Animal 18 (2024) 101087

Contents lists available at ScienceDirect

Animal

The international journal of animal biosciences

Short communication: Accuracy of whole-genome sequence imputation in Angus cattle using within-breed and multi breed reference populations

N. Kamprasert [*], H. Aliloo, J.H.J. van der Werf, S.A. Clark

School of Environmental and Rural Science, University of New England, 2351, Armidale, NSW, Australia

ARTICLE INFO

Article history:
Received 4 October 2023
Revised 16 January 2024
Accepted 19 January 2024
Available online 26 January 2024

Keywords:
Beef cattle
Genotype imputation
Imputation accuracy
Whole genome sequencing
1 000 bull genomes

ABSTRACT

Genotype imputation is a standard approach used in the field of genetics. It can be used to fill in missing genotypes or to increase genotype density. Accurate imputed genotypes are required for downstream analyses. In this study, the accuracy of whole-genome sequence imputation for Angus beef cattle was examined using two different ways to form the reference panel, a within-breed reference population and a multi breed reference population. A stepwise imputation was conducted by imputing medium-density (50k) genotypes to high-density, and then to the whole genome sequence (**WGS**). The reference population consisted of animals with WGS information from the 1 000 Bull Genomes project. The within-breed reference panel comprised 396 Angus cattle, while an additional 2 380 Taurine cattle were added to the reference population for the multi breed reference scenario. Imputation accuracies were variant-wise average accuracies from a 10-fold cross-validation and expressed as concordance rates (**CR**) and Pearson's correlations (**PR**). The two imputation scenarios achieved moderate to high imputation accuracies ranging from 0.896 to 0.966 for CR and from 0.779 to 0.834 for PR. The accuracies from two different scenarios were similar, except for PR from WGS imputation, where the within-breed scenario outperformed the multi breed scenario. The result indicated that including a large number of animals from other breeds in the reference panel to impute purebred Angus did not improve the accuracy and may negatively impact the results. In conclusion, the imputed WGS in Angus cattle can be obtained with high accuracy using a within-breed reference panel.

© 2024 The Author(s). Published by Elsevier B.V. on behalf of The Animal Consortium. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Implications

Genotype imputation is widely used in the genetic evaluation of livestock to amplify genotype data to higher densities. Accurate imputation is important because poorly imputed genotypes may reduce the potential benefit from the use of dense genotypes in genomic selection. This study reports that when imputing purebred Angus cattle highly accurate imputation can be achieved using a within-breed reference. Adding animals from additional breeds to a reference panel did not improve the results and, in some instances, negatively impacted imputation accuracy. Within-breed imputation procedure can be used to impute purebred Angus genotypes to the whole genome sequence for subsequent analyses.

Introduction

It is common for genetic evaluation schemes to use DNA-based genetic marker information, allowing for the prediction of genomic breeding values. Genomic evaluation and selection can accelerate the response to selection in livestock by shortening the generation interval and enabling the prediction of more accurate breeding values, especially in young animals and for difficult-to-measure traits. The density of marker information used has been previously shown to improve genomic selection. For instance, Druet et al. (2014) reported using simulations that in some scenarios, prediction accuracy was improved by up to 30% when using whole genome sequences (**WGS**); they also pointed out a potential benefit of sequence data for fine mapping of quantitative trait loci.

The advantage of WGS data over low- to medium-density marker data has also been demonstrated in real data. For example, van Binsbergen et al. (2014) showed that more accurate predictions of genomic breeding values and an improved power to detect quantitative trait loci were possible in dairy cattle when

[*] Corresponding author.
E-mail address: nkampras@myune.edu.au (N. Kamprasert).

https://doi.org/10.1016/j.animal.2024.101087
1751-7311/© 2024 The Author(s). Published by Elsevier B.V. on behalf of The Animal Consortium.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

using WGS instead of different densities of marker data. Moghaddar et al. (2019) also illustrated that the WGS could be used to improve genomic prediction accuracy in Australian sheep. Although the cost of genotype sequencing has declined considerably since the advent of this technology, it is still economically unfeasible to genotype all individuals at WGS level. An alternative method to obtain sequence data is genotype imputation, which is a common approach in sequence-based studies.

