1-step versus 2-step imputation: a case study in German Black Pied cattle

P. Korkuć[1], D. Arends[1], C. Scheper[2], K. May[2], S. König[2], G. A. Brockmann[1]

[1]Humboldt University Berlin, Albrecht Daniel Thaer- Institute for Agricultural and Horticultural Sciences, Animal Breeding Biology, and Molecular Genetics, Invalidenstr. 42, 10115 Berlin, Germany, paula.korkuc@hu-berlin.de (Corresponding Author: paula.korkuc@hu-berlin.de)
[2]Justus-Liebig University Gießen, Institute for Animal Breeding and Domestic Animal Genetics, Ludwigstr. 21 b, 35390 Gießen, Germany

Summary

Since the genotyping of cattle is usually performed using low or medium density SNP chips, the *in silico* imputation is a required step to generate genotypes on the whole-genome level using high density and whole-genome sequencing data.

Here, we investigate different imputation strategies using the recently available sequencing data from the 1000 bull genomes project including 30 sequenced DSN cattle (German Black Pied cattle, German: "Deutsches Schwarzbuntes Niederungsrind") that we have contributed. Our investigation compares a 1-step *versus* a 2-step imputation approach using the imputation software tool Beagle. In the 1-step approach, 50k genotypes are directly imputed to the level of whole-genome sequencing data. The 2-step approach differs by first imputing 50k genotypes to 700k and subsequently from 700k to sequence level. Additionally, we investigate the imputation accuracy with respect to different reference population sizes and composition. These are 1) only DSN cattle, 2) DSN and Holstein cattle and 3) all *Bos taurus* cattle from the 1000 bull genomes project. The imputation accuracy was assessed as relative Manhattan distance in a leave-one-out cross validation using our 30 sequenced DSN cattle as targets for imputation.

Imputation with Beagle showed increased performance with increasing population sizes of the reference population, however a significant drop in imputation performance was observed when imputing using a smaller reference population consisting of breeds highly related to DSN compared to a larger reference population of unrelated breeds. Both size and composition play an important role in imputation accuracy. Furthermore, when using a small reference population in the first round (from 50k to 700k), we observed lower imputation accuracies of the 2-step approach compared to the 1-step approach. However, using a 'big enough' reference population in the first round restored imputation accuracies of the 2-step approach *versus* the much simpler 1-step approach. Our hypothesis is that when a limited reference population is available the 2-step approach leads to lower accuracy of imputation because imputation errors in the first round propagate to the second round of imputation.

Keywords: Cattle, DSN, Holstein, sequencing, 1000 bull genomes project, SNP, SNP chip, imputation, Beagle

Introduction

In our research, we focus on the German Black Pied cattle breed (DSN, German: “Deutsches Schwarzbuntes Niederungsrind”), which is considered to be one of the founding breeds of Holstein Friesian (HF) cattle. Recently, we sequenced 30 DSN cattle, and contributed the data to the 1000 bull genomes project.

Since cattle are usually genotyped using low (3k or 10k) or medium (50k) density SNP chips, the *in silico* imputation from low or medium density SNP chip genotypes to the level of whole-genome sequencing data is a practical, cheap and fast method to generate high density genotypes for many individuals. However, the reliability of imputed genotypes has to be sufficiently high to improve the accuracy when performing a genome-wide association study. Previous research showed that imputation reliability is improved when the number of individuals in the reference population is increased and if the reference population consists of close relatives of the to be imputed individuals (van Binsbergen et al. 2014; Pausch et al. 2017). Thus, imputation studies are often performed in breeds such as HF, where a high number of sequenced individuals is available, which serve as a reference for imputation in the same breed (van Binsbergen et al. 2015; Pausch et al. 2017; Brøndum et al. 2014).

Furthermore, stepwise imputation (2-step imputation) was reported to be more accurate than direct imputation of genotypes (van Binsbergen et al. 2014; VanRaden et al. 2013). Stepwise imputation is an imputation strategy which imputes from the original density to an intermediate higher density, after which imputation is performed up to the requested density.

In this paper, we explore multiple strategies of imputation with regards to the size and composition of the reference population. Furthermore, 1-step imputation was compared to 2-step imputation. For the study, we used the recently available data from the 1000 bull genomes project including 30 sequenced DSN cattle, and 48 DSN cattle genotyped with a 700k SNP chip.

Material and methods

Genotypes of DSN and reference population

For this project, whole-genome sequencing data of 30 DSN cattle was available. In addition, 48 DSN cattle were genotyped with the Illumina® BovineHD Genotyping BeadChip (700k SNP chip). Since no genetic data was available, cows were selected from different herds and the available artificial insemination bulls were chosen. Thus, DSN animals that were sequenced as well as the DSN animals that were genotyped with the 700k SNP chip were not unrelated. Due to the limited population size of DSN which is about 2000 animals, family relationships exist between individuals.

