University of Sistan and Baluchestan

Journal of Epigenetics

Online ISSN: 2676 - 4350

Homepage: https://jep.usb.ac.ir

Iranian Genetics Society

Imputation of Ungenotyped Individuals Based on Genotyped Relatives Using Machine Learning Methodology

Naeem Rastin Bojnord[1], Mehdi Aminafshar[2*], Mahmood Honarvar[3], Nasser Emam Jomeh Kashan[4]

[1] Department of Animal Science, Science and Research Branch, Islamic Azad University, Tehran, Iran. E-mail: naeemrastin64@yahoo.com

[2] *Corresponding author*, Department of Animal Science, Science and Research Branch, Islamic Azad University, Tehran, Iran. E-mail: Aminafshar@srbiau.ac.ir, Tel.: +982144868559.

[3] Department of Animal Science, Shahr-e-Qods Branch, Islamic Azad University, Tehran, Iran. E-mail: honarvar.mahmood@gmail.com

[4] Department of Animal Science, Science and Research Branch, Islamic Azad University, Tehran, Iran. E-mail: nasser_ejk@yahoo.com

ARTICLE INFO

Article type:
Research Article

Article history:
Received: 25 June 2020
Revised: 14 December 2020
Accepted: 14 December 2020

Keywords:
Genomic, Imputation Accuracy, Machine Learning, Random Forest, Support Vector Machine, Ungenotype.

A B S T R A C T

1) Background: Machine learning methods have been used in genetic studies to build models that are capable of predicting missing genotypes for both human and animal genetic variations. Genotype imputation is an important process of predicting unknown genotypes. The present study aimed to investigate the idea of machine learning as an imputation to compare the family-based methods and improve the imputation performance in different scenarios. It also compared the accuracies of Support Vector Machine (SVM) and Random Forest (RF). 2) Methods: The final population were simulated in the form of 100 families, including one sire with different number of genotyped progenies (2, 3, 4, 5 or 7). The number of markers was set to 5000 for whole genome. The sires in families and other scenarios such as, both parents, sire/dam and one progeny, sire and maternal grandsire were defined to investigate the ability of learning machine algorithm for imputation. 3) Results: The imputation accuracy ranged from 0.78 to 0.99 in different scenarios. Also, the least amount of imputation accuracy was achieved for sire and maternal grand sire scenario with both methods. Increasing in number of progenies from 2 to 3 was considerably increased in imputation accuracy (SVM and RF). The imputation of non-genotyped individuals is possible based on parent-offspring trios and the paired close relatives. However, the use of child- one parent genotyped, both parents genotyped and sire and maternal grandsire genotyped, average imputation accuracy would not exceed 85%. While the genotype of progenies are the best source of predicted genotypes for ungenotyped individuals, if the number of progeny is more than 4, the imputation accuracy is increased more than 95%. According to the results, the performance of machine learning methods in family of trios has a good accuracy and computational speed, which can be used in estimated breeding value.

Corrected Proof

Introduction

Genomic selection (GS) is the basis for the selection of competent and superior animals whose breeding value (BV) is estimated via high-density single nucleotide polymorphisms (SNPs) (Erbe et al. 2012). GS combines the information on genotypes, phenotypes and pedigree to

DOI: 10.22111/jep.2021.34789.1023

© Mehdi Aminafshar* **Publisher**: University of Sistan and Baluchestan

How to Cite: Rastin Bojnord, N., Aminafshar, M., Honarvar, M., Emam Jomeh Kashan, N. (2021). Imputation of ungenotyped individuals based on genotyped relatives using Machine Learning Methodology. *Journal of Epigenetics*, 2(2), 13-22. DOI: 10.22111/jep.2021.34789.1023