Genotype imputation allows for unobserved genotypes in a sample of individuals to be predicted using a reference panel of haplotypes derived from observed sequence information (Ma et al., 2013; Rowan et al., 2019). Imputation can be used either to recover missing genotypes or amplify low-density genotype arrays to denser genotypes. Many studies have intended to achieve the most accurate imputed genotypes since inaccurate results can negatively influence the downstream analyses. Different factors affecting the performance of genotype imputation have been reported in previous studies (van Binsbergen et al., 2014; Das et al., 2016; Shi et al., 2018), for instance, the imputation algorithm implemented by software, reference panel selection, reference sample size, the relationship between reference and target sets, minor allele frequency and genotype quality.

A common theme across imputation studies is regarding the impact of the reference panel on the accuracy of genotype imputation. The questions are about which animals should be included in a reference panel, considering the size and composition, in order to achieve the most accurate imputation. The larger size of the reference panel gains more accurate imputation (Druet et al., 2014); however, the reference population from a single breed could be small because of the cost of sequencing. Hence, combining data from multiple breeds can offer an option to increase reference population size. Previous studies suggested that a multi breed reference panel can improve imputation performance (Brøndum et al., 2014). This study aimed to examine the effect of within-breed and multi breed reference panels on genotype imputation in Angus beef cattle.

Material and methods

Genotypes

Whole-genome sequence data were obtained from the 1000 Bull Genomes project (Hayes and Daetwyler, 2019). To investigate the imputation accuracy of Angus cattle, two different scenarios of imputation were examined. The first scenario included 396 Angus animals in the reference panel for a within-breed reference (**WB**). In the second scenario, 2 380 additional animals from 61 other Taurine breeds were added to the within-breed Angus reference panel for a multi breed reference scenario (**MB**). The selected samples had sequencing coverage greater than 10x. Only autosomal chromosomes were retained in this study. There were 163 156 536 genetic markers in autosomal chromosomes, of which 145 184 873 were single nucleotide polymorphisms (**SNPs**). SNP variants were called by using the Genome Analysis Toolkit and were filtered with the Variant Quality Score Recalibration (**VQSR**) (Hayes and Daetwyler, 2019). Briefly, the concept of VQSR is similar to machine learning, which requires truth and training datasets. The method learns the annotation profiles of good- and bad-quality variants from a truth dataset and then applies the rules to variants in a training set. After VQSR, each variant is assigned a score called variant quality score log-odds, and recalibration is applied using specified tranche sensitivity thresholds. The quality control process selected variants for which the variant quality score log-odds scores were greater than the threshold for the 90.0 tranche, which is the most stringent threshold. There were 39 173 127 SNPs that passed the VQSR criteria. Besides the VQSR, markers with a high missing genotype rate (> 0.1) or a low minor allele frequency (< 0.001) were also excluded. Then, only biallelic SNPs were kept for further analysis. Similar genotype quality control criteria were separately applied in the two imputation scenarios. This resulted in 13 965 792 and 26 171 746 polymorphic SNPs in the Angus and Taurus datasets (Table 1), respectively, with a total of 13 469 859 common SNPs between the two datasets.

Genotype imputation

Genotype imputation aimed to impute medium-density (**50K**) genotypes to WGS for the Angus cattle using WB and MB reference scenarios. A stepwise imputation strategy was utilized as it was recommended in the literature for genotype imputation from medium-density arrays to the sequence level (van Binsbergen et al., 2014). In the first step, the 50k genotypes were imputed up to high-density (**HD**), and then, the HD genotypes were imputed up to the WGS. A total of 46 469 and 562 155 SNPs were used for the 50K and HD genotypes, respectively, based on Illumina BovineSNP50 and BovineHD arrays. To assess imputation accuracy, the Angus WGS was masked to 50K, and the 440 Angus samples were randomly assigned to 10-folds in a cross-validation scheme, leaving 44 animals in a validation set and 396 animals in the reference panel for each iteration of imputation. In the MB scenario, 2 380 Taurine cattle were added to the reference panel, which summed up the reference size to 2 776, and the validation sets were the same as in the WB scenario. In each iteration, the same set of reference and validation samples were used for both HD and WGS imputation. Genotype imputations were performed using Minimac4 (version 1.0.3) (Das et al., 2016) with phased genotypes from Eagle (version 2.4.1) (Loh et al., 2016) with the default parameters**;** the detail of default parameters for Minimac4 (version 1.0.3) see Minimac4 Documentation (https://genome.sph.umich.edu/wiki/Minimac4_Documentation) and for Eagle (version 2.4.1) see Eagle v2.4.1 User Manual (https://alkesgroup.broadinstitute.org/Eagle/). Minimac4 operated with a population-based method, which was more suitable for this study due to the lack of pedigree information for all the animals used and the lack of pedigree linked across breeds in the case of MB.

