ABSTRACT: Sequence data were generated from 157 animals of the Fleckvieh population. A pre-phasing based approach was used to impute genotypes for 21,045,178 polymorphic sites into 10,363 target animals genotyped with high-density arrays. Imputed sequence variants were used in an association study with daughter-derived phenotypes for milk fat percentage. The association study identified ten QTL controlling fat percentage in Fleckvieh cattle. Two postulated causal variants in the DGAT1 and GHR genes yielded the most significant association signals. Sequence-based association studies for udder conformation traits demonstrated a complex genetic architecture of mammary gland development in cattle. The association studies identified eight, six and seven QTL underlying udder depth, teat length and teat thickness, respectively. Imputed sequence variants captured genetic effects at a better resolution than array-based genotypes. However, even when considering large-scale imputed sequence variants, a significant fraction of the heritability remains “missing”.

Keywords: cattle, whole-genome sequencing, QTL, causal variants

Introduction

The availability of reference genomes facilitated to assess the genomic variation of populations (e.g., Venter et al. (2001); Van Tassell et al. (2008)). Resulting variant collections were the basis for high-throughput genotyping arrays for several species including cattle (e.g., Matsuzaki et al. (2004); Gunderson et al. (2005); Matukumalli et al. (2009)). After the implementation of genome-assisted selection in many cattle populations, at least the male breeding animals are nowadays routinely genotyped with dense arrays. Access to population-wide genotype data is unique in livestock populations and allows to obtain an individual's genetic merit without relying on phenotype information from relatives (Meuwissen et al. (2001); VanRadten et al. (2011)).

As another application, using dense genotypes in genome-wide association studies (GWAS) facilitates to pinpoint trait-associated regions. GWAS using high-density genotypes allow for the rapid identification of genomic regions underlying mendelian traits (Charlier et al. (2008)) and permit insights into the genetic architecture of complex traits (Hayes et al. (2010); Pausch et al. (2012b)).

GWAS in cattle populations revealed quantitative trait loci (QTL) for a wide range of phenotypes. However, only few causal variants (i.e., quantitative trait nucleotides (QTN), Long et al. (1998)) have been identified and validated so far (Weller and Ron (2011)). It turned out that, even with the availability of high-density genotypes, the fine-mapping of QTL remains a difficult task. QTL in livestock populations often extend over several million basepairs due to the long-range linkage disequilibrium (LD). Numerous positional candidate genes may be located within associated regions, rendering the identification of underlying genomic variants a difficult task. Pinpointing causal variants is especially complicated if regulatory sites control the trait variation (Karim et al. (2011); Visser and Goddard (2011)).

Exploiting population-wide sequence information would allow for most comprehensive GWAS. Causal variants could be tested for association with traits of interest rather than relying on anonymous markers in LD. However, to obtain sufficient power to identify QTL in GWAS requires large sample sizes. Whole-genome sequencing of the required number of individuals is currently cost-prohibitive. A reasonably cheap strategy is to sequence key animals explaining the vast majority of a population's genomic variation (Jansen et al. (2013); Daetwyler et al. (2014)). As most key animals are males that have been widely used in artificial insemination, a large proportion of the genes of current populations can be traced back to these individuals. Given that sequence data of a population's key ancestors is available, their descendants need only to be genotyped with dense arrays to infer their entire sequence with high confidence using genotype imputation (Scheet and Stephens (2006); Pausch et al. (2013)).

Here we report the imputation of whole-genome sequence information of 157 Fleckvieh animals into 10,363 target animals. We use genotypes for 21 million variants in large-scale association studies with daughter-derived phenotypes for milk fat percentage and udder conformation traits. We exemplify the suitability of imputed sequence variants for the identification of causal nucleotides based on two well described QTN. We demonstrate that imputed sequence variants explain more of the genetic variation than array-derived genotypes.

