bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Evaluation of the accuracy of imputed sequence variants and their utility for causal variant detection in cattle

Hubert Pausch[1,$], Iona M MacLeod[1], Ruedi Fries[2], Reiner Emmerling[3], Phil J Bowman[1,4], Hans D Daetwyler[1,4], Michael E Goddard[1,5]

[1] Biosciences Research, Agriculture Victoria, Department of Economic Development, Jobs, Transport and Resources, AgriBio, 3083 Victoria, Australia
[2] Chair of Animal Breeding, Technische Universitaet Muenchen, 85354 Freising, Germany
[3] Institute of Animal Breeding, Bavarian State Research Center for Agriculture, 85586 Grub, Germany
[4] School of Applied Systems Biology, LaTrobe University, Bundoora, 3083 Victoria, Australia
[5] Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Melbourne, 3010 Victoria, Australia

[$]Corresponding author: hubert.pausch@ecodev.vic.gov.au

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Background: The availability of dense genotypes and whole-genome sequence variants from various sources offers the opportunity to compile large data sets consisting of tens of thousands of individuals with genotypes at millions of polymorphic sites that may enhance the power of genomic analyses. The imputation of missing genotypes ensures that all individuals have genotypes for a shared set of variants.

Results: We evaluated the accuracy of imputation from dense genotypes to whole-genome sequence variants in 249 Fleckvieh and 450 Holstein cattle using *Minimac* and *FImpute*. The sequence variants of a subset of the animals were reduced to the variants that were included in the Illumina BovineHD genotyping array and subsequently inferred *in silico* using either within- or multi-breed reference populations. The accuracy of imputation varied considerably across chromosomes and dropped at regions where the bovine genome contains segmental duplications. Depending on the imputation strategy, the correlation between imputed and true genotypes ranged from 0.898 to 0.952. The accuracy of imputation was higher with *Minimac* than *FImpute* particularly for rare alleles. Considering a multi-breed reference population increased the accuracy of imputation, particularly when *FImpute* was used to infer genotypes. When the sequence variants were imputed using *Minimac*, the true genotypes were more correlated to predicted allele dosages than best-guess genotypes. The computing costs to impute 23,256,743 sequence variants in 6958 animals were ten-fold higher with *Minimac* than *FImpute*. Association studies with imputed sequence variants revealed seven quantitative trait loci (QTL) for milk fat percentage. Two causal mutations in the *DGAT1* and *GHR* genes were the most significantly associated variants at two QTL on chromosomes 14 and 20 when *Minimac* was used to infer genotypes.

Conclusions: The population-based imputation of millions of sequence variants in large cohorts is computationally feasible and provides accurate genotypes. The accuracy of imputation is low at regions where the genome contains large structural variants. Using a reference population that includes individuals from many breeds increases the accuracy of imputation particularly at low-frequency variants.

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Considering allele dosages rather than best-guess genotypes as explanatory variables is advantageous to detect causal mutations in association studies with imputed sequence variants.

Background

Several genotyping arrays that comprise a varying number of single nucleotide polymorphisms (SNPs) are routinely used for genome-wide genotyping in cattle. Cows are usually genotyped using low-density genotyping arrays whereas bulls are mostly genotyped at higher density [1]. Moreover, the sequencing of important ancestors of many cattle breeds yielded genotypes at millions of polymorphic sites [2]. Combining the genotype data from various sources into a single large data set may enhance the power of genome-wide analyses. The imputation of missing genotypes is necessary to ensure that all individuals have genotypes for a shared set of variants. Genotype imputation may also infer dense genotypes *in silico* for animals that were genotyped at lower density using animals that were genotyped at a higher density as a reference [3].

Algorithms that infer missing genotypes apply family- (*e.g.*, [4]) or population-based (*e.g.*, [5-7]) imputation approaches or a combination thereof (*e.g.*, [8-10]). Family-based imputation approaches rely on Mendelian transmission rules in pedigrees to infer missing genotypes. Population-based imputation approaches exploit linkage disequilibrium (LD) between adjacent markers to predict missing genotypes using a probabilistic framework without (explicitly) considering pedigree information [11]. While population-based imputation approaches are accurate, their computing costs are too high to infer genotypes for a large number of animals and markers for routine applications [12, 13]. Methods that apply a combination of family- and population-based imputation approaches exploit shared haplotypes among relatives thereby enabling rapid imputation of genotypes for tens of thousands of individuals and millions of markers *in silico* [8, 9, 10, 14].

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The accuracy of imputation from low to higher density depends on the relationships between target and reference animals, genotype density in both panels, allele frequencies of the imputed variants, LD between adjacent markers and algorithms applied to infer missing genotypes (*e.g.*, [12, 15, 16]). These parameters also affect the accuracy of imputation from dense genotypes to sequence variants [2, 17, 18]. However, the accurate imputation of low-frequency variants is critical with sequence data because rare alleles are more frequent in sequence than array-derived variants and the LD between rare alleles and array-derived variants may be low [19, 20].

A number of studies with simulated and real sequence data in cattle indicated that using imputed sequence variants may improve genomic predictions and facilitate the detection of causal trait variants because the polymorphisms that underlie phenotypic variation are included in the data [2, 19, 21-23]. However, the benefits of dense marker maps for genome-wide analyses may be compromised when the accuracy of imputation is low [24-26].

In this paper, we evaluate the accuracy of imputation for sequence variants in two cattle breeds using different imputation algorithms and reference populations. We perform association studies between imputed sequence variants and milk fat percentage in 6958 Fleckvieh bulls and show that the imputation strategy is critical to pinpoint causal mutations.

Methods

Animal ethics statement

No ethical approval was required for this study.

Animals

We used whole-genome sequence data of 1577 taurine animals that were included in the fifth run of the of the 1000 bull genomes project [2]. The reads were aligned to the UMD3.1 bovine reference genome using the *BWA-MEM* algorithm [27, 28]. SNPs, short insertions and deletions were genotyped for all sequenced animals simultaneously using a multi-sample variant calling pipeline that was implemented

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

with the *mpileup* module of *SAMtools* [29] and that is described in Daetwyler et al. [2]. The variant calling yielded genotypes at 39,721,987 biallelic sites for 1577 animals. We considered genotypes at 22,737,136 autosomal sequence variants that had a minor allele frequency (MAF) higher than 0.5% to build the genomic relationship matrix among the sequenced animals using the *plink* (version 1.9) software tool [30]. Principal components of the genomic relationship matrix were calculated using the *GCTA* (version 1.25.3) software tool [31]. Following the inspection of the top principal components, animals whose breed was uncertain were removed leaving 249 Fleckvieh (FV) and 450 Holstein (HOL) cattle (see Additional file 1 Figure S1).

Design of cross-validation scenarios

The imputation from dense genotypes to full sequence variants was evaluated using 15-fold cross-validation in FV and HOL cattle. We considered only sequence variants that were polymorphic in 249 FV or 450 HOL animals. For each fold, the sequenced animals were divided into reference and validation animals. Forty-nine FV and 100 HOL animals, that were a random subset of the sequenced animals, were considered as validation animals. All remaining animals were considered as a multi-breed reference population. The within-breed reference populations consisted of 200 FV or 350 HOL animals. The genotypes of the validation animals were reduced to the variants that were included in the Illumina BovineHD genotyping array (HD) to mimic dense genotypes. The masked genotypes of the validation animals were subsequently inferred *in silico* using full sequence information from the reference animals. The selection of validation animals and subsequent imputation of genotypes was repeated 15 times for chromosomes 1, 5, 10, 15, 20 and 25.

Evaluation of the accuracy of imputation

The overall accuracy of imputation was the mean correlation between *in silico* imputed and true (sequenced) genotypes ($r_{IMP,SEQ}$) across 15 folds and six chromosomes analysed. We additionally grouped the imputed sequence variants into 50 classes with regard to their MAF in the reference population. The $r_{IMP,SEQ}$–value for each MAF class was the mean value across six chromosomes and 15 folds. To detect intra-chromosomal variations in the accuracy of imputation, we calculated

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$r_{IMP,SEQ}$-values for successive 1 Mb segments along each chromosome analysed. Differences between imputation scenarios were tested using two-tailed t-tests.