increase the accuracy of estimated breeding value (EBV) (Grossi et al. 2018). BV is calculated as the sum of the total SNPs effects. Hence, estimating the SNPs effects is absolutely essential in GS (Liesbeth. 2014). On the one hand, the ordinary method of controlling data quality in GS removes some SNPs from the dataset, causes data loss, and reduces EBV accuracy (Samorè and Fontanesi, 2016). The major challenge in GS is the cost of genotype determination in most species (Grossi et al. 2018). Therefore, it is not costeffective to determine all individual genotypes by using dense chips, and the value of information in the determined genotype should be more than the cost of genotype itself (Boison et al. 2015). Innovations in genomic technologies could provide new tools to enhance animal productivity and welfare (Mikhchi et al. 2016). Meanwhile, one of the technologies used in genomic evaluation is imputation, which identifies missing genotypes via the haplotype or genotype reference panel (Bai et al. 2019). Genotype imputation can facilitate the use of low-density chips, so it can predict missing genotypes for both human and animal variations, which may be more economical and facilitate the extensive implementation of GS (Boichard et al. 2012; Chen et al. 2013). Imputation also allows geneticists to accurately evaluate the evidence at genetic markers that are not directly genotyped. In GS programs in cattle industries, the use of dense chips increases the accuracy of candidate bulls for commercial sperm production and sales (Mrode et al. 2019). However, in order to take full advantage of GS, a large number of female animals must be determined. For this purpose, low density chips can be used for female animals, which reduces the cost of genotype (Dassonneville et al. 2012). There are different commercial SNP chips for dairy cattle, including Bovine3K for 2,900 SNPs, BovineSNP50 comprised of 54,001 SNPs and BovineHD (777,962 SNPs) produced by Illumina, the United States and 3 million SNPs from Bovine Genomic Database produced by Affymetrix (Berry et al. 2016). Also, an increase in marker density can raise the cost of genotyping. We can possibly determine the genotype of a group of animals with dense SNP chip. Then, via imputation, we can predict the genotype of other individuals in the population who were genotyped with low-dense chip (Bouwman et al. 2014; Sargolzaei et al. 2014). The properties of the population structure observed in many livestock species often include semi-relatives or all relatives and animals with several progenies. Imputation methods can be divided into family-based methods, which employ linkage information from close relatives, and population-based methods, which use population Linkage Disequilibrium (LD) information (Sargolzaei et al. 2014). A “trio” data consist of genotypes from sire-dam-progeny triplets. A child's genotype can be predicted via the genomic information obtained from its parents. Thus, some phasing algorithms are adapted to be used in this type of data (Lu & Cantor. 2014). Despite the lack of extensive research on the accuracy of imputing ungenotyped individuals, it is dependent on different factors, including the number of genotyped close relatives (Bouwman et al. 2014; Sargolzaei et al. 2014), the number of SNPs in the low-dense panel (Grossi et al. 2018), the rare allele frequency, LD between the markers (Bernardes et al. 2019), the number and type of reference population livestock (Butty et al. 2019), effective population size, and the analysis model (Heidaritabar et al. 2015). However, the level of LD plays an essential role in all methods. The results show that strong LD levels can increase the imputation accuracy (Bernardes et al. 2019; Heidaritabar et al. 2015). Also, the size of the reference population and the degree of kinship between the target and the reference population, the rate of genotyped close relatives should be considered in the design of the LD panel (Grossi et al. 2018; Herry et al. 2018). Butty et al. 2019. reported that increased number of reference animals increase the imputation accuracy. Imputation methods usually employ the haplotype or genotype reference panel estimating the missing genotypes in the sequenced samples with low-density genotype (Bai et al. 2019). Besides, researchers identified that the use of the 1000 Genomes Project that has more populations and haplotype numbers. Also, HapMap indicated the relationship between the sample size and sequencing coverage density to imputation accuracy (Shi et al. 2018). Thus, the reference panel size and marker density in the study sample always increase the imputation accuracy (Bai et al. 2019). In addition to the phenotypic records, increasing genetic improvement requires the use of GS programs with an optimal reference population. The main limitations of using GS include the large number of marker required and the cost of predicting genotype. Machine learning methods have been used in genetic studies to discover the underlying genetic defect and build models that are capable of predicting missing data of a marker (Bouwman et al. 2014; Sargolzaei et al. 2014). In this research, we performed the idea of using machine learning as imputation to compare the family-based methods and improve the imputation performance in different scenarios. We also, compared the accuracies of two different methods (Support vector Machine; SVM, Random forest; RF) in different scenarios for imputation of non-genotyped individuals using genotype information from relatives/parental or his/her progenies.