Imputation accuracy was assessed by comparing the observed and imputed genotypes, considering only the imputed loci. Accuracy of imputation was expressed as the percentage of correctly imputed genotypes, including genotype concordance rate (**CR**), and Pearson's correlation coefficient (**PR**). The accuracy was calculated as an average per SNP across all animals in the validation set and averaged across the 10-fold in the cross-validation. The CR was calculated as the percentage of correctly imputed genotypes. PC values were the correlation between the observed and imputed dosages, where genotypes were encoded as 0, 1 and 2 based on alternate allele counts. Imputation accuracies, both CR and PR, were reported for independent SNPs from each scenario and common SNPs overlapped between the two scenarios.

Results and discussion

In general, variant-wise imputation accuracies were moderate to high across all chromosomes and were similar between the two scenarios for both HD and WGS imputations (Fig. 1). A large difference in the number of imputed genotypes between imputation scenarios can bias the accuracy of imputation. Therefore, only common SNPs between the two scenarios were considered to compare their imputation performance. The average CRs were constantly higher than the PRs. Previous studies have shown that CR often gives higher results than PR because markers with low minor

Table 1
Number of cattle and SNPs in a reference and validation set for imputation scenarios.

	High-density imputation		Whole genome sequence imputation		
Items	WB	MB	WB	MB	common SNPs
Number of cattle in each iteration					
Reference set	396	2 776	396	2 776	
Target set	44	44	44	44	
Number of SNPs					
Genotyped SNPs	36 714	35 880	523 221	511 266	510 712
Imputed SNPs	486 507	475 386	13 442 571	25 660 480	12 959 147
Total	523 221	511 266	13 965 792	26 171 746	13 469 859

Abbreviations: SNPs = single nucleotide polymorphisms; WB = a within-breed reference scenario; MB = a multi breed reference scenario.