Imputation methods

The performance of the *FImpute* (version 2.2) [9] and *Minimac3* (version 2.0.1, henceforth referred to as *Minimac*) [7] software tools was evaluated using default parameter settings. The algorithm implemented in *FImpute* uses family and population-based information to infer haplotypes and missing genotypes. A pedigree consisting of 47,012 FV animals tracing back ancestry information to animals born in 1925 was considered when genotypes were inferred using *FImpute*. Pedigree records for sequenced animals from other breeds were also included. However, their pedigrees included only first-degree relatives. Genotypes for the imputed sequence variants were coded as 0, 1 and 2 for homozygous, heterozygous and alternative homozygous animals, respectively.

The algorithm implemented in the *Minimac* software tool uses previously phased genotypes, *i.e.*, it takes haplotypes as input for reference and validation animals. Haplotype phases were estimated separately for the reference and validation animals using the phasing algorithm implemented in the *Eagle* (version 2.3) software tool [32]. Both *Eagle* and *Minimac* do not consider pedigree information to infer haplotypes and missing genotypes. *Minimac* provides best-guess genotypes (coded as 0, 1 and 2 for homozygous, heterozygous and alternative homozygous animals, respectively) and allele dosages (continuously distributed values ranging from 0 to 2) for the imputed sequence variants. The $r_{IMP,SEQ}$–values were calculated for best-guess genotypes ($Minimac_{BG}$) and allele dosages ($Minimac_{DOS}$).

Imputation of sequence variants in 6958 Fleckvieh animals

We imputed genotypes for 23,256,743 autosomal sequence variants that were polymorphic in 249 sequenced FV animals for 6958 FV bulls that had (partially imputed) array-derived genotypes at 603,662 autosomal variants (see [33]). Genotypes were imputed with either *FImpute* or *Minimac* using either within- or multi-breed reference populations that consisted either of 249 FV animals or 1577 animals from various bovine breeds (see above). Haplotype phases for reference and target animals were estimated using the *Eagle* software tool (see above). The LD

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

among imputed sequence variants was calculated as the squared correlation between predicted allele dosages.

Accuracy of imputation at two known causal mutations

712 and 902 FV animals that had imputed sequence variants also had direct genotypes for two mutations in the *DGAT1* (rs109234250 and rs109326954, p.A232K) and *GHR* (rs385640152, p.F279Y) genes that were obtained for previous studies using TaqMan® genotyping assays (Life Technologies) [34-37]. The accuracy of imputation at both sites was the proportion of correctly imputed genotypes.

Association analyses in Fleckvieh cattle

Association tests between 23,256,743 imputed sequence variants and milk fat percentage were carried out in 6958 FV bulls using a variance components-based approach that was implemented in the *EMMAX* software tool and that accounts for population stratification by fitting a genomic relationship matrix [38]. The genomic relationship matrix was built using array-derived genotypes of 603,662 (partially imputed) autosomal SNPs [39]. Daughter yield deviations (DYDs) for milk fat percentage were the response variables. The genotypes were coded as 0, 1 and 2 for homozygous, heterozygous and alternative homozygous animals, respectively, when sequence variants were imputed using *FImpute*. When the sequence variants were imputed using *Minimac*, we considered both best-guess genotypes and allele dosages as explanatory variables for the association tests. Sequence variants with P values less than 2.1×10^{-9} (Bonferroni-corrected significance threshold for 23,256,743 tests) were considered as significantly associated.

Computing environment

All computations were performed on the Biosciences Advanced Scientific Computing (BASC) cluster that is located at AgriBio, Centre for AgriBiosciences, VIC 3083, Bundoora. The memory usage and process time required to infer genotypes for 23,256,743 sequence variants in 6958 animals was quantified on 12-core Intel® Xeon® processors rated at 2.93 Ghz with 96 gigabytes of random-access memory (RAM).

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

We evaluated the accuracy of imputation from dense genotypes to sequence variants in 249 FV and 450 HOL animals using sequence data on bovine chromosomes (BTA for Bos taurus) 1, 5, 10, 15, 20 and 25. The number of polymorphic sites ranged from 413,371 (BTA25) to 1,444,299 (BTA1) and from 383,072 (BTA25) to 1,382,987 (BTA1) in FV and HOL, respectively, indicating that genetic diversity is higher in FV than HOL cattle (Table 1). Rare variants were more frequent among the sequence than HD variants; between 58.12 and 60.55% of the sequence variants and between 14.27 and 18.55% of the HD variants had a MAF lower than 10%.

Table 1 - Number of polymorphic sites in Fleckvieh and Holstein cattle

Chr	Chr. Length (Mb)	Fleckvieh		Holstein	
		SEQ (MAF<0.1)	700K (MAF<0.1)	SEQ (MAF<0.1)	700K (MAF<0.1)
1	158.32	1,444,299 (60.55)	38,009 (18.40)	1,382,987 (59.11)	37,397 (17.01)
5	121.18	1,098,976 (59.64)	28,173 (18.55)	1,073,964 (59.12)	27,667 (15.90)
10	104.30	959,206 (59.13)	25,787 (17.09)	917,799 (59.48)	25,497 (16.12)
15	85.27	866,151 (58.54)	20,617 (17.24)	827,526 (58.12)	20,360 (14.38)
20	71.98	679,738 (59.94)	18,466 (18.18)	649,768 (59.18)	18,264 (18.13)
25	42.85	413,371 (59.08)	11,370 (14.69)	383,072 (58.14)	11,179 (14.27)

Number of sequence (SEQ) and HD (700K) variants located on six chromosomes that were polymorphic in 249 Fleckvieh and 450 Holstein animals. The proportion of variants with a minor allele frequency (MAF) lower than 10% is given in parentheses.

Evaluation of the accuracy of imputation

The correlation between imputed and sequenced genotypes ($r_{IMP,SEQ}$) was higher in HOL than FV cattle (Table 2) likely reflecting that more HOL animals were in the reference population. When within-breed reference populations were considered, the $r_{IMP,SEQ}$–values were 0.898, 0.908 and 0.934 for *FImpute*, *Minimac*$_{BG}$ and *Minimac*$_{DOS}$ in FV and 0.912, 0.929 and 0.951 in HOL, respectively. Adding animals from various breeds to the reference panel increased $r_{IMP,SEQ}$ in FV, particularly when *FImpute* was used to impute missing genotypes ($P=7.5x10^{-11}$, Table 2). The highest accuracy of imputation in FV ($r_{IMP,SEQ}$=0.939) was achieved using *Minimac*$_{DOS}$ with a multi-breed reference population. In HOL, a multi-breed reference population increased $r_{IMP,SEQ}$ when *FImpute* was used to infer missing genotypes ($P=4.0x10^{-4}$). However, the use of a multi-breed reference population had little effect on the $r_{IMP,SEQ}$-values in HOL when *Minimac* was used (P>0.12). Using *Minimac*$_{DOS}$ and a within-breed reference

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

population provided the highest accuracy of imputation ($r_{IMP,SEQ}$=0.951) in HOL. Regardless of the reference population used, the $r_{IMP,SEQ}$–values were higher for *Minimac*$_{DOS}$ than *Minimac*$_{BG}$ in both breeds ($P_{FV}<2.9x10^{-6}$, $P_{HOL}<3.9x10^{-6}$).

