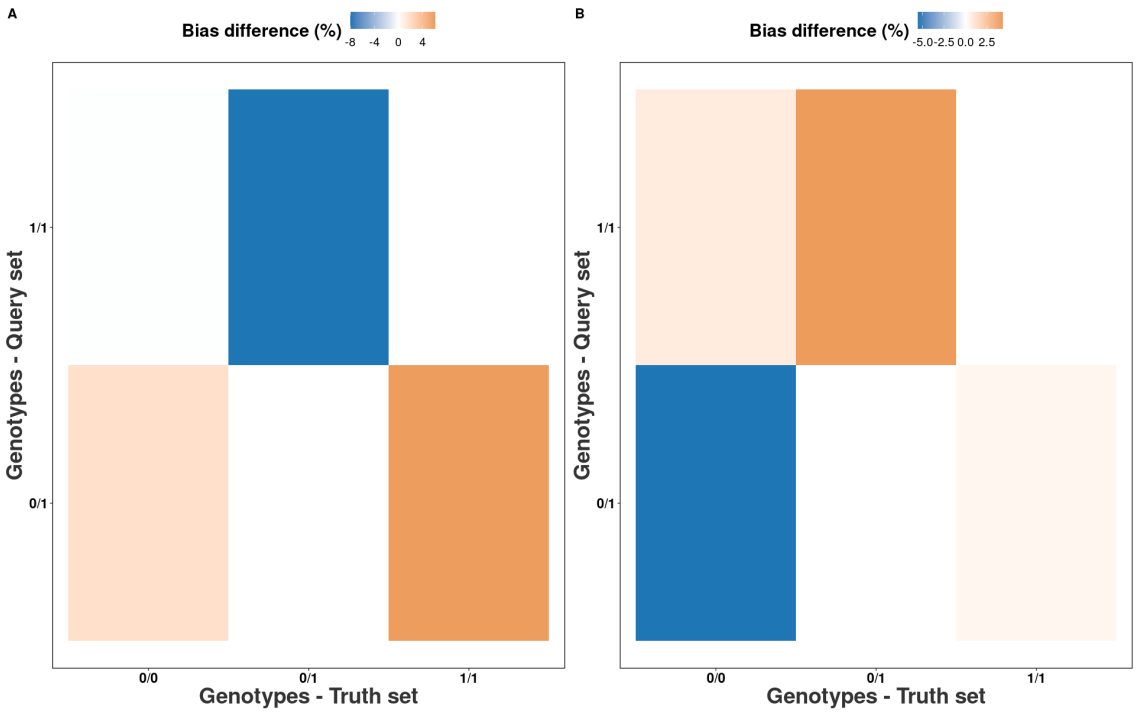
MAFs greater than 0.05 are bolded, and HOMALT samples refers to the number of samples with 1/1 genotypes for the variant (out of a maximum of 33 genotyped samples). Some variants were present in the other caller but as a different variant than predicted by truth. Truly missing variant calls are indicated by "-".

Variant caller	Position	Predicted impact	MAF	HOMOALT samples	Reason not intersecting
GATK	9:78221981	Moderate	0.78	21	INDEL
GATK	19:46232626	High	0.02	0	-
DV	5:116965831	Low	0.09	1	-
DV	6:37374718	Low	1	33	-
DV	11:15257933	Low	0.57	10	INDEL
DV	15:1094876	Low	0.75	16	-
DV	25:20777430	Low	0.13	1	-
DV	15:49709135	Moderate	0.06	2	-
DV	18:61285359	Moderate	0.08	0	Multiallelic
DV	4:49174950	High	0.02	0	INDEL
DV	7:41125658	High	0.26	1	Multiallelic
DV	10:27971463	High	0.01	0	INDEL
DV	15:49811331	High	0.29	3	-

Additional file 20

Genotyping accuracy of variant calls.

Categorisation and comparison of filtered (a) and imputed (b) genotypes detected with hap.py, where the colour intensity indicates the percentual differences between GATK and DeepVariant (DV).



Additional file 21

F1, recall and precision scores when comparing the truth set and the query sets (different coverages and panels).