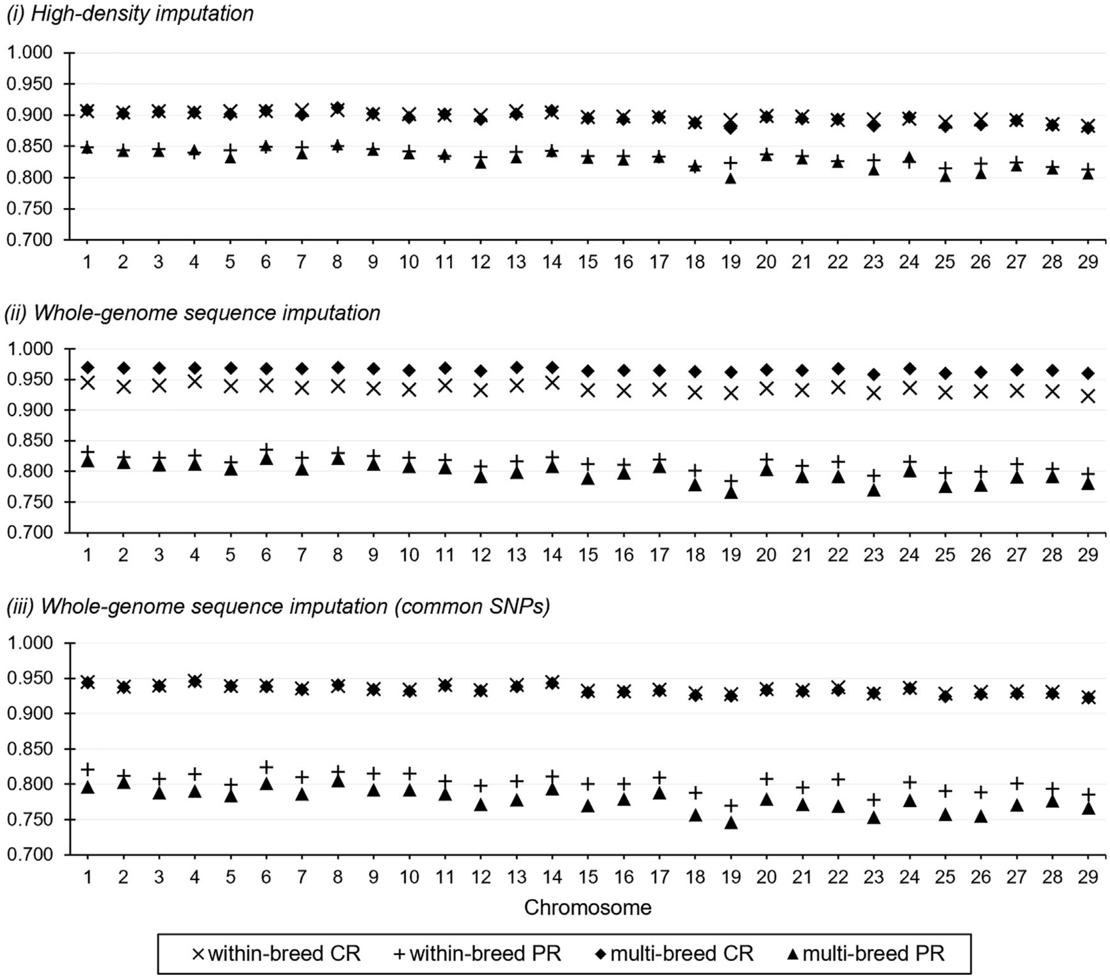

Fig. 1. Variant-wise mean imputation accuracies by chromosome in Angus cattle using within-breed reference panel and multi breed reference panel, where (*i*) 50k to HD imputation, (*ii*) HD to WGS imputation, and (*iii*) HD to WGS imputation considering only common SNPs. The imputation accuracies were expressed as concordance rates (CR) and Pearson's correlations (PR). Abbreviations: HD = high-density, WGS = whole genome sequence, SNPs = single nucleotide polymorphisms.

allele frequency appear to be well imputed by chance. The CR depends on the genotype probabilities and inflates imputation accuracy for low-frequency SNPs. On the other hand, PR accounts for filling by chance since the correlation formula considers the incorrect genotypes by the covariance component (Ma et al., 2013; Ramnarine et al., 2015; Kai-li et al., 2022). For the 50k to HD imputation, the average CR and PR were 0.899 and 0.834, respectively, for the WB scenario and 0.896 and 0.829 for the MB scenario (Table 2). A similar pattern of accuracy statistics between different reference panel imputations was also observed by Rowan et al. (2019) but higher values were reported. Their study performed genotype imputation in Gelbvieh cattle with a larger reference panel for both scenarios. The accuracies ranged from 0.993 to 0.999 for CR and 0.984 to 0.992 for PR in the within-breed scenario. In the multi breed scenario, the averages ranged from 0.996 to 0.999 for CR and 0.982 to 996 for PR.

The average imputation accuracies from HD to WGS for the WB scenario were 0.935 and 0.814 for CR and PR, respectively. The average CR and PR obtained from the MB scenario were 0.966 and 0.789. However, the results were biased because the number of imputed SNPs between the two scenarios was massively different. There were 25 660 480 imputed SNPs in the MB, while only

Table 2
Variant-wise mean imputation accuracy[1] in Angus cattle using two different reference panels.

	within-breed		multi breed		
Scenarios	CR	PR	CR	PR	*P*-value
HD imputation	0.899 ± 0.001	0.834 ± 0.002	0.896 ± 0.002	0.829 ± 0.003	0.166
WGS imputation[2]	0.935 ± 0.001	0.814 ± 0.002	0.966 ± 0.001	0.798 ± 0.003	–
WGS imputation (common SNPs)	0.935 ± 0.001	0.803 ± 0.002[a]	0.933 ± 0.001	0.779 ± 0.003[b]	< 0.0001

Abbreviations: CR = concordance rates; PR = Pearson's correlations; HD = high-density; WGS = whole genome sequence; SNPs = single nucleotide polymorphisms.
[1] The values were an average (mean ± SEM) from 10-fold cross-validation.
[2] There was no statistical comparison since a large difference in the number of imputed SNPs.
[a,b] Values within a row with different superscripts differ significantly at *P* < 0.05.