Materials and Methods

Generation of sequence data. Genomic DNA of 263 animals representing seven breeds was prepared from semen and blood samples, respectively, following standard protocols using proteinase K digestion and phenol-chloroform extraction. Of those, 157 are considered as key animals (Goddard and Hayes (2009)) of the current Fleckvieh population. DNA-concentration was set to 250 ng/μl. Paired-end libraries were prepared using the paired-end TruSeq DNA sample prep kit (Illumina inc., San Diego, CA, USA) and sequenced using the Illumina HiSeq 2500 instrument (Illumina inc., San Diego, CA, USA). The reads were processed with the Illumina Base Caller during the sequencing step. The alignment of the reads to the University of Maryland reference sequence (UMD3.1, Zimin et al. (2009)) was performed with BWA (version...
0.6.2, Li and Durbin (2009)) using default parameters. The resulting per individual SAM files were converted into BAM files with SAMtools (version 0.1.18, Li et al. (2009)). Duplicate reads were identified and marked with Picard tools (http://picard.sourceforge.net/).

**Variant calling.** Polymorphic sites including short insertions and deletions (Indels) were identified in 263 sequenced animals simultaneously using the multi-sample approach implemented in SAMTools' mpileup along with BCFTools. Beagle phasing and imputation (Browning and Browning (2009)) was used to improve the primary genotype calling by SAMtools. Sequence-derived genotypes were compared with array-derived genotypes to assess the quality of SAMtools' genotype calling. A detailed overview of the variant calling pipeline is given in Jansen et al. (2013).

**Genotyping animals of the target population.** The target animals were genotyped with medium and high-density arrays, respectively. Of the target population, 3,332 animals were genotyped with the Illumina BovineHD Bead chip comprising 777,962 SNPs. Another 7,073 animals were genotyped with the BovineSNP50 Bead chip comprising ~54,000 SNPs. After applying standard quality filters (see Pausch et al. (2014) for more details), the high density and the medium density dataset included genotypes for 652,856 and 42,907 SNPs, respectively. Genotype imputation was performed to extrapolate medium-density genotypes to higher density using a pre-phasing based approach. Haplotypes were inferred for the two datasets separately using Beagle. Subsequent haplotype-based imputation was performed with Minimac (Howie et al. (2012)). This multi-stage imputation approach provides high genotype quality at comparatively low computational costs in cattle populations (Pausch et al. (2013)). The imputed dataset comprised 10,363 animals with genotypes for 652,856 SNPs.

**Imputation of full sequence information.** We extracted sequence-derived genotypes for 21,045,178 polymorphic sites (19,086,476 SNPs, 1,958,702 Indels) that were segregating in 157 sequenced Fleckvieh animals. Genotypes were phased using Beagle. Thus obtained haplotypes were imputed into 10,363 target animals using Minimac.

**Evaluation of SNP-specific imputation accuracy.** Genotypes for polymorphic sites (N=15) were obtained in a large number of Fleckvieh animals using TaqMan® genotyping assays (Life Technologies) in order to validate putatively causal variants for economically important traits and mendelian disorders (e.g., Pausch et al. (2011); Pausch et al. (2014)). We compared thus obtained genotypes with imputed genotypes in up to 1,825 animals to assess the quality of imputed sequence variants.

**Animals and phenotypes.** Two missense mutations in the DGAT1 (Grisart et al. (2002); Winter et al. (2002)) and GHR genes (Blott et al. (2003)) are postulated to underly variation in bovine milk composition. As a proof of concept, we performed an association study with daughter-derived phenotypes (estimated breeding values, EBVs) for milk-fat percentage in 7,527 bulls with imputed sequence variants. The imputed sequence data includes the two postulated missense mutations. The reliability (r²) of the EBVs ranged from 0.51 to 0.99 with an average of 0.91. Given the high reliability, these EBVs should be suitable phenotypes for GWAS (Ekine et al. (2010)).

**Sequence-based association studies.** We used EMMAX (Kang et al. (2010)) to perform a single-marker based regression. The mixed model Y = Xb + u + e was fitted, where Y is a vector of daughter-derived phenotypes, b is the SNP effect, X is a design matrix of allele dosage data, u is a vector of additive genetic effects assumed to be normally distributed with mean 0 and (co)variance Gσ²a, with σ²a being the additive genetic variance and G is the realized genomic relationship matrix (GRM) built based on 635,224 autosomal SNPs (VanRaden (2008)) and e is a vector of random residual deviates ~N(0, I σ²e).

**Variance components estimation.** The pedigree relationship among 7,110 animals was used to estimate variance components for three udder conformation traits. Genomic relationship matrices (GRM) were built using three subsets of SNPs. A 50K and a 700K GRM were built using only SNPs interrogated with the 50K and the 777K array, respectively. A sequence-derived GRM was built using genotypes for 21,045,178 imputed sequence variants. The REML-algorithm implemented in GCTA (Yang et al. (2011a)) was used to estimate the variance components.