Table 2 - Cross-validation imputation accuracy in Fleckvieh and Holstein cattle

Breed	Within-breed reference populations			Multi-breed reference populations		
	Fimpute	*Minimac*$_{BG}$	*Minimac*$_{DOS}$	*Fimpute*	*Minimac*$_{BG}$	*Minimac*$_{DOS}$
Fleckvieh	0.898	0.908	0.934	0.921	0.916	0.939
	(0.674)	(0.735)	(0.771)	(0.774)	(0.812)	(0.838)
Holstein	0.912	0.929	0.951	0.926	0.928	0.948
	(0.681)	(0.772)	(0.798)	(0.732)	(0.797)	(0.820)

The values represent the correlation between true and imputed genotypes. The correlation between true and imputed genotypes for sequence variants with a minor allele frequency lower than 10% is given in parentheses. BG: best-guess, DOS: allele dosage

A decline in the accuracy of imputation was evident for low-frequency variants across all scenarios tested. While $r_{IMP,SEQ}$ was high for variants with a MAF higher than 10%, it was considerably less for rare variants (Figure 1a-b). In FV, the accuracy of imputation was higher for rare alleles using multi- than within-breed reference populations ($P<6.6x10^{-6}$, Table 2). The benefit of multi-breed reference panels was less pronounced in HOL and it was only significant (P=0.013) when *FImpute* was used. The $r_{IMP,SEQ}$–values for low-frequency variants (MAF<0.1) were higher with *Minimac* than *FImpute* ($P_{FV}<0.005$, $P_{HOL}<0.002$, Table 2, Figure 1a-b).

The accuracy of imputation varied considerably across chromosomes. The $r_{IMP,SEQ}$-values were high for sequence variants located on BTA1, BTA20 and BTA25. However, the accuracy of imputation was lower for sequence variants located on BTA5, BTA10 and BTA15 (Figure 1c-d). This pattern was observed regardless of the reference population and imputation software used. To investigate the reason for the variable accuracy of imputation across chromosomes, we calculated $r_{IMP,SEQ}$–values for successive 1 Mb windows along the six chromosomes analysed. Multiple (partially extended) segments with high imputation errors were located on chromosomes 5, 10 and 15 (see Additional file 2 Figure S2) at positions where the bovine genome contains large segmental duplications [40, 41]. Such segments with

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

low imputation quality were not detected along chromosomes 1, 20 and 25, that had an overall higher accuracy of imputation.

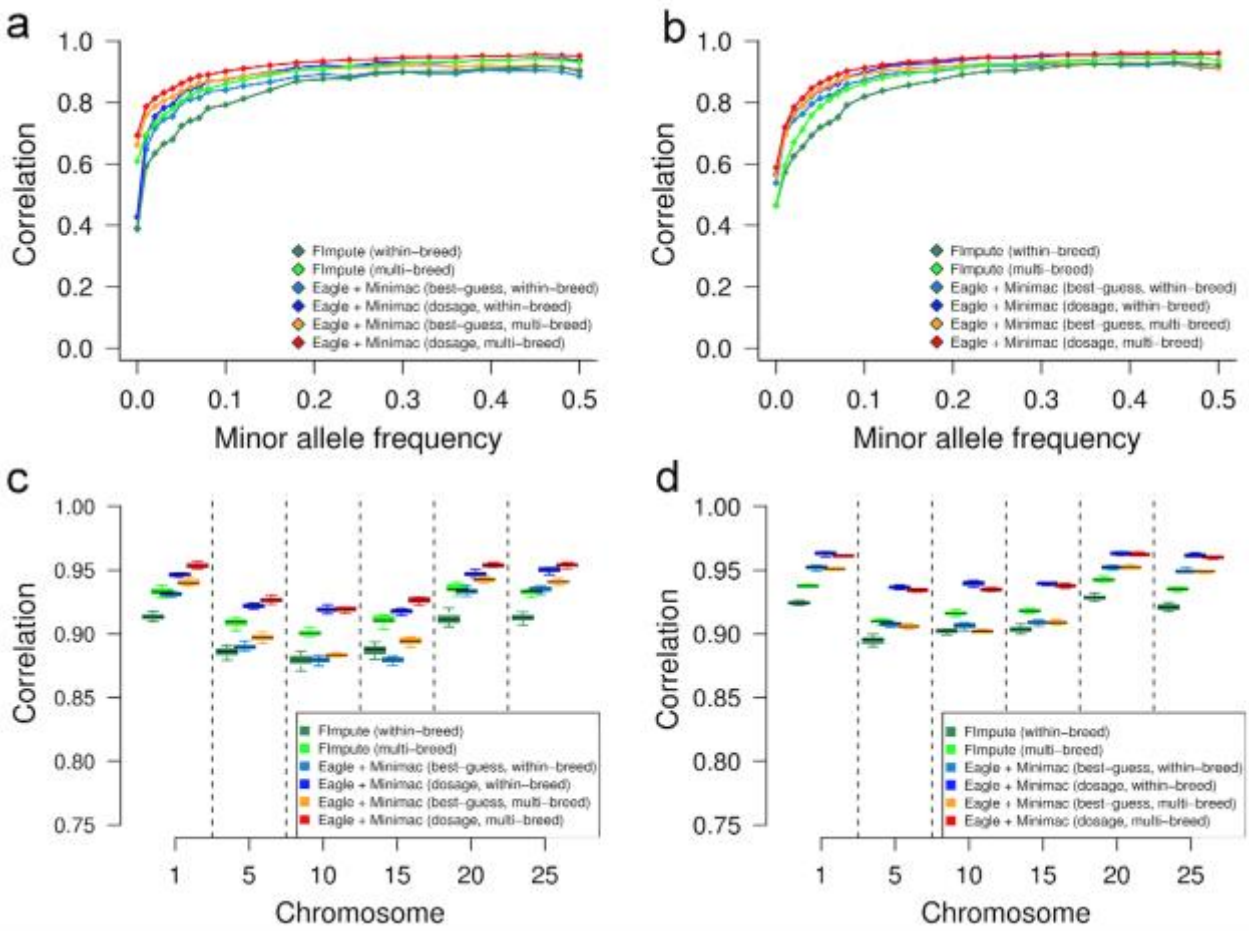

Figure 1 - The accuracy of imputation in Fleckvieh and Holstein cattle.
(a-b) The correlation between true and imputed genotypes as a function of the frequency of the minor allele in Fleckvieh (a) and Holstein (b) animals. (c-d) The boxplots represent the correlation between true and imputed genotypes for sequence variants on six chromosomes in Fleckvieh (c) and Holstein (d) animals.

Pinpointing causal mutations among imputed sequence variants

We imputed genotypes for 756,135 and 679,738 sequence variants that were located on BTA14 and BTA20, respectively, in 6958 FV animals using either *FImpute* or *Minimac* and considering either within- or multi-breed reference populations. To evaluate the ability to detect causal mutations, the imputed genotypes were tested for association with milk fat percentage. Milk fat percentage was the target trait because

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

two mutations in the *DGAT1* (rs109326954, pA232K) and *GHR* (rs385640152, p.F279Y) genes have large effects on the milk composition in many cattle breeds including FV. In 249 sequenced FV animals, the minor allele frequencies of rs109326954 and rs385640152 were 6.2 and 7.2%. Both variants were more frequent in the multi-breed reference population (17.3 and 12.1%, Table 3). Depending on the reference population and imputation algorithm used, the proportion of correctly imputed genotypes was between 99.3 and 99.9% and between 86.1 and 99.4% for rs109326954 and rs385640152, respectively (Table 3). The imputation error rates were lower with *Minimac* than *FImpute*. Using *Minimac* and a multi-breed reference population provided the most accurate genotypes at both variants (99.9 and 99.4%).

Table 3 – Accuracy of imputation for two causal variants in Fleckvieh cattle

Chr	Position	NCBI identifier	effect	Minor allele frequency			Proportion of correctly imputed genotypes			
							Fimpute		*Minimac*	
				TaqMan	Within-breed	Multi-bred	Within-breed	Multi-breed	Within-breed	Multi-breed
14[a]	1,802,266	rs109326954	p.A232K	0.074	0.062	0.183	0.993	0.998	0.999	0.999
20	31,909,478	rs385640152	p.F279Y	0.074	0.072	0.121	0.861	0.931	0.868	0.994

The minor allele frequencies were calculated using TaqMan-derived and sequence genotypes in either within- or multi-breed reference populations.
[a] The p.A232K-variant in the *DGAT1* gene results from two adjacent SNPs in LD located at 1,802,265 (rs109234250) and 1,802,266 bp (rs109326954). In the present study, we considered only the variant at 1,802,266 bp.