Whole-genome sequencing data from *Bos taurus* animals was provided by the 1000 bull genomes project (Run 6.0) consisting of 2,333 animals including our 30 DSN cattle. Alignment, SNP calling, and quality control were done as described in Daetwyler et al. 2014. Animals with no breed specification, belonging to crossbreeds or to breeds with less than 10 individuals were removed from the dataset. Relative Manhattan distance between all individuals was calculated as a measurement of genetic similarity (see below, Equation 1). We noticed animals that showed an unusually high amount of genetic similarity (relative Manhattan distance > 0.99) which were subsequently removed from the dataset. The resulting dataset contains 2,145 cattle from 30 breeds, including 541 HF and the 30 DSN cattle mentioned earlier. For the analysis, only SNPs were used that were polymorphic (heterozygous or homozygous for the alternative allele) in at least one DSN animal.

To generate a dataset for imputation, the 30 sequenced DSN cattle were scaled down to the level of the Illumina® Bovine50SNP chip (50k SNP chip). Similarly, the whole-genome sequencing genotypes of the 1000 bull genomes project were scaled down to the level of the 700k SNP chip in order to generate reference populations for the imputation from 50k to 700k in the 2-step approach.

To obtain correct genome positions for each SNP, SNP chip probe sequences of the 50k and 700k SNP chips (obtained from Illumina) were remapped against the *Bos taurus* genome version UMD3.1 using blastn. Probes for SNPs that did not map to a single genomic location were excluded from further analyses. SNPs that did not yield genotype calls in at least 95% of animals were removed. In total, 49,106 SNPs were left in the 50k dataset and 602,587 SNPs in the 700k dataset.

Imputation strategies

Imputation of genotypes was done using either a 1-step or a 2-step approach. In the 1-step approach, the scaled down 50k genotypes were directly imputed to sequence level using whole-genome sequencing data. In the 2-step approach, the scaled down 50k genotypes were first imputed to 700k and subsequently imputed to sequence level using whole-genome sequencing data.

Reference populations differing in size and composition were evaluated. For the 1-step imputation from 50k to sequence level, three reference populations were generated which are composed of 1) DSN cattle (30 sequenced individuals), 2) DSN and HF cattle (30 + 541 sequenced individuals) and 3) all *Bos taurus* cattle from the 1000 bull genomes project (2145 sequenced individuals which include DSN and HF) (Figure 1). The same reference populations were used for the imputation from 700k to sequence level in the 2-step approach. The downscaled 700k genotypes were used together with the genotypes of the 48 DSN, which were

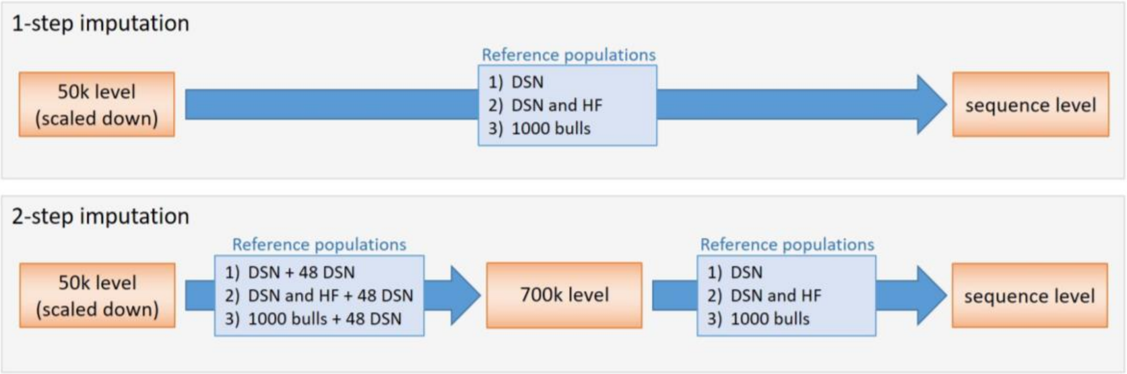

Figure 1: Imputation strategies for the 1- and 2-step imputation approaches showing the reference populations used.

available from the 700k SNP chip, as reference populations for the first step (50k to 700k) in the 2-step imputation (Figure 1). The 50k data used as the target set for imputation was unphased. The output of the 50k to 700k imputation of the 2-step imputation approach was used as the input for the imputation from 700k to sequence level. All reference populations were phased using Beagle.