13 442 571 imputed variants in the WB scenario. The vast difference in the number of imputed variants was caused by the different genetic backgrounds between breeds represented in the dataset. Due to multiple breeds in the Taurus dataset, more genetic variants passed the quality control process.

Considering only the common SNPs between the two datasets, there were 12 959 147 imputed SNPs for the HD to WGS imputation. Average imputation accuracy expressed by CR and PR were 0.935 and 0.803 for the WB scenario, while these values were 0.933 and 0.779 for the MB scenario. There was no difference in CR between the imputation scenarios, but the PR from the WB scenario was significantly ($P < 0.01$) higher than that from the MB scenario. Rowan et al. (2019) suggested that if a within-breed reference panel imputation performed well, no significant improvement was expected when a multi breed reference panel was used for imputation. Moreover, introducing a large number of individuals from different breeds into the reference panel might introduce genetic variates, which did not exist in the target population, and may negatively impact the accuracy. A contrary pattern was observed by Frischknecht et al. (2017) who investigated imputation accuracy with different reference panels in Brown Swiss cattle. The study reported that average accuracies for a within-breed scenario ranging between 0.963 and 0.973 for CR and between 0.905 and 0.924 for PR. On the other hand, 0.983 and 0.987 for CR and between 0.927 and 0.943 for PR were average accuracies for a multi breed scenario. This study also emphasized that using a multi breed reference slightly improved imputation accuracy compared to a within-breed reference.

Our results for both CR and PR were reasonably high. In contrast to the previous studies of Frischknecht et al. (2017); Rowan et al. (2019), our accuracies were lower in comparison. These could be the reason for several factors affecting the imputation accuracy. Firstly, the reference size in the current study was smaller than in other studies. From the study of Rowan et al. (2019), 522 and 35 401 samples were used to construct a reference panel for a within-breed and multi breed scenario, respectively, and all of those animals were from the same region. While there was no recommended number of samples in the reference, the larger reference size was generally preferable to provide a more accurate imputation (Rowan et al., 2019). The second factor could be a weak relationship between animals in the dataset. The imputation accuracy has been proved to rely on the relationship between the reference and target sets (Brøndum et al., 2014). Although the reference size in the study of Frischknecht et al. (2017) was smaller than the current study, the relationship between the reference and target sets appeared stronger, especially in the within-breed scenario. The study informed pedigree-based relationship analysis, which included 123 sequenced animals from the within-breed scenario and their ancestors; therefore, it was assumed that the samples were from the same population. Nevertheless, in the current study, the Angus samples were pooled from different regions and were selected under different breeding programs. Hence, the relationship between the Angus cattle in the dataset might be weaker than it is expected in a purebred population. Additionally, Minimac4, the software used in the current study, exploited a summary of haplotypes in the reference panel to impute the genotypes. As a result, when these factors combined, the haplotype library from the reference inadequately represented potential haplotypes to match with genotypes in the target set. The factors outlined above may cause our imputation accuracies not to be as high as the previous studies.

In conclusion, a within-breed reference scenario can achieve similar performance to a multi breed reference scenario when an adequate number of the target breed are represented in a reference panel. In this case, introducing a considerable number of other breeds into a reference panel did not improve the accuracy of imputation. A potential explanation for this result is that the introduction of non-existing genetic variants from other breeds into the target population may negatively affect the imputation accuracy.

Ethics approval

Not applicable.

Data and model availability statement

The data that support the study findings are available from the 1000 bull genomes project regarding the agreement with the project. Information can be made available from the authors upon request.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) did not use any AI and AI-assisted technologies.

Author ORCIDs

N. Kamprasert: https://orcid.org/0000-0001-5747-7043.
S.A. Clark: https://orcid.org/0000-0001-8605-1738.
H. Aliloo: https://orcid.org/0000-0002-5587-6929.
J.H.J. van der Werf: https://orcid.org/0000-0003-2512-1696.