Association tests between the imputed genotypes and milk fat percentage revealed that the *DGAT1*:p.A232K-variant was among the most significantly associated variants across all scenarios tested reflecting high accuracy of imputation regardless of the reference population and imputation algorithm used (Figure 2, Table 3). However, several adjacent variants (±90 kb) were in complete or near complete LD (r^2>0.99) with rs109326954 and had P values that were slightly higher, identical, or slightly lower. When the genotypes were imputed using *Minimac* and considering a within-breed reference population, five variants in complete LD including rs109326954 were the most significantly associated (Figure 2e,f).

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

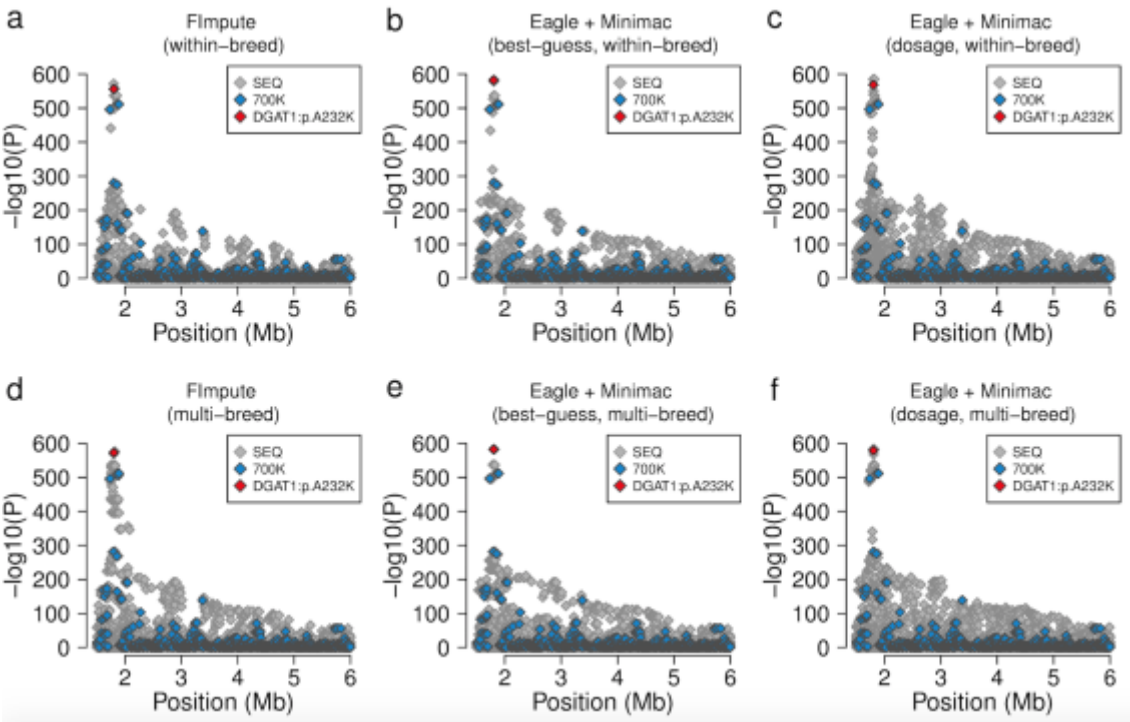

Figure 2 - Fine-mapping of a fat percentage QTL on bovine chromosome 14. (a-e) Association between 48,641 imputed sequence variants located at the proximal end of bovine chromosome 14 and milk fat percentage in 6958 Fleckvieh animals. Genotypes for the association studies were imputed using either *FImpute* (a, d) or *Minimac* (b, c, e, f) with either within- (a-c) or multi-breed reference populations (d-f). Grey and blue colours represent sequence and array-derived variants. The red symbol represents the p.A232K-mutation in the *DGAT1* gene.

The imputation algorithm and composition of the reference population had a large effect on the ability to detect an association between the *GHR*:p.F279Y-variant and milk fat percentage (Figure 3). The *GHR*:p.F279Y-variant was not significantly associated ($P>1x10^{-8}$) when the genotypes were imputed using *FImpute* or $Minimac_{BG}$ and a within-breed reference population reflecting high imputation error rates (Table 3). There were 350 and 395 variants detected, respectively, that had lower P values than rs385640152. Association tests with genotypes that were obtained using $Minimac_{DOS}$ (within-breed) or *FImpute* (multi-breed) revealed significant association of rs385640152 ($P=1.0x10^{-15}$, $3.0x10^{-10}$). However, six and 1089 variants had lower P values than rs385640152. When the genotypes were inferred using *Minimac* and a multi-breed reference population, rs385640152 was the most significantly associated variant reflecting higher accuracy of imputation (Figure 3e,f, Table 3).

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

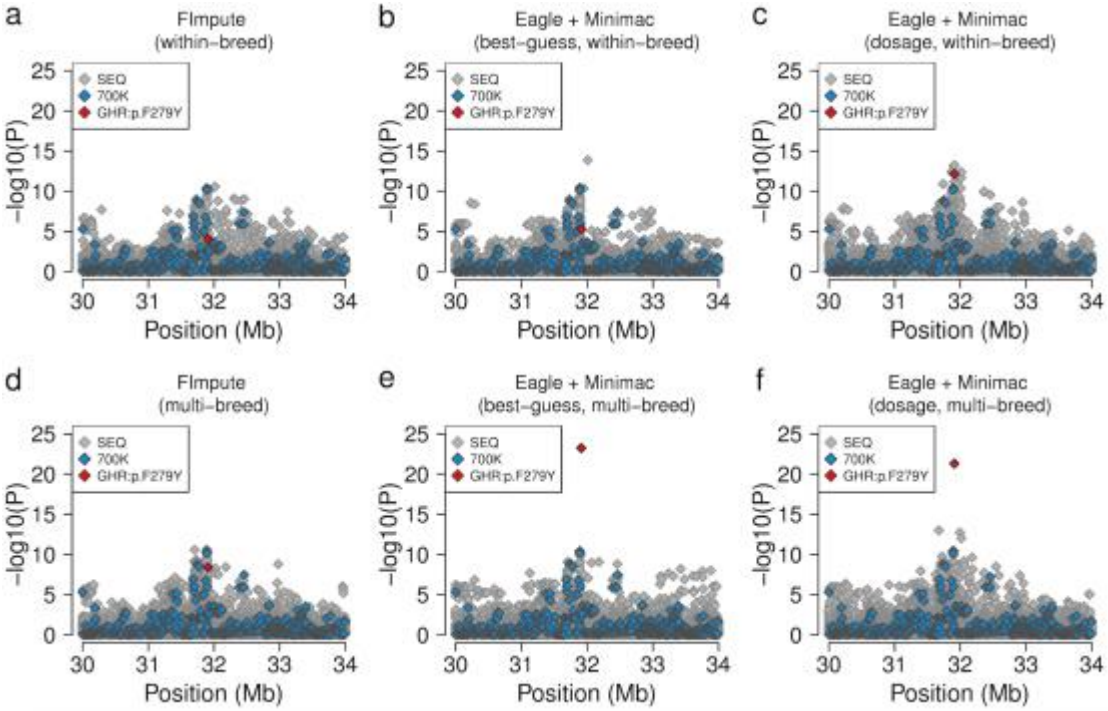

Figure 3 - Fine-mapping of a fat percentage QTL on bovine chromosome 20. (a-e) Association between 29,205 imputed sequence variants on bovine chromosome 20 and milk fat percentage in 6958 Fleckvieh animals. Genotypes for the association studies were imputed using either *FImpute* (a, d) or *Minimac* (b, c, e, f) with either (a-c) or multi breed reference populations (d-f). Grey and blue colours represent sequence and array-derived variants, respectively. The red symbol represents the p.F279Y-mutation in the *GHR* gene.