**Results and Discussion.**

**Sequence-derived genotype quality.** Multi-sample variant calling yielded genotypes at 25,426,490 sites. Among them, 21,045,178 were segregating in 157 Fleckvieh animals. The sequence-derived genotype quality was assessed in 133 Fleckvieh animals for which array-derived genotypes (777K) were available. We compared array-derived and sequence-derived genotypes for 39,327 SNPs located on bovine chromosome 1. The genotypic concordance averaged 96.19% and 99.39% before and after Beagle imputation, respectively (Figure 1). Beagle imputation improved the genotype quality especially in low-coverage sequence data which agrees with Jansen et al. (2013) and Daetwyler et al. (2014). Our results demonstrate that sequencing a significant number of animals at low to medium coverage followed by multi-sample variant calling and Beagle imputation provides high quality sequence-derived genotypes in cattle populations.

**Evaluation of imputation accuracy.** Sequence-derived genotypes for 21,045,178 variants were imputed into 10,363 target animals using a pre-phasing based approach. Imputation quality was assessed for 15 variants which have been genotyped in up to 1,825 animals using 5'-exonuclease assays (Table 1). Concordance between
imputed and genotyped sites averaged 97.3%. Imputation accuracy is clearly lower if the reference population’s allele frequency strongly deviates from the target population’s allele frequency. Imputation accuracy is likely to increase as the number of reference haplotypes increases (Pausch et al. (2013)). However, sequencing a highly informative subset of animals, i.e. animals with similar allele frequency spectrum as the target population, seems most crucial to obtain maximum imputation accuracy.

**Figure 1: Sequence-based genotype quality.** Array-based genotypes (777K) of 133 sequenced Fleckvieh animals were compared with the sequence-based genotypes before (orange) and after (blue) Beagle imputation.

**Table 1: Imputation accuracy.**

<table>
<thead>
<tr>
<th>Chr</th>
<th>bp</th>
<th>rsID</th>
<th>N</th>
<th>MAFp</th>
<th>MAFseq</th>
<th>CON</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>111129964</td>
<td>rs208582518</td>
<td>1130</td>
<td>0.04</td>
<td>0.04</td>
<td>99.0</td>
</tr>
<tr>
<td>6</td>
<td>88739941</td>
<td>rs109611915</td>
<td>1142</td>
<td>0.48</td>
<td>0.49</td>
<td>98.9</td>
</tr>
<tr>
<td>11</td>
<td>103304757</td>
<td>rs109625649</td>
<td>1525</td>
<td>0.45</td>
<td>0.45</td>
<td>98.6</td>
</tr>
<tr>
<td>14</td>
<td>1802265</td>
<td>rs109234250</td>
<td>925</td>
<td>0.08</td>
<td>0.06</td>
<td>99.7</td>
</tr>
<tr>
<td>14</td>
<td>23884989</td>
<td>rs135061217</td>
<td>812</td>
<td>0.1</td>
<td>0.01</td>
<td>80.5</td>
</tr>
<tr>
<td>14</td>
<td>24759178</td>
<td>rs132707603</td>
<td>824</td>
<td>0.02</td>
<td>0.01</td>
<td>99.3</td>
</tr>
<tr>
<td>14</td>
<td>24954981</td>
<td>rs108949025</td>
<td>815</td>
<td>0.17</td>
<td>0.11</td>
<td>97.5</td>
</tr>
<tr>
<td>19</td>
<td>27042848</td>
<td>rs381722524</td>
<td>763</td>
<td>0.1</td>
<td>0.07</td>
<td>99.6</td>
</tr>
<tr>
<td>19</td>
<td>27570146</td>
<td>rs385135118</td>
<td>753</td>
<td>0.13</td>
<td>0.08</td>
<td>93.9</td>
</tr>
<tr>
<td>19</td>
<td>27689622</td>
<td>rs378652941</td>
<td>1644</td>
<td>0.08</td>
<td>0.07</td>
<td>99.6</td>
</tr>
<tr>
<td>19</td>
<td>27956790</td>
<td>rs383770500</td>
<td>1737</td>
<td>0.01</td>
<td>0.02</td>
<td>98.4</td>
</tr>
<tr>
<td>20</td>
<td>31909478</td>
<td>rs385640152</td>
<td>722</td>
<td>0.07</td>
<td>0.08</td>
<td>95.7</td>
</tr>
<tr>
<td>20</td>
<td>39115344</td>
<td>rs135164815</td>
<td>546</td>
<td>0.27</td>
<td>0.24</td>
<td>99.5</td>
</tr>
<tr>
<td>21</td>
<td>71001232</td>
<td>rs378824791</td>
<td>1653</td>
<td>0.01</td>
<td>0.02</td>
<td>99.9</td>
</tr>
<tr>
<td>29</td>
<td>43599204</td>
<td>rs385444696</td>
<td>1825</td>
<td>0.06</td>
<td>0.06</td>
<td>99.9</td>
</tr>
</tbody>
</table>