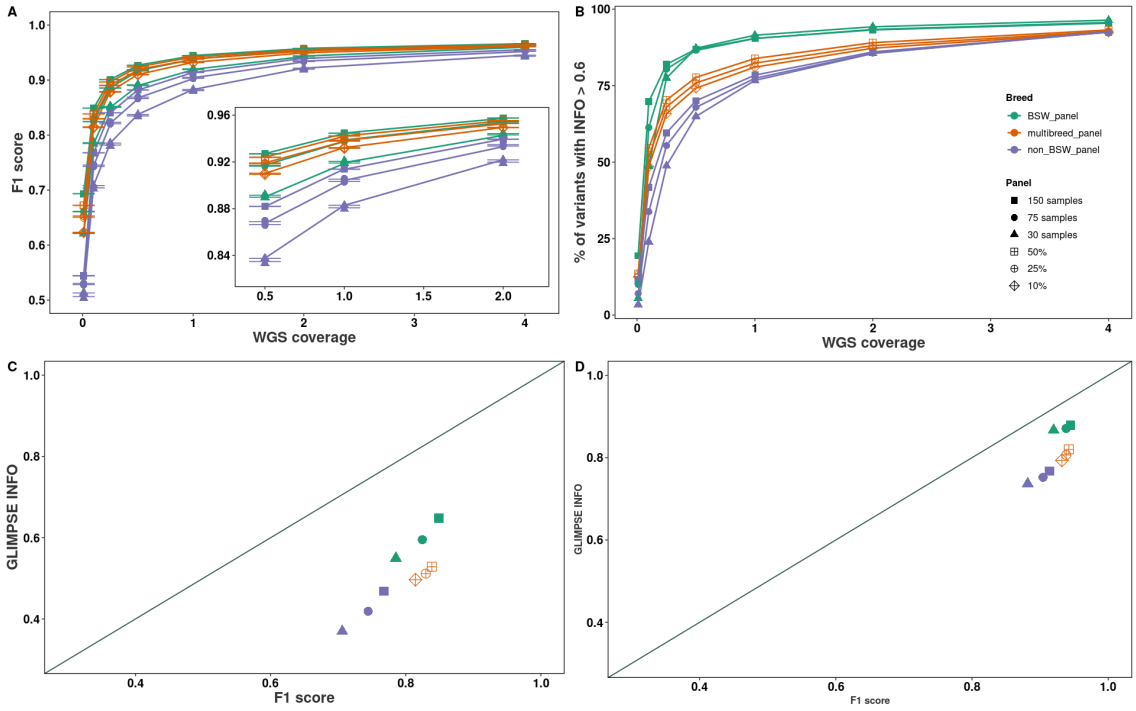
Numbers in pure panels indicate the number of samples. Percentages in multibreed panels indicate the proportion of BSW samples. The top values for each metrics and coverage are highlighted.

Metric	Panel	4x	2x	1x	0.5x	0.25x	0.1x	0.01x
F1	BSW (150)	0.9663	0.9575	0.9456	0.9269	0.9009	0.8490	0.6934
	BSW (75)	0.9655	0.9543	0.9381	0.9167	0.8855	0.8245	0.6609
	BSW (30)	0.9608	0.9436	0.9198	0.8906	0.8513	0.7853	0.6215
	Multibreed (50%)	0.9642	0.9554	0.9422	0.9239	0.8966	0.8385	0.6722
	Multibreed (25%)	0.9629	0.9529	0.9382	0.9188	0.8901	0.8296	0.6518
	Multibreed (10%)	0.9614	0.9494	0.9320	0.9099	0.8786	0.8144	0.6229
	Non-BSW (150)	0.9552	0.9394	0.9138	0.8819	0.8405	0.7677	0.5442
	Non-BSW (75)	0.9526	0.9342	0.9042	0.8676	0.8218	0.7443	0.5291
	Non-BSW (30)	0.9444	0.9208	0.8818	0.8361	0.7837	0.7061	0.5097
Recall	BSW (150)	0.9700	0.9565	0.9380	0.9144	0.8810	0.8167	0.6423
	BSW (75)	0.9677	0.9515	0.9293	0.9008	0.8607	0.7858	0.6064
	BSW (30)	0.9591	0.9367	0.9059	0.8686	0.8189	0.7394	0.5679
	Multibreed (50%)	0.9665	0.9528	0.9338	0.9092	0.8736	0.8018	0.6170
	Multibreed (25%)	0.9650	0.9500	0.9293	0.9034	0.8665	0.7921	0.5975
	Multibreed (10%)	0.9628	0.9459	0.9222	0.8934	0.8535	0.7756	0.5702
	Non-BSW (150)	0.9523	0.9317	0.8993	0.8596	0.8100	0.7243	0.4961
	Non-BSW (75)	0.9457	0.9222	0.8849	0.8394	0.7835	0.6939	0.4760
	Non-BSW (30)	0.9282	0.8995	0.8526	0.7974	0.7350	0.6475	0.4535
Precision	BSW (150)	0.9626	0.9586	0.9512	0.9398	0.9218	0.8839	0.7534
	BSW (75)	0.9632	0.9571	0.9472	0.9331	0.9118	0.8672	0.7262
	BSW (30)	0.9625	0.9507	0.9340	0.9138	0.8863	0.8373	0.6862
	Multibreed (50%)	0.9620	0.9580	0.9507	0.9392	0.9208	0.8788	0.7383
	Multibreed (25%)	0.9609	0.9559	0.9473	0.9346	0.9152	0.8709	0.7171
	Multibreed (10%)	0.9599	0.9530	0.9419	0.9269	0.9051	0.8572	0.6864
	Non-BSW (150)	0.9583	0.9473	0.9287	0.9054	0.8746	0.8167	0.6027
	Non-BSW (75)	0.9597	0.9465	0.9245	0.8978	0.8641	0.8026	0.5956
	Non-BSW (30)	0.9611	0.9431	0.9133	0.8788	0.8392	0.7763	0.5818