Fine-mapping of five fat percentage QTL with imputed sequence variants

The fine-mapping of two known QTL indicated that genotypes that are imputed using *Minimac* and considering a multi-breed reference population are an accurate source for association tests, particularly when predicted allele dosages rather than best-guess genotypes are used as explanatory variables. We imputed 23,256,743 sequence variants in 6958 FV animals using *Minimac* and a multi-breed reference population and performed association tests between imputed allele dosages and fat percentage. Our association study revealed seven QTL, including two QTL on BTA14 and BTA20 that encompassed the *DGAT1* and *GHR* genes (Figure 4a, Figure 2f, Figure 3f). The top variants were imputed sequence variants at all QTL.

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A total of 239 variants located between 91,857,670 and 93,955,207 bp on BTA5 were significantly associated ($P<2.1x10^{-9}$) with fat percentage (Figure 4b, see Additional file 7 Table S1). Thirteen variants in the first intron of the *MGST1* gene (*microsomal glutathione S-transferase 1*) were in high LD ($r^2>0.95$) with each other and had markedly lower P values ($P<6.5x10^{-21}$) than all other variants. These 13 variants were also associated with milk composition in New Zealand dairy cows [23]. The top variant (rs208248675) was located at 93,945,991 bp. The rs208248675 A-allele had a frequency of 19.4% in FV cattle and it decreased fat percentage.

A QTL for fat percentage on BTA6 encompassed the casein gene complex (Figure 4c, see Additional file 7 Table S1). One hundred and sixty-four variants with $P<2.1x10^{-9}$ were located between 87,084,144 and 87,296,017 bp. The most significantly associated variant (rs109193501 at 87,154,594 bp, $P=7.6x10^{-14}$) was located in an intron of the *CSN1S1* gene (*casein alpha s1*).

Two hundred and sixty-two variants located within a 50 kb interval (between 103,274,736 and 103,324,728 bp) on BTA11 were associated with fat percentage (Figure 4d, see Additional file 7 Table S1). The top variant (rs381989107 at 103,296,192 bp, $P=1.5x10^{-15}$) was located 5472 bp upstream of the translation start site of the *LGB* gene (*beta-lactoglobulin*, also known as *progestagen associated endometrial protein* (*PAEP*)). Two missense variants (p.G80D and p.V134A) in the *LGB* gene (rs110066229 at 103,303,475 bp and rs109625649 at 103,304,757 bp) that distinguish the LGB protein variants A and B [42] were in high LD with rs381989107 and had P values less than $4.5x10^{-12}$ (see Additional file 7 Table S1). A variant in the promoter of the *LGB* gene that causes an aberrant LGB expression in Brown Swiss cattle (OMIA 001437-9913) [43] was not polymorphic in the sequenced FV animals.

A QTL on BTA16 included 14 significantly associated variants located between 481,671 and 3,265,708 bp (see Additional file 7 Table S1). The top variant (rs380194132, $P=1.6x10^{-10}$) was located at 3,242,688 bp. Since the genomic region that included associated sequence variants with similar P values extended over almost 3 Mb, the identification of putative candidate gene(s) underlying the QTL was not attempted (see Additional file 3 Figure S3).

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

On BTA19, a single variant located at 51,380,692 bp was significantly associated (rs208289132, P=4.1x10^{-10}) with fat percentage (Figure 4e, see Additional File 7 Table S1). The associated variant was located 4230 bp upstream of the translation start site of the *FASN* gene (*fatty acid synthase*).

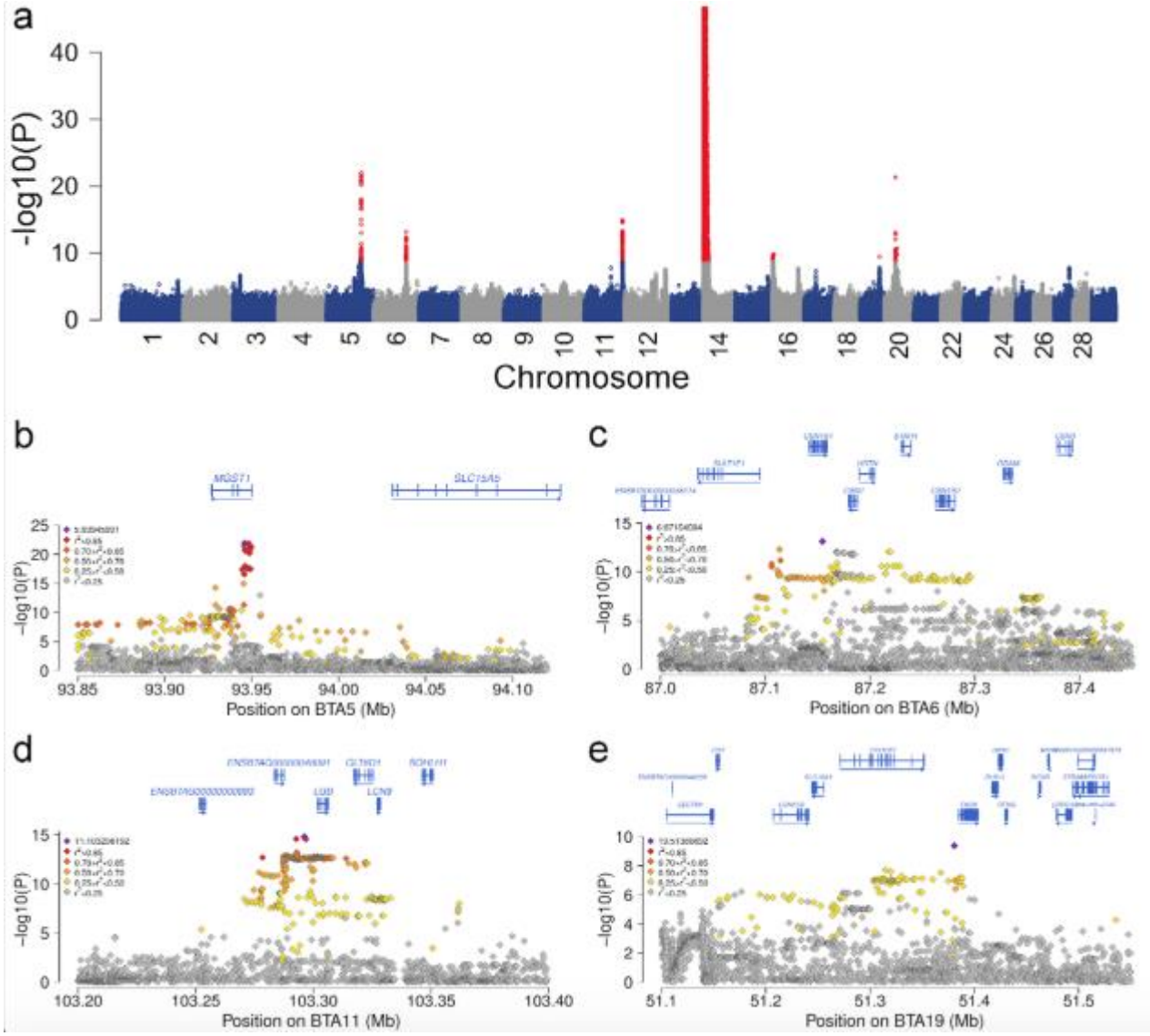

Figure 4 - Fine-mapping of fat percentage QTL in Fleckvieh cattle
(a) Manhattan plot representing the association of 23,256,743 imputed sequence variants with milk fat percentage in 6958 Fleckvieh animals. Red colours represent significantly associated variants (P<2.15x10^{-9}). The y-axis is truncated at –log10(10^{-40}). (b-e) Fine-mapping of four QTL for milk fat percentage on bovine chromosomes 5, 6, 11 and 19. Different colours represent the linkage disequilibrium between the most significantly associated variant (violet) and all other variants. Blue arrows indicate the direction of the gene transcription.

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Computing resources required

We assessed the computing resources required to impute genotypes at 23,256,743 sequence variants for 6958 target animals using 1577 sequenced animals as a reference. *Eagle* and *Minimac* were run on 10 processors per chromosome, whereas *FImpute* was run on a single processor. *FImpute* ran out of memory and failed to impute genotypes for chromosomes 12 and 23 likely because of large structural variants located on both chromosomes (see Additional file 4 Figure S4). The process (CPU) time to infer haplotypes and impute sequence variants for 27 chromosomes was 1037, 490 and 146 hours, respectively, for *Eagle*, *Minimac* and *FImpute*.