Camera Ready

2. Methods

2.1. The data simulation

Due to the lack of real data, the population of trio families in 100 replications were simulated using the statistical software package R version 3.1.2, (R Development. 2014). The R package hypred (Technow, 2015) was modified to suit the needs of sampled from the source of male and female gametes. After generating the base population, the final population were simulated in the form of different family structures.

2.2. Simulation of genomic population

The genome comprised five chromosomes of 1 M each and the number of SNPs per chromosome were 10000. The number of markers was set as 5000 for the whole genome. A historic population consisted of 100 individuals (50 males, 50 females) randomly mated during 50 generations using a mutation rate of 2.5×108 per site. The reference population was generated from the historic population by mating parent groups. The parent groups were randomly selected from the last generation of the historic population in which half of the animals were female. This structure was followed by 100 generations of random mating. Mating was performed by drawing the parents of an animal randomly from the animals of the previous generation. After these 100 generations, the population expanded to 1000 individuals. 100 sires were randomly selected and randomly mated with females, thereby stimulating 100 families including trio families or one sire with a different number of progenies (2, 3, 4, 5, or 7). The sires in families are considered as ungenotyped individuals.

2.3. Imputation methods

2.3.1 SNP window

The SNP window is defined by a fixed number of SNPs to the left and right ($L+R$) of the un-typed SNP (except when the un-typed SNP is near the end of a chromosome). An SNP window of size L corresponds to considering $L/2$ SNPs to the left and $L/2$ SNPs to the right of the target SNP. In all imputation methods, we define an SNP window of size L centered at marker i to extend L markers left and right. For SNPs less than L markers from the beginning or end of a chromosome, the window extends L SNPs in one direction and to the boundary of the chromosome in the other.

2.3.2 SNP distance

Distance could be defined in terms of the index of the SNP, the physical position on the chromosome, or the genetic distance. We fitted a distance measure to the observed correlation matrix between markers and selected the best window size of 10 for imputation by scanning over a large range of windows (Roberts et al. 2007). If we have *N*SNPs, we can construct the N by N correlation matrix C whose $_{ij}$'th element C_{ij} equals the correlation (We use a Pearson's correlation coefficient between the SNPs) between the phased, binary SNPs with indices $_i$ and $_j$. We then model $|c_{ij}|$ by $\exp\left(-\sum_{k=i}^{j-1} d_k\right)$

For positive distances $d_k > 0$. we created a distance function d (i, j) which is increasing in $|i-j|$ and takes into account the correlation structure of the data yet is still fast to compute. By differentiating the squared error criteria f (d), we get a gradient descent formula.

(1)

$$f(d) = \sum_{i,j} \left(|C^{ij}| - \exp\left(-\sum_{k=i}^{j} d_k \right) \right)^2$$

(2)

$$\rightarrow \frac{\partial f(d)}{\partial d^s} = \Delta 2 \sum_{ij:i \le s < j} \exp\left(-\sum_{k=i}^{j-1} d_k \right) \sum_{i,j} \left(|C^{ij}| - \exp\left(-\sum_{k=i}^{j-1} d_k \right) \right)$$