Imputation was performed using Beagle (version 4.1) (Browning and Browning 2016) with the default settings. The accuracy of imputation was assessed based on the 30 sequenced DSN cattle using a leave-one-out cross validation. That means that in each imputation round the individual, which should be imputed was left out of the reference population. The imputed genotypes of this individual were then compared to the known genotypes from whole-genome sequencing. We define the accuracy of imputation by calculating the relative Manhattan distance *d* as:

$$d_{obs,imp} = 1 - \frac{\sum_{i=1}^{n} |obs_i - imp_i|}{2n}, \quad (1)$$

where *obs* is the observed variant and *imp* the imputed variant. Genotypes were coded as 0, 1, and 2 corresponding to genotypes homozygous to the reference allele, heterozygous, and homozygous to the alternative allele. *n* corresponds to the number of variants at sequence level. Accuracy values range between 0, all SNPs are different, and 1, all SNPs are identical between observed and imputed variants.

To reduce the computational burden the leave-one-out cross validation was performed only for chromosomes 1 through 5 of the *Bos taurus* genome consisting of 29 autosomal chromosomes. However, similar results are to be expected for the other chromosomes.

Results

Comparison of imputation strategies

The best median imputation accuracy (89.8%) was observed in the 2-step imputation approach with the 1000 bull genomes dataset as reference for both imputation rounds (from 50k to 700k and from 700k to sequence level) (Figure 2). Also for the 1-step imputation approach, the best accuracy was observed when the 1000 bull genomes dataset was used as the reference (89.5%). The use of only DSN or DSN together with HF cattle as a reference population in the imputation from 50k to 700k level in the 2-step imputation approach significantly reduces the accuracy of imputation, even when using the *Bos taurus* cattle of the 1000 bull genomes data for the subsequent imputation from 700k to sequence level (Figure 2). The use of only DSN or DSN together with HF cattle in the 1-step approach led also to much lower imputation accuracy. We can conclude that (when using Beagle) increasing the number of individuals in the reference population leads to increased imputation accuracy, independent of the composition of the reference population.

From the literature, the 2-step approach has been suggested to provide more accurate imputation compared to the 1-step approach (van Binsbergen et al. 2014; VanRaden et al. 2013). However, our results show that if only a small number of individuals is available for the first round of imputation (from 50k to 700k) the overall imputation accuracy (from 50k to sequence level) is impaired. Our hypothesis is that this loss in accuracy results from erroneous genotype calls in the first imputation round that are carried over into the second round of imputation.