The wall-clock time and RAM usage for *FImpute* was between 2.9 and 9.6 hours and 6.7 and 29.7 gigabytes per chromosome, respectively (see Additional file 5 Figure S5). The estimation of haplotype phases took between 1.6 and 5.9 hours using *Eagle* and it required between 1.8 and 3.4 gigabytes of RAM with between 11,178 and 38,009 SNPs per chromosome for 6958 animals. The imputation of sequence variants using *Minimac* took between 2.5 and 9.4 hours and it required between 6.3 and 21 gigabytes of RAM per chromosome.

The estimation of haplotype phases for 1577 reference animals using *Eagle* (multi-threaded on 10 CPUs) took between 4.2 and 15.1 hours with between 413,371 and 1,444,299 sequence variants, respectively, and it required between 15.4 and 51.7 gigabytes of RAM per chromosome.

Discussion

We evaluated the accuracy of imputation from dense genotypes to sequence variants in 1577 sequenced animals which is a four- to eight-fold increase in sample size compared to previous studies in cattle [2, 17, 18]. Our findings are likely to be relevant for many cattle populations because we considered animals from two breeds with different effective population sizes as validation panel and inferred sequence variants on six chromosomes that reflect a broad spectrum of LD [44]. The accuracy of imputation was the correlation between imputed and sequenced genotypes ($r_{IMP,SEQ}$), which is a measure that depends less on the allele frequencies of the

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

sequence variants than *e.g.*, the proportion of correctly imputed genotypes [45]. However, the cross-validation procedure that was applied in our study yields reliable results only when the quality of the sequence data is high. Low fold sequencing coverage and sequencing or assembly errors may result in flawed genotypes [46]. Genotype imputation might infer true genotypes at such positions that differ from the sequenced genotypes thereby underestimating the accuracy of imputation. Considering that the genotype error rates were low in the sequenced FV and HOL animals [2, 47], it is unlikely that our findings are biased due to flawed genotypes at some sequence variants.

We assessed the performance of *FImpute* and *Minimac* because both methods can impute millions of polymorphic sites in large populations within a reasonable time. Previous studies showed that *FImpute* is more accurate than other tools with similar computing costs [48, 49]. The population-based genotype imputation algorithm implemented in *Minimac* is highly accurate and fast because it takes previously phased genotypes as input [11, 50]. We phased the target population using *Eagle* because it enables timely and accurate haplotype inference [32]. Although the computing costs to impute 23 million sequence variants in 6958 animals were more than ten-fold higher with *Eagle* and *Minimac* than *FImpute*, the wall-clock time was only 1.67-fold higher because *Eagle* and *Minimac* enable multi-threading to reduce the process time [32, 50]. The difference in computing costs is greater when haplotypes are not available for the reference animals. Since many variant detection pipelines include phasing and imputation algorithms [2, 47, 51], the computing time required to phase the sequence data was not considered in our study. However, the phasing of the reference panel was necessary because the variant detection pipeline of the 1000 bull genomes project was applied to successive 5 Mb segments rather than whole chromosomes [2].

The accuracy of imputation was higher with *Minimac* than *FImpute*, although *Minimac* does not consider pedigree information. Most individuals of our study were selected in a way that they represent a large proportion of the genetic variation of current populations [2, 52]. They are likely less related with each other than a random subset from the population, although some parent-offspring pairs were included. The accuracy of imputation might be higher when reference and target animals are closely

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

related particularly when *FImpute* is used [49]. However, dense genotypes for reference and target animals, *e.g.*, sequence and HD-derived genotypes, can be used to identify short shared haplotypes among apparently unrelated animals that originate from ancient ancestors possibly predating the separation of breeds [9, 53]. Including reference animals from various breeds increased the accuracy of imputation, particularly in FV cattle. The principal components analysis revealed that animals from many breeds clustered nearby the FV population indicating that they are distantly related whereas the HOL animals formed a cluster that was separated from all other breeds. The benefit of a multi-breed reference population on the accuracy of imputation was less pronounced in HOL cattle. Since the number of reference animals was nearly twice as high in HOL, our results may also corroborate the suggestion that multi-breed reference panels are particularly useful to impute genotypes when the reference population is small [18, 54, 55]. In agreement with previous studies in cattle [2, 18], a multi-breed reference population increased the accuracy of imputation at rare variants likely reflecting the shared ancestry of many cattle populations and the limited number of sequenced animals per breed. However, it seems advisable to periodically evaluate different imputation strategies because multi-breed reference populations can also compromise the imputation of sequence variants when the diversity of the reference panel increases [26].

The population-based imputation of sequence variants was particularly advantageous at rare alleles corroborating previous findings [12, 49, 56]. The true genotypes were even more correlated to predicted allele dosages than best-guess genotypes which is in line with the findings of Brøndum et al. [18]. An association study between allele dosages and milk fat percentage revealed that two rare (MAF<10%) alleles in the *DGAT1* and *GHR* [34-36] genes were the most significantly associated variants at two QTL on BTA14 and BTA20. This finding demonstrates that utilizing predicted allele dosages in association studies with imputed sequence variants can pinpoint causal mutations which agrees with Zheng at al. [57] and Khatkar et al. [58]. So far, imputed sequence variants did not substantially improve the accuracies of genomic predictions in real data which might partly result from flawed genotypes at low-frequency variants [17, 59]. We show that allele dosages are more accurate than best-guess genotypes particularly at rare variants because they better reflect imputation uncertainty. However, the analysis of allele dosages for millions of sequence variants

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

in tens of thousands of individuals for genomic predictions is computationally costly and has not been attempted so far [17, 60]. Since the proportion of rare alleles is more than three-fold higher in sequence than HD variants, further research is warranted to investigate if allele dosages may enhance genomic predictions with imputed sequence variants.

Although the overall accuracy of imputation was high, the $r_{IMP,SEQ}$–values varied considerably across chromosomes and dropped at several positions along the genome. Previous studies using array-derived genotypes showed that a decline in imputation accuracy may indicate intra- or inter-chromosomal misplacement of SNPs [12, 61, 62]. We detected low $r_{IMP,SEQ}$–values at regions where the bovine genome contains large segmental duplications [40, 41]. The population-wide imputation of sequence variants for chromosomes 12 and 23 was not possible using *Fimpute*. Assessing the accuracy of imputation along these chromosomes using cross-validation revealed large segments with low imputation quality on both chromosomes including the highly variable major histocompatibility complex [63]. We were able to eventually impute genotypes for BTA12 and BTA23 using *FImpute* when those segments were excluded from the reference panel. Imputing genotypes within both segments was possible using *Minimac*. However, the inferred genotypes are flawed and must be treated with caution as our results show (see Additional file 4 Figure S4). While an improved assembly of the reference genome may resolve some of these problems, a better resolution of large structural elements is not possible using short paired-end DNA sequencing [64]. Since flawed genotypes compromise downstream analyses [24, 25], it is advisable to exclude sequence variants within segments of low imputation quality from the target population and retain the array-derived genotypes only.

Our findings show that the composition of the reference population and the choice of the imputation algorithm are critical to infer accurate genotypes thereby enabling us to pinpoint causal mutations in association studies with imputed sequence variants. When the accuracy of imputation is low, causal variants may be “buried” among other variants in LD (*e.g.*, Figure 3a,d). In such situations, association testing will not reveal true causal mutations because variants with lower P values are likely to be prioritized as candidate causal variants [23]. However, true causal variants are not necessarily the

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

top variants in association studies even when their genotypes are almost perfectly imputed. Although the imputation accuracy for the causal p.A232K-polymorphism in the *DGAT1* gene was greater than 99.5%, several adjacent variants had identical or slightly lower P values likely indicating sampling errors [65] or synthetic associations [66, 67]. Nevertheless, the mutations in the *DGAT1* and *GHR* genes offer a convenient approach to evaluate the ability to detect causal mutations with imputed sequence variants in cattle, because both variants segregate in many breeds and phenotypes for fat percentage are readily available.