CRediT authorship contribution statement

N. Kamprasert: Writing – original draft, Methodology, Formal analysis. **H. Aliloo:** Writing – review & editing, Supervision, Methodology, Conceptualization. **J.H.J. van der Werf:** Writing – review & editing, Supervision. **S.A. Clark:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of interest

There is no conflict of interest involved.

Acknowledgements

The authors gratefully acknowledge the use of data from the 1000 Bull Genomes project.

Financial support statement

This research received no specific grant from any funding agency, commercial or not-for-profit section.

References

Brøndum, R.F., Guldbrandtsen, B., Sahana, G., Lund, M.S., Su, G., 2014. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics 15, 728. https://doi.org/10.1186/1471-2164-15-728.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W.G., Swaroop, A., Scott, L.J., Cucca, F., Kronenberg, F., Boehnke, M., Abecasis, G.R., Fuchsberger, C., 2016. Next-generation genotype imputation service and methods. Nature genetics 48, 1284–1287. https://doi.org/10.1038/ng.3656.

Druet, T., Macleod, I.M., Hayes, B.J., 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 112, 39–47. https://doi.org/10.1038/hdy.2013.13.

Frischknecht, M., Pausch, H., Bapst, B., Signer-Hasler, H., Flury, C., Garrick, D., Stricker, C., Fries, R., Gredler-Grandl, B., 2017. Highly accurate sequence imputation enables precise QTL mapping in brown Swiss cattle. BMC Genomics 18, 999. https://doi.org/10.1186/s12864-017-4390-2.

Hayes, B.J., Daetwyler, H.D., 2019. 1000 bull genomes project to map simple and complex genetic traits in cattle: applications and Outcomes. Annual Review of Animal Biosciences 7, 89–102. https://doi.org/10.1146/annurev-animal-020518-115024.

Kai-li, Z., Xia, P., Sai-xian, Z., Hui-wen, Z., Jia-hui, L.U., Sheng-song, X.I.E., Shu-hong, Z., Xin-yun, L.I., Yun-long, M.A., 2022. A comprehensive evaluation of factors affecting the accuracy of pig genotype imputation using a single or multi-breed reference population. Journal of Integrative Agriculture 21, 486–495. https://doi.org/10.1016/S2095-3119(21)63695-X.

Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., Reshef, Y.A., Finucane, H.K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., Durbin, R., Price, A.L., 2016. Reference-based phasing using the haplotype reference consortium panel. Nature Genetics 48, 1443–1448. https://doi.org/10.1038/ng.3679.

Ma, P., Brøndum, R.F., Zhang, Q., Lund, M.S., Su, G., 2013. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish red cattle. Journal of Dairy Science 96, 4666–4677. https://doi.org/10.3168/jds.2012-6316.

Moghaddar, N., Khansefid, M., van der Werf, J.H.J., Bolormaa, S., Duijvesteijn, N., Clark, S.A., Swan, A.A., Daetwyler, H.D., MacLeod, I.M., 2019. Genomic prediction based on selected variants from imputed whole-genome sequence data in Australian sheep populations. Genetics Selection Evolution 51, 72. https://doi.org/10.1186/s12711-019-0514-2.

Ramnarine, S., Zhang, J., Chen, L.-S., Culverhouse, R., Duan, W., Hancock, D.B., Hartz, S.M., Johnson, E.O., Olfson, E., Schwantes-An, T.-H., Saccone, N.L., 2015. When does choice of accuracy measure alter imputation accuracy assessments? PloS One 10, e0137601.

Rowan, T.N., Hoff, J.L., Crum, T.E., Taylor, J.F., Schnabel, R.D., Decker, J.E., 2019. A multi-breed reference panel and additional rare variants maximize imputation accuracy in cattle. Genetics Selection Evolution 51, 77. https://doi.org/10.1186/s12711-019-0519-x.

Shi, S., Yuan, N., Yang, M., Du, Z., Wang, J., Sheng, X., Wu, J., Xiao, J., 2018. Comprehensive assessment of genotype imputation performance. Human Heredity 83, 107–116.

van Binsbergen, R., Bink, M.C.A.M., Calus, M.P.L., van Eeuwijk, F.A., Hayes, B.J., Hulsegge, I., Veerkamp, R.F., 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution 46, 41. https://doi.org/10.1186/1297-9686-46-41.