Imputation in Swiss Cattle Breeds

B. Gredler[1], F. R. Seefried[1], U. Schuler[1], B. Bapst[1], U. Schnyder[1,2], J. M. Hickey[3]
[1]Qualitas AG, Chamerstrasse 56, 6300 Zug, Switzerland
[2]swissherdbook cooperative, Schuetzenstrasse 10, 3052 Zollikofen, Switzerland
[3]School of Environment and Rural Science, University of New England, Armidale, NSW, 2351, Australia
birgit.gredler@qualitasag.ch

Abstract

Imputation from Illumina 3k to 54k was carried out using AlphaImpute and Findhap V2 for Swiss cattle breeds. Genotypes of Original Braunvieh and Brown Swiss were combined on the one hand, and on the other hand Simmental, Swiss Fleckvieh and Holstein (mixture data set). Accuracy of imputation was slightly better for the Brown Swiss data set than for the mixture data set. AlphaImpute outperformed Findhap V2 when close relatives of imputation candidates were 54k-genotyped. Findhap V2 resulted in higher imputation accuracy when candidates were less related to 54k reference animals.

Introduction

Genomic selection can be more cost effective when Single Nucleotide Polymorphisms (SNP) panels are available at different densities and prices. Low-density SNP panels can make genotyping of thousands of cows affordable, for example. Missing SNP on the low-density chip can be filled using imputation, given a reasonable number of animals is genotyped for a high-density panel, which builds the reference population for imputation. Very likely, it will be possible to impute very high-density panels or even whole genome re-sequence data from medium SNP panels. Imputation methods make use of information on linkage (e.g Habier *et al.*, 2009) or linkage disequilibrium among SNP (e.g. fastPHASE; Sheet and Stephenson, 2006), while ignoring information from known pedigrees. The imputation program AlphaImpute (Hickey *et al.*, 2011a) makes use of linkage in combination with pedigree information. The software Findhap V2 (Van Raden *et al.*, 2011) combines family with population haplotyping. Both AlphaImpute and Findhap V2 are expected to be appropriate for imputation in dairy cattle populations.

The objective of this study was to compare the accuracy of genotype imputation from Illumina 3k to 54k SNP chip information in Swiss cattle breeds.

Material

Genotypes of Original Braunvieh (OB), Brown Swiss (BS), Swiss Fleckvieh (SF), Simmental (SI) and (Red) Holstein (HO) were used in this study. Swiss Fleckvieh is a composite of SI and Red Holstein. Bulls were genotyped using the Illumina Bovine SNP50^{TM} Beadchip. Genotypes of OB and BS (data set BSW) were merged and analysed together as well as SI, SF, and HO (data set MIX). Population structures of BSW and MIX, which was inferred from principal component analysis on the SNP genotypes, are shown in Figure 1 and Figure 2. Bulls and SNPs with a call rate < 90 % were removed from the analysis. SNPs significantly deviating from Hardy-Weinberg-Equilibrium ($p<0.00001$) and with a minor allele frequency of < 0.005 were discarded. After filtering, BSW and MIX included 3,738 animals genotyped for 39,743 SNPs and 4,753 animals genotyped for 39,841 SNPs, respectively. The pedigrees for BSW and MIX

included 20.743 and 28.202 animals. To be in step with actual breeding practice, 20 % of the most recently born animals (birth year 2008-2011) in BSW and MIX were selected as the imputation candidates. They were assumed to be genotyped with the Illumina 3k SNP Chip, a subset of Illumina 54k SNP chip, comprising 3,268 and 3,145 SNP for BSW and MIX, respectively. All other genotypes were masked and considered to be unknown. In conclusion, the set remaining animals built the 54k-genotyped reference group for imputation.

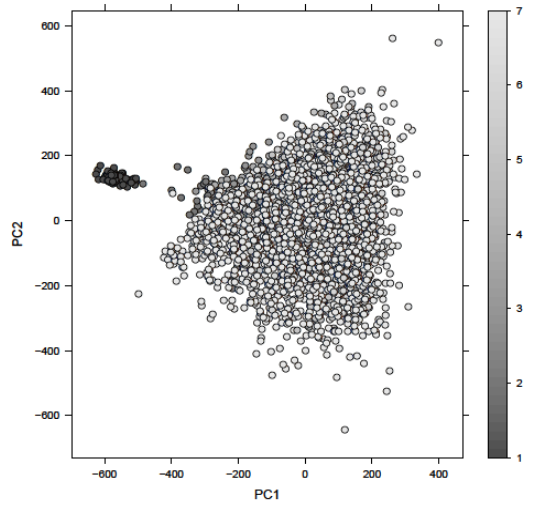

Figure 1. Population structure of BSW according to proportion of OB genes (class 1 = 100% OB, class 7 = 0% OB).