Our association study with more than 23 million sequence variants detected seven QTL that were significantly associated with milk fat percentage. Applying a less stringent significance threshold [68] would reveal additional QTL (see Additional file 6 Figure S6). Since we were able to pinpoint the causal mutations at two QTL, it is likely that causal mutations for other QTL are among the most significantly associated variants. However, the QTL identified in our study include several sequence variants in high LD and with similar P values rendering the identification of causal sites difficult. Association studies in animals from multiple breeds may facilitate differentiation between causal and non-causal sites in LD [2]. Causal sites may be located on ancient haplotypes that still segregate across multiple breeds as is the case for the fat percentage QTL on BTA5. Our association study revealed thirteen candidate causal variants in high LD in an intron of the *MGST1* gene that had markedly lower P values than all other variants. It is very likely that one of those variants is the actual causal polymorphism because they were also highly significantly associated with milk composition in another dairy cattle breed [23]. An improved functional annotation of the bovine genome and a fine-mapping strategy that incorporates biological information may facilitate to pinpoint causal mutations at such QTL [22, 69].

Conclusions

Genotypes for millions of sequence variants can be accurately inferred in large bovine cohorts using population-based imputation algorithms. The accuracy of imputation for rare alleles is higher with multi-breed reference populations. The accuracy of

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

imputation is not constant across chromosomes and may drop at regions where the genome contains large structural variants. Considering predicted allele dosages rather than best-guess genotypes as explanatory variables is beneficial for association studies with imputed sequence variants. The composition of the reference population and the choice of the approach used to infer genotypes are critical to detect causal mutations in association studies with imputed sequence variants.

Declarations

Ethics approval and consent to participate

Not applicable

Consent to publication

Not applicable

Availability of data and material

All relevant data are included in the manuscript and its supporting files

Competing interests

The authors declare that they have no competing interests.

Funding

HP was financially supported by a Deutsche Forschungsgemeinschaft (DFG) postdoctoral fellowship.

Authors' contributions

HP, IMM, HDD and MEG designed the experiments, HP analyzed the data, PJB provided support in high performance computing, RF and RE provided genotype and phenotype data, HP wrote the manuscript. All authors read and approved the final manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements

We thank Arbeitsgemeinschaft Süddeutscher Rinderzüchter und Besamungsorganisationen e.V. (ASR), Arbeitsgemeinschaft österreichischer Fleckviehzüchter (AGÖF) and Förderverein Bioökonomieforschung e.V. (FBF) for providing genotyping data. We acknowledge the 1000 bull genomes project for providing sequence variants.

References

1. Wiggans GR, Cooper TA, VanRaden PM, Van Tassell CP, Bickhart DM, Sonstegard TS. Increasing the number of single nucleotide polymorphisms used in genomic evaluation of dairy cattle. J Dairy Sci. 2016;99:4504–11.

2. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46:858–65.

3. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 2006;78:629–44.

4. Burdick JT, Chen WM, Abecasis GR, Cheung VG. In silico method for inferring genotypes in pedigrees. Nat Genet. 2006;38:1002–4.

5. Browning BL, Browning SR. A Unified Approach to Genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009;84:210–23.

6. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLOS Genet. 2009;5:e1000529.

7. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012;44:955–9.

8. Hickey JM, Kinghorn BP, Tier B, van der Werf JH, Cleveland MA. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet Sel Evol. 2012;44:9.

9. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15:478.

10. VanRaden PM, O'Connell JR, Wiggans GR, Weigel KA. Genomic evaluations with many more genotypes. Genet Sel Evol. 2011;43:10.

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

11. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11:499–511.

12. Pausch H, Aigner B, Emmerling R, Edel C, Götz KU, Fries R. Imputation of high-density genotypes in the Fleckvieh cattle population. Genet Sel Evol. 2013;45:3.

13. Browning BL, Browning SR. Genotype imputation with millions of reference samples. Am J Hum Genet. 2016;98:116–26.

14. Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, Thorleifsson G, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet. 2008;40:1068–75.

15. Druet T, Schrooten C, de Roos APW. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. J Dairy Sci. 2010;93:5443–54.

16. Zhang Z, Druet T. Marker imputation with low-density marker panels in Dutch Holstein cattle. J Dairy Sci. 2010;93:5487–94.

17. van Binsbergen R, Bink MC, Calus MP, van Eeuwijk FA, Hayes BJ, Hulsegge I, et al. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol. 2014;46:41.

18. Brøndum RF, Guldbrandtsen B, Sahana G, Lund MS, Su G. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics. 2014;15:728.

19. Qanbari S, Pausch H, Jansen S, Somel M, Strom TM, Fries R, et al. Classic selective sweeps revealed by massive sequencing in cattle. PLoS Genet. 2014;10:e1004148.

20. Pausch H, Emmerling R, Schwarzenbacher H, Fries R. A multi-trait meta-analysis with imputed sequence variants reveals twelve QTL for mammary gland morphology in Fleckvieh cattle. Genet Sel Evol. 2016;48:14.

21. Meuwissen T, Goddard M. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics. 2010;185:623–31.

22. MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ, et al. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics. 2016;17:144.

23. Littlejohn MD, Tiplady K, Fink TA, Lehnert K, Lopdell T, Johnson T, et al. Sequence-based association analysis reveals an MGST1 eQTL with pleiotropic effects on bovine milk composition. Sci Rep. 2016;6:25376.

24. Khatkar MS, Moser G, Hayes BJ, Raadsma HW. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. BMC Genomics. 2012;13:538.

25. Ertl J, Edel C, Emmerling R, Pausch H, Fries R, Götz K-U. On the limited increase in validation reliability using high-density genotypes in genomic best linear

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

unbiased prediction: Observations from Fleckvieh cattle. J Dairy Sci. 2014;97:487–96.

26. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3. 2011;1:457–70.

27. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at: http://arxiv.org/abs/1303.3997. Accessed at 4 July 2016.

28. Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, et al. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol. 2009;10:R42.

29. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

30. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4:7.

31. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82.

32. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016; advance online publication.

33. Pausch H, Kölle S, Wurmser C, Schwarzenbacher H, Emmerling R, Jansen S, et al. A nonsense mutation in TMEM95 encoding a nondescript transmembrane protein causes idiopathic male subfertility in cattle. PLoS Genet. 2014;10:e1004044.

34. Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, et al. 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 Res. 2002;12:222–31.

35. Winter A, Krämer W, Werner FAO, Kollers S, Kata S, Durstewitz G, et al. Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA:diacylglycerol acyltransferase (DGAT1) with variation at a quantitative trait locus for milk fat content. Proc Natl Acad Sci USA. 2002;99:9300–5.

36. Blott S, Kim J-J, Moisio S, Schmidt-Küntzel A, Cornet A, Berzi P, et al. 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. 2003;163:253–66.

37. Pausch H, Wurmser C, Reinhardt F, Emmerling R, Fries R. Short communication: Validation of 4 candidate causative trait variants in 2 cattle breeds using targeted sequence imputation. J Dairy Sci. 2015;98:4162–7.

38. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies.

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Nat Genet. 2010;42:348–54.

39. VanRaden PM. Efficient Methods to Compute Genomic Predictions. J Dairy Sci. 2008;91:4414–23.

40. Liu G, Ventura M, Cellamare A, Chen L, Cheng Z, Zhu B, et al. Analysis of recent segmental duplications in the bovine genome. BMC Genomics. 2009;10:571.

41. Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, et al. Copy number variation of individual cattle genomes using next-generation sequencing. Genome Res. 2012;22:778–90.

42. Ganai NA, Bovenhuis H, van Arendonk J a. M, Visker MHPW. Novel polymorphisms in the bovine beta-lactoglobulin gene and their effects on beta-lactoglobulin protein concentration in milk. Anim Genet. 2009;40:127–33.

43. Braunschweig MH, Leeb T. Aberrant low expression level of bovine beta-lactoglobulin is associated with a C to A transversion in the BLG promoter region. J Dairy Sci. 2006;89:4414–9.