Chromosomal position (Chr, basepairs (bp)) of 15 variants for which genotypes were obtained with 5’-exonuclease assays. The minor allele frequency is presented based on genotyped (MAFp) and sequence-derived (MAFseq) genotypes for up to 1,825 animals (N). The quality of imputed sequence variants is given as the proportion of concordant genotypes (CON).

**Sequence-based association study for fat percentage.** To investigate the suitability of imputed sequence data for causal variant detection, we performed an association study using 21,045,178 imputed variants with estimated breeding values for milk fat percentage in 7,527 Fleckvieh bulls. Red symbols indicate genome-wide significantly associated variants. The y-axis is truncated at 40. Close-up of two QTL-regions on BTA14 (B) and BTA20 (C). Black and orange symbols represent markers of the 700K-panel and imputed sequence variants, respectively. The red arrows indicate two postulated causal missense mutations in the DGAT1 and GHR genes, respectively.

**Figure 2: Sequence-based association study for milk fat percentage.** Association of 21,045,178 imputed variants with estimated breeding values for milk fat percentage in 7,527 Fleckvieh bulls (A). Red symbols indicate genome-wide significantly associated variants. The y-axis is truncated at 40. Close-up of two QTL-regions on BTA14 (B) and BTA20 (C). Black and orange symbols represent markers of the 700K-panel and imputed sequence variants, respectively. The red arrows indicate two postulated causal missense mutations in the DGAT1 and GHR genes, respectively.

The association study identified significantly associated variants nearby MGST1, CSN1S2, PAEP, FASN and AGPAT6 indicating that these genes might control milk composition in Fleckvieh cattle. These regions coincide with QTL for milk composition traits identified in several cattle breeds (Wang et al. (2012); Caroli et al. (2009), Braunschweig and Leeb (2006); Bouwman et al. (2011)). The positions of the most significantly associated variants indicate that regulatory sites might cause FP variation at these QTL. Investigating the effects of such sites on gene
expression might provide insights into the biological mechanisms underlying FP variation (Lappalainen et al. (2013)). However, identifying causal variants in non-coding elements awaits a better functional annotation of genomes.

**Sequence-based association studies for udder conformation traits.** Udder conformation traits represent an important information source in cattle populations as they are highly predictive for longevity (Larroque and Ducrocq (2001); Vollema and Groen (1997)). Sequence-based association studies in 7,110 animals identified eight, six and seven QTL for udder depth (UD), teat length (TL) and teat thickness (TT) (Figure 3). Only three QTL were associated with more than one phenotype. The multitude of QTL for udder conformation traits indicates a very complex genetic architecture underlying mammary gland development in cattle. There was little overlap between the QTL identified in the present study and QTL controlling udder conformation traits in other cattle breeds (Wu et al. (2013); Cole et al. (2011)). A QTL for teat morphology on BTA17 is nearby TBX3, TBX5 and RBM19. This QTL coincides with a region causing mammary gland malformations in cattle and other species (Pausch et al. (2012a); Ihara et al. (2007)).

![Figure 3: Sequence-based association studies for udder conformation traits.](image)

**Figure 3: Sequence-based association studies for udder conformation traits.** Association of 21,045,178 imputed variants with daughter yield deviations for udder depth (A), teat length (B) and teat thickness (C) in 7,110 Fleckvieh bulls. Red symbols indicate genome-wide significantly associated variants.