Additional file 22

Genotyping accuracy from low-pass whole-genome sequencing.

a) F1 score between truth and imputed variants, with error bars accounting for the three replicas. b) Percentage of imputed variants with a GLIMPSE's estimated accuracy higher than 0.6. Relationship between F1 score and estimated imputation accuracy for lcWGS at 0.1x (c) and 1x (d). Panels are indicated with colours and number/percentage of BSW samples in different point shapes.



Additional file 23

Compute resources used by DeepVariant (DV) and GATK to pre-process aligned BAM files, call variants per sample (gVCF stage), and jointly genotype and filter variants (pVCF stage).

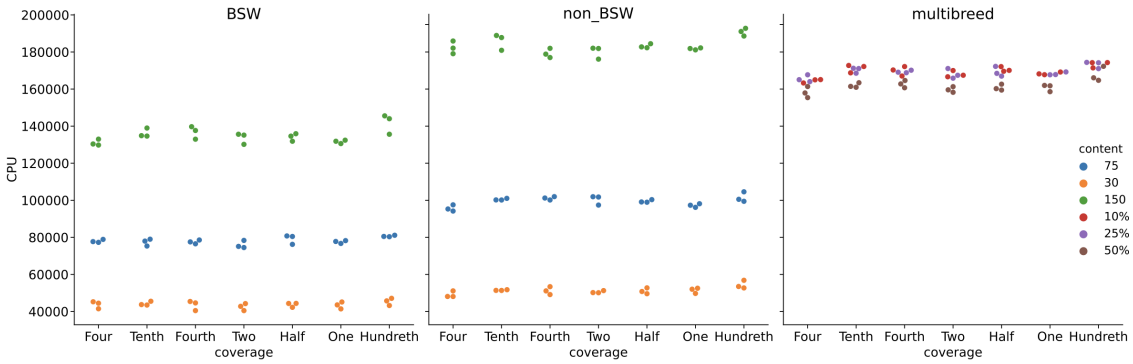
Times are listed as CPU hours, and peak memory usage across all stages is given in gigabytes. DV does not require pre-processing. Jobs were submitted to nodes with different CPUs and non-exclusive use, and so these figures do not represent precise benchmarking. However, any node/usage variability is minor compared to the differences in total CPU hours used between DV and GATK.

Variant caller	Preprocessing (h)	gVCF stage (h)	pVCF stage (h)	Peak memory (GB)
DV	-	577	3	81
GATK	820	999	233	49

Additional file 24

CPU hours required to impute different coverages and panels for 3 replicates.

Compute time was dominated by panel size followed by panel composition, while the input coverage had limited effect.



Acknowledgements

First, I would like to thank Prof. Dr. Hubert Pausch for providing me with the opportunity to pursue a doctorate at such world-renowned university. I am deeply grateful for the unvaluable mentorship and support, instrumental throughout my academic journey.

Thanks to Prof. James Prendergast who has agreed to join the examination committee and review this thesis.

My gratitude goes to both the current and former members of the Animal Genomics group, from whom I learnt tremendously and with whom I have felt incredibly comfortable during my time here. I would like to extend a special acknowledgment to Dr. Alexander S. Leonard for his exceptional advices and for dedicating his valuable time to review and guide my research.

Thanks to Laura Olivares Boldú, Jessica Deuber and Dr. Danang Crysnanto for helping with the design and enhancing the overall quality of this thesis.

Als meus pares i als meus germans, per l'amor i la pedagogia.

Elle ne me quitte pas d'un pas, fidèle comme une ombre. Elle m'a suivi çà et là aux quatre coins du monde. Non, je ne suis jamais seul avec ma solitude.