44. Boitard S, Rodríguez W, Jay F, Mona S, Austerlitz F. Inferring population size history from large samples of genome-wide molecular data - an approximate bayesian computation approach. PLOS Genet. 2016;12:e1005877.

45. Hickey JM, Crossa J, Babu R, de los Campos G. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 2012;52:654.

46. Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok P-Y, Schaefer C, et al. Estimating genotype error rates from high-coverage next-generation sequence data. Genome Res. 2014;24(11):1734–9.

47. Jansen S, Aigner B, Pausch H, Wysocki M, Eck S, Benet-Pagès A, et al. Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. BMC Genomics. 2013;14:446.

48. Sun C, Wu X-L, Weigel KA, Rosa GJM, Bauck S, Woodward BW, et al. An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. Genet Res (Camb). 2012;94:133–50.

49. Ma P, Brøndum RF, Zhang Q, Lund MS, Su G. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. J Dairy Sci. 2013;96:4666–77.

50. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48:1284–7.

51. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.

52. Goddard ME, Hayes BJ. Genomic selection based on dense genotypes inferred

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

from sparse genotypes. Proc Adv Anim Breed Genet. 2009;18:26–9.

53. Schwarzenbacher H, Burgstaller J, Seefried FR, Wurmser C, Hilbe M, Jung S, et al. A missense mutation in TUBD1 is associated with high juvenile mortality in Braunvieh and Fleckvieh cattle. BMC Genomics. 2016;17:400.

54. Li H, Sargolzaei M, Schenkel F. Accuracy of whole- genome sequence genotype imputation in cattle breeds. In: Proceedings of the 10th World Congress on Genetics Applied to Livestock Production; 2014;191.

55. Bouwman AC, Veerkamp RF. Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. BMC Genetics. 2014;15:105.

56. Frischknecht M, Neuditschko M, Jagannathan V, Drögemüller C, Tetens J, Thaller G, et al. Imputation of sequence level genotypes in the Franches-Montagnes horse breed. Genet Sel Evol. 2014;46:63.

57. Zheng J, Li Y, Abecasis GR, Scheet P. A Comparison of approaches to account for uncertainty in analysis of imputed genotypes. Genet Epidemiol. 2011;35:102–10.

58. Khatkar MS, Thomson PC, Raadsma HW. Utility of imputed SNP genotypes for genome-wide association studies in dairy cattle. Proc Adv Anim Breed Genet. 2013;554–7.

59. Gonzalez-Recio O, Daetwyler HD, MacLeod IM, Pryce JE, Bowman PJ, Hayes BJ, et al. Rare variants in transcript and potential regulatory regions explain a small percentage of the missing heritability of complex traits in cattle. PLoS One. 2015;10:e0143945.

60. Wang T, Chen Y-PP, Bowman PJ, Goddard ME, Hayes BJ. A hybrid expectation maximisation and MCMC sampling algorithm to implement Bayesian mixture model based genomic prediction and QTL mapping. BMC Genomics. 2016;17:744.

61. Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012;95:4114–29.

62. Utsunomiya ATH, Santos DJA, Boison SA, Utsunomiya YT, Milanesi M, Bickhart DM, et al. Revealing misassembled segments in the bovine reference genome by high resolution linkage disequilibrium scan. BMC Genomics. 2016;17:705.

63. Emam M, Tabatabaei S, Sargolzaei M, Cartwright SL, Schenkel FS, Miglior F, et al. Evaluating the accuracy of imputation in the highly polymorphic MHC region of genome. J Anim Sci. 2016;94:174–5.

64. English AC, Salerno WJ, Hampton OA, Gonzaga-Jauregui C, Ambreth S, Ritter DI, et al. Assessing structural variation in a personal genome—towards a human reference diploid genome. BMC Genomics. 2015;16:286.

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

65. Barendse W. The effect of measurement error of phenotypes on genome wide association studies. BMC Genomics. 2011;12:232.

66. Bickel RD, Kopp A, Nuzhdin SV. Composite effects of polymorphisms near multiple regulatory elements create a major-effect QTL. PLoS Genet. 2011;7:e1001275.

67. Kühn C, Thaller G, Winter A, Bininda-Emonds ORP, Kaupe B, Erhardt G, et al. Evidence for multiple alleles at the DGAT1 locus better explains a quantitative trait locus with major effect on milk fat content in cattle. Genetics. 2004;167:1873–81.

68. Gao X, Becker LC, Becker DM, Starmer JD, Province MA. Avoiding the high Bonferroni penalty in genome-wide association studies. Genet Epidemiol. 2010;34:100–5.

69. Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW, et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biol. 2015;16:57.

Additional files

Additional file 1 Figure S1
File format: tif
Title: **Principal component analysis of 1577 sequenced animals.**
Description: (a-c) Plot of the top two principal components of the genomic relationship matrix. Different colours and symbols represent different breeds. The partners of the 1000 bull genomes consortium assigned the animals to breeds (b) Non-grey symbols indicate 249 animals that were considered as Fleckvieh animals. (c) Green symbols indicate 450 Holstein animals.

Additional file 2 Figure S2
File format: tif
Title: **Imputation accuracy along six chromosomes in Fleckvieh cattle.**
Description: (a-f) The correlation between true and imputed genotypes for sequence variants located within successive 1 Mb windows on six chromosomes. Different colours and symbols represent correlation coefficients obtained using different imputation algorithms and reference populations.

Additional file 3 Figure S3
File format: tif
Title: **Detailed view of a milk fat percentage QTL on bovine chromosome 16**
Description: Different colour represents the linkage disequilibrium between the most significantly associated variant (violet) and all other variants. Blue arrows indicate the direction of the gene transcription.

bioRxiv preprint doi: https://doi.org/10.1101/085399; this version posted November 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Additional file 4 Figure S4
File format: tif
Title: **Imputation accuracy along chromosomes 12 and 23 in Fleckvieh cattle.**
Description: The correlation between true and imputed genotypes for sequence variants located within successive 1 Mb windows on six chromosomes. Different colours and symbols represent correlation coefficients obtained using different imputation algorithms and reference populations. We were eventually able to impute sequence variants for BTA12 and BTA23 using *FImpute* when we discarded variants that were located between 70 and 77 Mb and between 25 and 30 Mb, respectively, from the reference panel.

Additional file 5 Figure S5
File format: tif
Title: **Computing resources required to impute 23,256,742 sequence variants in 6958 animals.**
Description: The wall-clock times (a) and random-access memory (RAM) usage (b) required to infer haplotypes and genotypes with *FImpute* (green), Eagle (dark blue) and *Minimac* (light blue) were assessed on 12-core Intel® Xeon® processors rated at 2.93 Ghz with 96 gigabytes of RAM. *FImpute* ran out of memory and did not finish when we attempted to infer genotypes for BTA12 and BTA23. *FImpute* was ran on a single processor whereas *Eagle* and *Minimac* used 10 processors per chromosome.

Additional file 6 Figure S6
File format: tif
Title: **Detailed view of a milk fat percentage QTL on bovine chromosome 27**
Description: Different colour represents the linkage disequilibrium between the top variant (violet) and all other variants. Blue arrows indicate the direction of the gene transcription. The top variant (36,211,258 bp) was associated with fat percentage ($P=1.9x10^{-8}$) albeit not at genome-scale. Twenty-two variants in high LD ($r^2>0.68$) with the top variant were located between 36,200,888 and 36,253,406 bp and had P values less than $7.7x10^{-7}$. Among those were three candidate causal variants (36,211,252 bp with $P=2.4x10^{-8}$, 36,211,708 bp with $P=2.8x10^{-8}$, 36,209,319 bp with $P=3.3x10^{-8}$) for fat content in the early lactation that were reported in Daetwyler et al. [2].

Additional file 7 Table S1
File format: tif
Title: **Significantly associated variants at five fat percentage QTL**
Description: Variants located on chromosomes 5, 6, 11, 16 and 19 with P values less than $2.1x10^{-9}$. The positions of the variants correspond to the UMD3.1 assembly of the bovine genome. The substitution effects (beta, standard error of beta) are given for the alternative allele.