Accuracy of Imputation to Whole-Genome Sequence in Nelore Cattle

Roberto Carvalheiro[1,2], Gerardo A. Fernandes Júnior[1], Mehdi Sargolzaei[3,4], Ricardo V. Ventura[5], Henrique N. Oliveira[1,2], Lucia G. Albuquerque[1,2]

[1]School of Agricultural and Veterinarian Sciences, São Paulo State University (Unesp), Jaboticabal - SP, Brazil
[2]National Council for Scientific and Technological Development (CNPq), Brasília - DF, Brazil
[3]University of Guelph, Department of Pathobiology & CGIL, Guelph - ON, Canada
[4]Select Sires Inc., Plain City - OH, USA
[5]School of Veterinary Medicine and Animal Science - FMVZ/USP, Pirassununga - SP, Brazil

SCOPE AND RELEVANCE

Advances in next-generation sequencing techniques associated to a drastically decrease in sequencing costs is favoring the development of strategies to explore the complete DNA sequence in genetic evaluations. However, cost is still a limitation to sequence a large number of animals. An alternative and cost-effective strategy would be to sequence key ancestors of the population and impute to whole-genome sequence the genotypes of the remaining animals genotyped with SNP arrays. Our research group has a database with more than 60,000 Nelore animals, from different breeding programs, genotyped with Illumina SNP arrays of varying densities. Using the sequence data of 151 influential Nellore bulls, we will carry out the imputation of these genotypes to whole genome sequence, which will increase the amount of genomic information per animal. Investigation of the imputation accuracy is essential to determine the feasibility of this process.

OBJECTIVES

Investigate the imputation accuracy from Illumina BovineHD Beadchip (~777K) to whole-genome sequence in a Nelore beef cattle population, to assess the feasibility of the imputation process in this breed; compare imputation accuracy of two software (FImpute3 and Minimac4); and develop a pipeline to impute sequence data of our existing genotype database.

MATERIAL AND METHODS

- ✓ Whole-genome sequencing of 151 influential Nelore sires, using Illumina HiSeq X™ Ten (52) and Illumina NovaSeq™ (99) platforms, at an overall average sequence coverage (after quality control) of 14.5 (7.8-26.3).
- ✓ The sequenced sires were chosen based on a *k-means* cluster analysis using the genomic relationship matrix of all genotyped animals in our database. The number of clusters was set equal to 151 and within each cluster the sire with the highest number of genotyped progeny was chosen to be sequenced. This strategy aimed to optimize the imputation accuracy of our genotype database.
- ✓ Variant calling and quality control procedures were carried out following the guidelines provided by the 1000 bull Genomes Project (www.1000bullgenomes.com).
- ✓ After quality control, a total of 30,394,484 SNPs located on autosomes were left.
- ✓ Imputation was carried out under a 5-fold cross-validation scheme. The 151 sires with sequence information were randomly divided into five groups, and 5 imputation analyses were performed using one group at a time as target population and sires from the remaining groups as reference population. Sires from the target population had all their genotypes masked, except those that overlapped with the HD chip (565,035 SNPs), pretending they were genotyped with the HD chip.
- ✓ Imputation was carried out using two software: FImpute v3 (Sargolzaei et al., 2014) and Minimac4 (Howie et al., 2012). For Minimac4, the phasing step was done by using the software Eagle (Loh et al., 2016).
- ✓ Imputation accuracy was assessed through Pearson's correlation between observed and imputed genotypes (CORR) and percentage of correctly imputed genotypes (PERC) of the target population.

RESULTS AND DISCUSSION

- ✓ The concordance rate between the commercial HD genotypes and the set of markers obtained from the sequencing process was, on average, equal to 99.64% (ranging from 97.34 to 99.96%), indicating good quality of sequencing information.
- ✓ Imputation accuracies were high and consistent across methods and statistics. The average (minimum and maximum) sample-wise CORR and PERC (%) statistics were, respectively, 0.97 (0.94 to 0.99) and 96.57 (93.97 to 98.55) for FImpute3, and 0.97 (0.95 to 0.99) and 97.14 (94.97 to 98.88) for Minimac4. .
- ✓ On average, common variants (MAF>0.03) were more accurately imputed by Minimac4 and low-frequency variants (MAF≤0.03) were imputed with higher accuracy by FImpute3 (Figure 1).
- ✓ The observed FImpute's advantage for imputing SNP with low MAF could be due to the fact that most rare variants are recent and located on long haplotypes, which are quite efficiently exploited by the FImpute software.
- ✓ Regarding computing time, FImpute3 was much more efficient than Eagle+Minimac4. For each 5-fold imputation analysis, FImpute3 took around 30 minutes whereas Eagle+Minimac4 run in approximately 40 hours.