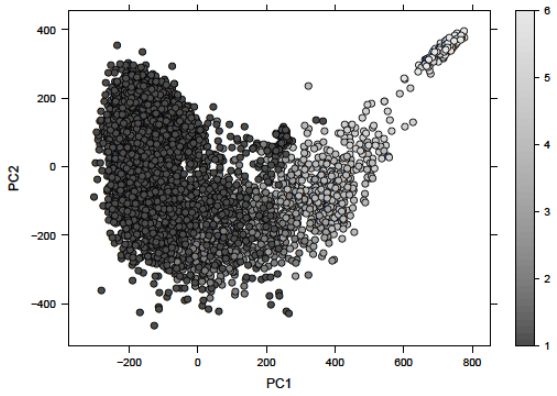

Figure 2. Population structure of MIX according to proportion of HO genes (class 1 = 100% HO, class 6 = 0% HO).

Methods

Imputation from Illumina 3k to 54k was performed using AlphaImpute (Hickey *et al.*, 2011a) and Findhap V2 (Van Raden *et al.*, 2011). AlphaImpute implements an imputation algorithm based on segregation analysis and long haplotype library imputation (SALHI). SALHI includes three major steps: (1) allele probabilities for each locus of each animal in the pedigree are calculated using segregation analysis (Kerr and Kinghorn, 1996), (2) phasing of individuals genotyped for the high density chip based on the long range phasing algorithm by Hickey et al. (2011b) and storage of identified haplotypes in a haplotype library, (3) missing alleles are imputed by matching allele probabilities (step 1) to identified haplotypes (step 2).

Hickey *et al.* (2011a) give a detailed description of AlphaImpute.

The software Findhap V2 is described in detail by Van Raden *et al.* (2011). Findhap imputes unknown genotypes by combining population-based with pedigree haplotyping. Chromosomes are first divided into segments of "n" SNP. Genotypes are lined up to a library of known haplotypes. Pedigree data is then used to correct Mendelian errors between parent and offspring haplotypes.

Accuracy of imputation of AlphaImpute and FindhapV2 was evaluated by comparing the original and imputed genotypes according to percentage of SNP imputed correctly, incorrectly and not imputed.

Results and Discussion

Imputation accuracy using AlphaImpute and Findhap for BSW and MIX is shown in Table 1 with accuracies being analysed according to ancestors' genotyping status.

Table 1. Mean, standard deviation (sd), minimum (min), and maximum (max) percentage of SNP imputed correctly, incorrectly, and not imputed using AlphaImpute and Findhap V2 for Brown Swiss (BSW) and the mixture data set (MIX).

Relatives HD genotyped		**AlphaImpute**				**Findhap V2**			
					BSW				
Imputed correctly	N	mean	sd	min	max	mean	sd	min	max
Both parents	27	97.74	2.74	85.82	99.11	94.74	4.07	76.92	98.33
Sire and MGS[1]	573	94.34	2.88	74.47	98.35	92.95	2.68	69.45	96.30
Sire	55	92.12	5.87	55.13	96.60	91.39	3.99	66.99	94.76
Other relatives	68	86.69	8.50	38.1	96.36	89.71	2.95	73.46	94.01
Imputed incorrectly									
Both parents	27	1.65	2.24	0.71	11.49	5.27	4.07	1.67	23.08
Sire and MGS[1]	573	3.65	1.88	1.13	17.58	7.04	2.57	3.70	26.81
Sire	55	4.01	1.73	1.44	12.75	8.51	3.37	5.24	27.28
Other relatives	68	7.48	2.85	0.99	18.61	10.29	2.95	6.00	26.54
Not imputed									
Both parents	27	0.62	0.54	0.18	2.69	0	-	0	0
Sire and MGS[1]	573	2.01	1.13	0.44	14.09	0	0.01	0	0.09
Sire	55	3.87	4.55	1.01	32.11	0.01	0.03	0	0.19
Other relatives	68	5.84	8.47	1.06	60.91	0	0	0	0
					MIX				
Imputed correctly	N	mean	sd	min	max	mean	sd	min	max
Both parents	1	98.93	-	98.93	98.93	97.81	-	97.81	97.81
Sire and MGS[1]	331	93.23	2.59	79.06	98.36	91.19	2.29	77.37	95.80
Dam and PGS[2]	1	95.94	-	95.94	95.94	91.86	-	91.86	91.86
Sire	59	91.41	4.24	73.44	97.39	90.06	3.47	77.24	94.24
Other relatives	544	82.53	9.60	24.52	95.36	87.51	2.86	68.39	93.20
Imputed incorrectly									
Both parents	1	0.57	-	0.57	0.57	2.19	-	2.19	2.19
Sire and MGS[1]	331	4.03	1.52	0.91	13.49	8.81	2.29	4.20	22.63
Dam and PGS	1	3.16	-	3.16	3.16	8.14	-	8.14	8.14
Sire	59	4.57	1.72	1.66	11.71	9.94	3.47	5.76	22.76
Other relatives	544	8.49	2.87	0.13	22.86	12.49	2.86	6.81	31.61
Not imputed									
Both parents	1	0.50	-	0.50	0.50	0	-	0	0
Sire and MGS[1]	331	2.74	1.25	0.73	10.49	0	0	0	0
Dam and PGS[2]	1	0.91	-	0.91	0.91	0	-	0	0
Sire	59	4.03	3.27	0.95	20.39	0	0	0	0
Other relatives	544	8.99	9.79	1.64	74.63	0	0	0	0