These distances allow the window of predictive SNP's to use the left more than the right if they are more correlated overall. In the SNP Prediction Problem, our methods builds a model after given ($L+R$) SNP of parents as the training set. When an unknown SNP is given as a test sample, the model can predict the unknown SNP value. We employed SVM and RF methods to impute the SNPs in ungenotyped sires. For all of these methods, the genotype data set consisted of a matrix P with m individuals and n SNP loci that P (i, j) records the genotype of individual sat locus i. In the following, our target missing value is P (i, j) = NaN. We further assumed that individuals has a known value at locus i. Otherwise, it is excluded from the imputation but to be imputed in exactly the same way as sample j. Then we assume a genetic distance for each locus. Every other individuals has a known value at locus i, for otherwise it is excluded from the imputation but to be imputed in the same way as

individuals *j*. In our imputation methods, only parent genotype values at nearby SNP loci are used in the inference of *P (i, j*) in offspring. Our study "nearby" is characterized as a window centering at locus *i*, whose radius is set to a fixed genetic distance d. In other words, only SNP loci within distance *d* to the target locus *i* are used in the inference.

2.4 Supervised methods

Learning algorithms were used to impute untyped SNPs using SVM and RF methods.n2.4.1. Support vector machine (SVM) In machine learning, support-vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data for classification or regression analysis. Both training and target sample is represented as an (L+R)-dimensional binary vector. We use the SVM function of the R language for imputation. The essence of SVM is a kernel function K (xi, x) between any two SNP x and xi, which is the measure of similarity between two SNPs. Given a set of m SNP x1…xm, SVM estimates a scoring function for any new SNP x of the form using the following equation:

(3)

$$f(x) = \sum_{i=1}^{m} \alpha i K(xi, x) + c$$

Where, "αi" signifies the weights in the expression to be optimized by the SVM by maximizing the large positive scores for SNP in the training set. "C" is so-called complexity parameter that needs to be optimized for predication performance. It also controls possible over fitting of the training set. We used Polynomial kernel function, which has two additional parameters, d indicating the degree of freedom (also known as order of polynomial) and slope of alpha. In our experiments, the function of polynomial kernel is as follows:

(4)

$$k(x, y) = (\alpha x' y + c)^d$$

The parameters of SVM are optimized via grid search. Data is not normalized before processing to the SVM classifier (Johnson, 2008).

Data and then the classification trees to impute the un-typed SNP by taking the mode of the labels predicted by these trees. Random decision forests are the mode of the classes (in classification problems) or mean prediction (in regression problems) of the individual trees. More specifically, for a given training dataset, B bootstrapped datasets are used to build decision trees where a random subset of predictors are evaluated at each node (Ogutu et al. 2011).

2.4.2 Random forest (RF)

RF is an ensemble learning method that grows a forest of classification trees based on the training The Random Forest, which consists of B prediction trees, is used for classifying potential samples. We tested B = 100 in our experiments. For a test sample, each prediction tree will assign it to one of the K classes, and the class label of this sample is then determined by a majority vote from the B decision trees (Ogutu et al. 2011).

3. Results and discussion

The following items were considered to calculate the efficiency and imputation accuracy; 1. The correlation between true genotypes and imputed genotype (r). 2. The genotype concordance rate that described as the average proportion of correctly imputed genotypes within animal (CRg). 3. The allele concordance rate is described as the average proportion of correctly imputed alleles within animal (CRa); in this case, a genotype was imputed to be heterozygote but was truly homozygote was assumed to have one correct allele imputed.

Scenario 1

Table 1 shows the results of average imputation accuracy of ungenotyped individuals by using both parents genotyped. In S1, imputation accuracy and other criteria of accuracy in both methods were obtained different slightly.