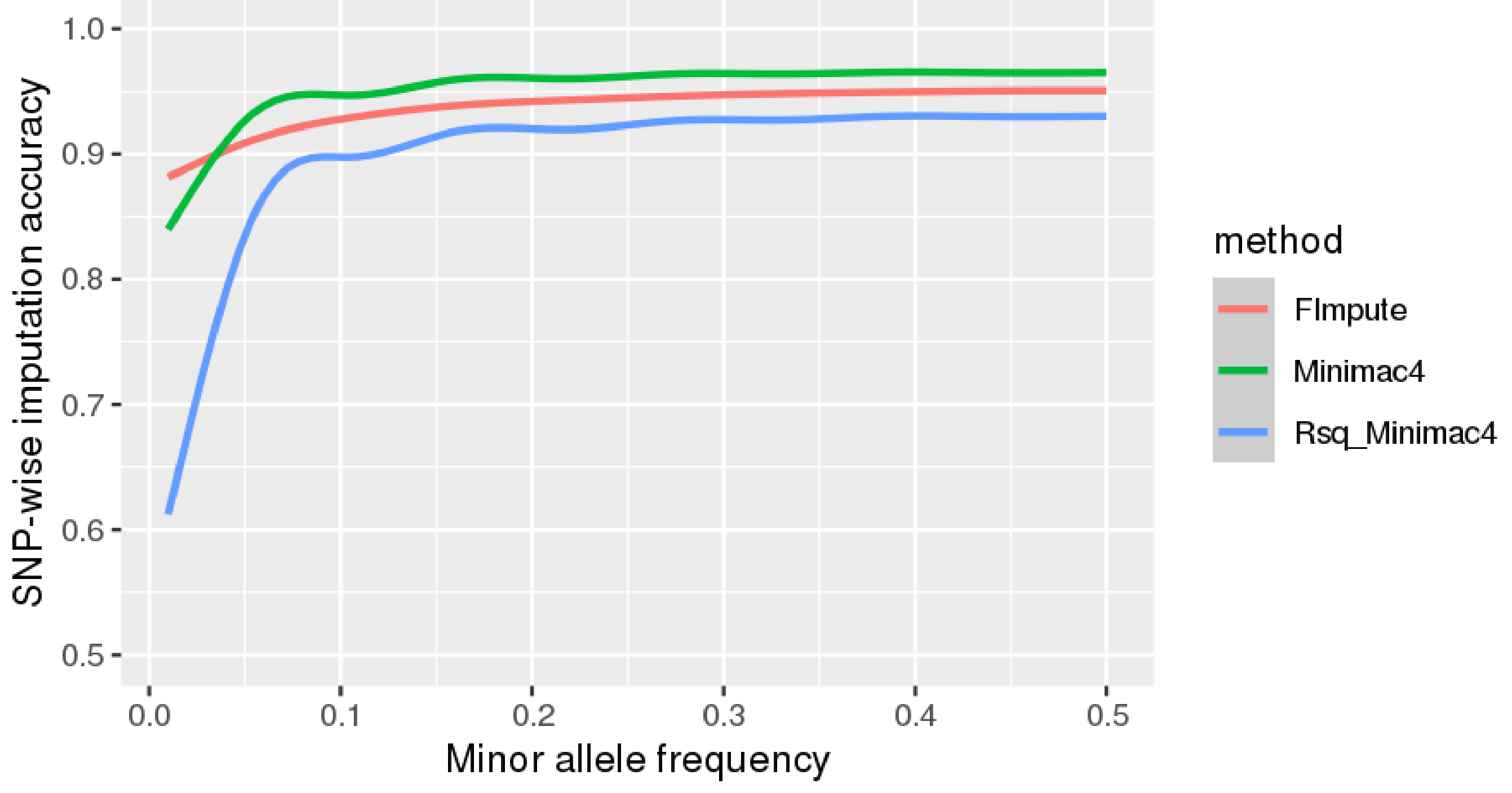

Figure 1. SNP-wise imputation accuracies by minor allele frequency.

Imputation accuracy for FImpute and Minimac4 correspond to Pearson's correlation between observed and imputed genotypes (CORR); Rsq_Minimac4 is an estimate of the squared correlation between imputed and true unobserved genotypes, provided by Minimac4.

CONCLUSIONS

Our results indicate that whole-genome sequence imputation is feasible in Nelore cattle since high imputation accuracies were achieved regardless the imputation software tested. The overall SNP-wise imputation accuracy was software-dependent. On average, Minimac4 provided higher imputation accuracies for common variants and FImpute3 had higher accuracies for rare variants. FImpute3 was much faster than Eagle+Minimac4 regarding computing time.

ACKNOWLEDGEMENTS

FAPESP

09/16118-5 17/10630-2
18/10109-3 19/12434-1

CAPES

CNPq