[1] MGS = Maternal grandsire
[2] PGS = Paternal grandsire

On average 97.74 % of all SNPs were imputed correctly for imputing candidates of BSW where both parents were genotyped with high-density SNP panel using AlphaImpute. When sire and maternal grandsire (MGS), sire only, or other relatives were HD genotyped 94.34 %, 92.12 %, and 86.69 % of all SNPs were imputed correctly, respectively. Lower percentage of SNP imputed correctly were observed using Findhap V2 for BSW when both parents were HD genotyped. Likewise, accuracy of imputation was lower for candidates when sire and MGS or sire only were HD genotyped compared to AlphaImpute. In contrast, when other relatives were HD genotyped Findhap V2 outperformed AlphaImpute (89.71 % vs. 86.69 % SNP imputed correctly). A similar trend was observed for MIX for percentage of SNPs imputed correctly, though imputation accuracy was slightly lower than for animals in the BSW data set. Percentage of SNPs imputed correctly decreased with decreasing relationship between imputing candidates and animals in the HD genotyped reference using both methods.

Average percentage of SNP imputed incorrectly were 1.65 and 5.27 for BSW when both parents were HD genotyped using AlphaImpute and Findhap V2, respectively. AlphaImpute gave lower percentage of SNP imputed incorrectly than Findhap V2 in all relationship classes for BSW and MIX.

A small percentage of SNP is not imputed at all when using AlphaImpute. Genotypes of these SNP are filled using genotype probabilities derived in step (1). Almost all SNP are imputed using Findhap V2.

Conclusions

AlphaImpute and Findhap V2 delivered accurate genotype imputation results for BSW and MIX. When close relatives of imputation candidates are HD genotyped AlphaImpute outperforms Findhap V2. In contrast, when the HD reference population is less related to imputation candidates Findhap V2 was found to give better results. Further studies will be carried out to examine the effect of imputed genotypes from AlphaImpute and Findhap V2 on accuracy of genomic breeding values.

Acknowledgements

We thank the Swiss Brown Cattle Breeders Federation, swissherdbook cooperative and Holstein Switzerland for providing genotypes.

References

Habier, D., Fernando, R.L. & Dekkers, J.C. 2009. Genomic selection using low-density marker panels. *Genetics 182,* 343-353.

Hickey, J.M., Cleveland, M., Gorjanc, G., Tier, B., van der Werf, J.H.J. & Kinghorn, B. 2011a. An imputation strategy which results in an alternative parameterization of the Single Step Genomic Evaluation. *Interbull Bulletin 44,* 38-41.

Hickey, J.M., Kinghorn, B.P., Tier, B., Wilson, J.F., Dunstan, N. & van der Werf, J.H.J. 2011b. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genet. Sel. Evol. 43,* 12.

Kerr, R.J & Kinghorn, B.P. 1996. An efficient algorithm for segregation analysis in large populations. *J. Anim. Breed. Genet. 113,* 457-469.

Sheet, P. & Stephens, M.A. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics 78,* 629-644.

Van Raden, P.M., O'Connell, J.R., Wiggans, G.R. & Weigel, K.A. 2011. Genomic evaluation with many more genotypes. *Genet. Sel. Evol. 43,* 10.