Camera Ready

Table 1. Comparison of the imputation accuracy for ungenotyped individuals based on parents genotyped (sire and dam), using SVM and RF methods

	r		CRa		CRg	
M	Mean	SD	Mean	SD	Mean	SD
SVM	0.8197	0.0092	0.8840	0.0063	0.7681	0.0126
RF	0.8237	0.0105	0.8856	0.0080	0.7713	0.0161

Method (M), Standard deviation (SD), correlation between the true and imputed genotyped (r) allele concordance rate (CRa) and genotype concordance rate (CRg). Because each offspring inherits one haplotype from the male parent and the other from the female parent, it is possible to impute the genotype of a non-genotyped progeny from the genotypes of its parent. Bouwman et al. 2014. used the software AlphaImpute to the imputation of individuals with both parents genotyped and no offspring. They reported the imputation accuracy of 0.66 with AlphaImpute phased and of 0.70 with segregation analysis.

Scenario2

Table 2 presents the results of average imputation accuracy of ungenotyped individuals based on using genotyped sires/dams with a genotyped progeny. The imputation accuracy and different criteria of accuracy was considerably increased with SVM.

Table 2. Comparison of the imputation accuracy for ungenotyped individuals using one genotyped parent (sire or dam) with a genotyped progeny, by SVM and RF methods.

	r		CRa		CRg	
M	Mean	SD	Mean	SD	Mean	SD
SVM	0.8877	0.0134	0.9205	0.0087	0.8412	0.0174
RF	0.8617	0.0104	0.9055	0.0061	0.8156	0.0117

Method (M), Standard deviation (SD), correlation between the true and imputed genotyped (r) allele concordance rate (CRa) and genotype concordance rate (CRg).

According to Bouwman et al.2014. if only one offspring was genotyped, the average animalspecific imputation accuracy across all animals would increase from 0.57 (±0.12) to 0.73 (±0.07). Chen et al. 2014. investigated the impact of genotype imputation on the performance of GBLUP and Bayesian method. Their results showed that imputation accuracy was 0.7738 when only sires were genotyped on the 50 k panel and 0.6213 when no parents had 50 k genotypes.

Shashkova et al. 2020. used FImpute to investigate the imputation quality using ADR and allelic correlation in animals that have no relatives, one parent, both parents, only one parent and grandfather, and only grandfather in the reference. Their results confirmed 86% parentages for the genotyped animals, for which parent genotypes were available. They reported that sufficiently high imputation accuracy was only achieved when both parents were genotyped via HD panels (Shashkova et al. 2020). Kamaei et al. 2017. examined the imputation accuracy of parent-offspring trios under different scenarios via simulated datasets and Beagle 3.3.2 program for imputation of parent-offspring trios. They showed that the imputation accuracy ranged from 0.67 to 0.96 for genotyped individuals and imputation accuracies decreased when dam was non-genotyped and both of offspring were lowdensity genotyped (Kamaei et al. 2017).

Scenario 3

Table 3 displays the results of average imputation accuracy for ungenotyped individuals via genotyped sires with a genotyped progeny. Imputation accuracy and (r, CRa and CRg) decreased compared to S2. However, the accuracy of RF in these scenarios was calculated slightly higher in comparison to SVM method.

Table 3. Comparison of the imputation accuracy for ungenotyped individuals using genotyped sires with a genotyped progeny by SVM and RF methods.

	r		CRa		CRg	
M	Mean	SD	Mean	SD	Mean	SD
SVM	0.8364	0.0168	0.8896	0.0112	0.7792	0.0225
RF	0.8404	0.0208	0.8923	0.0145	0.7845	0.0291

Method (M), Standard deviation (SD), correlation between the true and imputed genotyped (r) allele concordance rate (CRa) and genotype concordance rate (CRg).

Bouwman et al. 2014. found that the mean of imputation accuracy using AlphaImput method for animals that only their dams genotyped was 0.70 as the most recent ancestor, which was higher than expected. Also, Heidaritabar et al. 2015. showed a slight difference in imputation accuracy between individuals which had only their sire, only their maternal grandsire, or both these ancestors in the reference. Imputation accuracy in the 48K to 60K scenario for these groups of animals was always within 0.02 of the accuracy observed across the whole validation population (Heidaritabar et al. 2015).