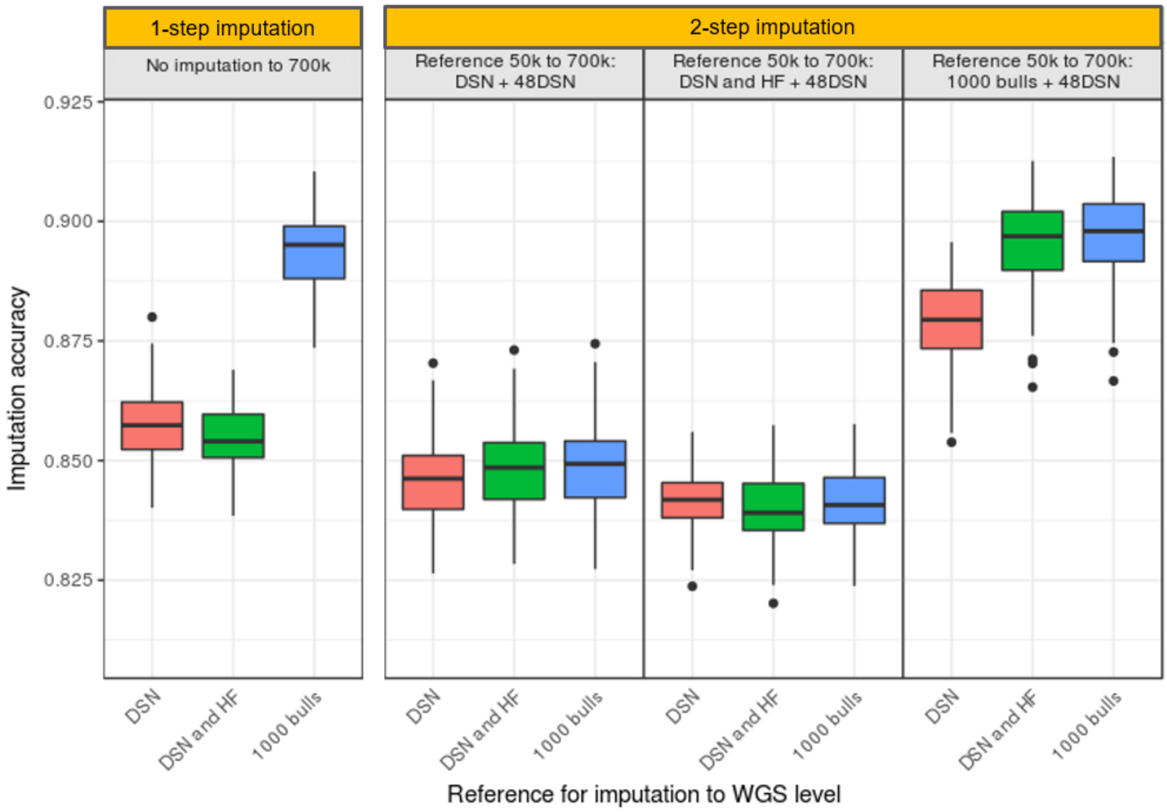

Figure 2: Comparison of imputation accuracies of the 1- to 2-step imputation from 50k to sequence level using Beagle with regard to different reference populations. We observe increased imputation accuracies with an increasing number of individuals in the reference population.

Conclusions

In this paper, we explored the differences between two approaches for the imputation of 50k genotypes to sequence level using diverse sizes and composition of the reference population for imputation.

Imputation with Beagle showed a significant drop in imputation accuracy when using a small reference population consisting of breeds highly related to the target breed compared to using a larger reference population of unrelated breeds. Furthermore, imputation with Beagle is best performed using as many individuals as possible in the reference population.

When using a ‘big enough’ reference population, we did not observe an improvement of the 2-step approach *versus* the much simpler 1-step approach. However, when only a limited reference population is available, it seems that the 2-step approach leads to lower accuracy of imputation because errors in the first round propagate to the second round of imputation.

Acknowledgments

The authors acknowledge the Rinderzucht Berlin-Brandenburg (RBB) for cooperation and providing blood, semen, and ear punch tissue samples from DSN cattle.

Bundesministerium
für Ernährung
und Landwirtschaft

This project is funded by the German Federal Ministry of Food and Agriculture (BLE) and Federal Program of Ecological Agriculture (BÖLN) (Funding number: 2815NA010).

BÖLN
Bundesprogramm Ökologischer Landbau
und andere Formen nachhaltiger
Landwirtschaft

List of references

Binsbergen, R. van, M. Bink, M. Calus, F. van Eeuwijk, B. Hayes, I. Hulsegge, and R. Veerkamp. 2014. "Accuracy of Imputation to Whole-Genome Sequence Data in Holstein Friesian Cattle." *Genetics Selection Evolution* 46 (1): 41. doi:10.1186/1297-9686-46-41.

Binsbergen, R. van, M. Calus, M. Bink, F. van Eeuwijk, C. Schrooten, and R. F. Veerkamp. 2015. "Genomic Prediction Using Imputed Whole-Genome Sequence Data in Holstein Friesian Cattle." *Genetics Selection Evolution* 47 (1). Genetics Selection Evolution: 71. doi:10.1186/s12711-015-0149-x.

Brøndum, R., B. Guldbrandtsen, G. Sahana, M. Lund, and G. Su. 2014. "Strategies for Imputation to Whole Genome Sequence Using a Single or Multi-Breed Reference Population in Cattle." *BMC Genomics* 15 (1): 728. doi:10.1186/1471-2164-15-728.

Browning, B. L. L., and S. R. R. Browning. 2016. "Genotype Imputation with Millions of Reference Samples" 98 (1). Cell Press: 116–26. doi:10.1016/j.ajhg.2015.11.020.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, Rasmus F Brøndum, Xi. Liao, et al. 2014. "Whole-Genome Sequencing of 234 Bulls Facilitates Mapping of Monogenic and Complex Traits in Cattle." *Nature Genetics* 46 (8). Nature Publishing Group: 858–65. doi:10.1038/ng.3034.

Pausch, H., I. MacLeod, R. Fries, R. Emmerling, P. Bowman, H. D. Daetwyler, and M. Goddard. 2017. "Evaluation of the Accuracy of Imputed Sequence Variant Genotypes and Their Utility for Causal Variant Detection in Cattle." *Genetics Selection Evolution* 49 (1). BioMed Central: 24. doi:10.1186/s12711-017-0301-x.

VanRaden, P, D. Null, M Sargolzaei, G. Wiggans, M. Tooker, J. Cole, T. Sonstegard, et al. 2013. "Genomic Imputation and Evaluation Using High-Density Holstein Genotypes." *Journal of Dairy Science* 96 (1): 668–78. doi:10.3168/jds.2012-5702.