The GWAS for UD revealed strong association in immediate vicinity to PLAG1 (Figure 3A). Variation affecting the expression of PLAG1 primarily controls stature and other growth-related traits in cattle (Karim et al. (2011); Pausch et al. (2011)). Phenotypes for UD, i.e. the interspace between ankle and udder base, are visually examined. We hypothesise that the interspace might be overestimated in tall animals. We assume that the association of the PLAG1 region reflects variation in body size rather than pleiotropic effects on UD, although pleiotropism has been reported for that QTL (Fortes et al. (2013)).

Two QTL for UD and TT indicate members of the ADAM-family (ADAMTS3, ADAM12) as underlying genes for udder conformation traits. ADAMs are transmembrane proteins that contain a disintegrin and metalloprotease domain (Primakoff and Myles (2000)). Mutations in genes encoding ADAMs cause connective-tissue disorders in various species (Fernandes et al. (2001)). As a deep udder base in cattle is the result of a weakness of the connective tissue, genes encoding ADAMs are excellent candidates to harbour variants affecting UD.

**Genetic variation explained by imputed sequence variants.** The fraction of genetic variation explained increased as the marker density increased (Table 2). Compared to the 50K subset, the GRM built upon 700K and 21M genotypes additionally explained 3.7% and 4.5% of the DYD variation, respectively. These results agree with findings in Nordic Holstein (Jensen et al. (2012)) and Japanese cattle (Ogawa et al. (2014)). Denser marker panels seem to capture genetic effects at a better resolution implying that the accuracy of genomic prediction slightly increased validation reliability in Fleckvieh cattle (Ertl et al. (2014)).

![Table 2: Proportion of variation explained.](image)

<table>
<thead>
<tr>
<th>Trait</th>
<th>Ped</th>
<th>50K</th>
<th>700K</th>
<th>Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>UD</td>
<td>0.915</td>
<td>0.702 (0.808)</td>
<td>0.743 (0.812)</td>
<td>0.751 (0.821)</td>
</tr>
<tr>
<td>TL</td>
<td>0.908</td>
<td>0.730 (0.804)</td>
<td>0.769 (0.847)</td>
<td>0.780 (0.859)</td>
</tr>
<tr>
<td>TT</td>
<td>0.811</td>
<td>0.652 (0.804)</td>
<td>0.682 (0.841)</td>
<td>0.687 (0.847)</td>
</tr>
</tbody>
</table>

The proportion of variance explained was estimated for udder depth (UD), teat length (TL) and teat thickness (TT) using a pedigree-based (Ped), a 50K SNP-based (50K), a 700K SNP-based (700K) and a sequence-based (Seq) relationship matrix. Values in brackets indicate the proportion of heritability explained.

The missing heritability amounts to 17.9%, 14.1% and 15.3% for UD, TL and TT, respectively, using imputed genotypes for 21 million variants. Although these values are clearly lower than those obtained for complex traits in humans (Yang et al. (2011b)), a substantial fraction of the genetic variation still remains unexplained. We assume that part of the missing heritability is attributable to low-frequency variants that were missed due to our key ancestor sequencing strategy. Additionally, there might be a significant number of variants with poor imputation quality compromising genomic relationship coefficients. Further research is necessary to quantify different elements of missing heritability in whole-genome sequence data in cattle populations.

**Conclusion**

Whole genome sequence data of a population's key ancestors provides a valuable source for genome-wide analyses of complex traits. Using sequence data of 157 key
animals of the Fleckvieh breed as reference population facilitated to accurately extrapolate 21 million variants for more than 10,000 animals. Using imputed sequence variants in large scale association studies allows to fine-map QTL at maximum resolution. Causal variants are in the data and can immediately be tested for association in large cohorts. However, identifying causal variants in non-coding elements remains a difficult task and awaits a better functional annotation of genomes. Imputed sequence variants explain more of the genetic variation of complex traits than array-based genotypes. However, even when using large-scale imputed sequence data, a substantial fraction of the heritability remains missing.

Acknowledgments

This study was funded by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr “Synbreed” - Synergistic plant and animal breeding (grant-ids: 0315528A and 0315527B). We acknowledge Qualitas AG (CH-Zug) and the Swiss Cattle Breeder Association (ASR) for making sequence data of animals of the Fleckvieh breed as reference population kindly provided by ZuchtData (A-Vienna).

Literature Cited