Scenario 4

This option is close to reality because bulls have been proofed in the dairy breeding sector, and the use of genomic sperm has become more common in recent years (Taylor et al. 2018). 100 families were simulated, including sires and maternal grandsire (MGS) to evaluate this option. Table 4 shows the average imputation accuracy of ungenotyped individuals based on sires and MGS. Also, the least amount of imputation accuracy was achieved for this scenario using both methods.

Table 4. Comparison of the imputation accuracy for ungenotyped individuals using genotyped sires and maternal grandsire by SVM and RF methods.

	r		CRa		CRg	
M	Mean	SD	Mean	SD	Mean	SD
SVM	0.7888	0.0152	0.8520	0.0089	0.7184	0.0161
RF	0.8001	0.0265	0.8592	0.0139	0.7318	0.0223

Method (M), Standard deviation (SD), correlation between the true and imputed genotyped (r) allele concordance rate (CRa) and genotype concordance rate (CRg).

Scenario5

Among the haplotypes of a person's offspring is a correlation or allele concordance. In the first step, similar haplotypes, which had a common genetic origin and inherited from an ungenotyped parent were imputed by sliding window (Sargolzaei et al. 2014). Table 5 and 6 show the average imputation accuracy for ungenotyped parent with a different number of genotyped progenies (2, 3, 4, 5, 6 or 7) via SVM and RF methods. Increasing the number of progenies from 2 to 3 raised the imputation accuracy in both methods (SVM and RF). According to table 5 and 6, when two genotyped progenies were calculated, the accuracy of RF was higher than SVM.

Table 5. Average imputation accuracy for an ungenotyped parent with a different number of genotyped progenies (2, 3, 4, 5, 6 or 7) by SVM method.

	r		CRa		CRg	
NP	Mean	SD	Mean	SD	Mean	SD
2	0.8342	0.0223	0.8849	0.0158	0.7798	0.0316
3	0.9400	0.0251	0.9590	0.0214	0.9179	0.0449
4	0.9565	0.0241	0.9709	0.0190	0.9419	0.0399
5	0.9751	0.0219	0.9840	0.0171	0.9679	0.0242
6	0.9829	0.0172	0.9921	0.0072	0.9853	0.0105
7	0.9906	0.0156	0.9941	0.0714	0.9882	0.0104

Number of Progeny (NP), Standard deviation (SD), correlation between the true and imputed genotyped (r) allele concordance rate (CRa) and genotype concordance rate (CRg).

In the study of Bouwman et al. 2014. the genotypes of sires and MGS improved imputation accuracy; a correlation between the true and imputed genotyped, corrected for mean gene content, across all animals increased from 0.47 (real situation) to 0.60. Also, only small differences in imputation accuracies using different methods were observed for the animals with both their sire and MGS genotyped, and IMPUTE2, BEAGLE, and FImpute performed better than findhap and AlphaImpute (Ma et al. 2013). Moreover, an inability to impute the genotype of a large number of sires when a paternal half- sib family size of just three was presented previously in Berry et al. 2014.

Table 6. Average imputation accuracy for an ungenotyped parent with a different number of genotyped progenies (2, 3, 4, 5, 6 or 7) by RF method.

	r		CRa		CRg	
NP	Mean	SD	Mean	SD	Mean	SD
2	0.8447	0.0303	0.8913	0.0193	0.7803	0.0392
3	0.9394	0.0334	0.9580	0.0260	0.9159	0.0520
4	0.9643	0.0307	0.9763	0.0223	0.9526	0.0446
5	0.9765	0.0201	0.9849	0.0132	0.9699	0.0264
6	0.9874	0.0164	0.9937	0.0104	0.9874	0.0208
7	0.9970	0.004	0.9981	0.0029	0.9963	0.0059

Number of Progeny (NP), Standard deviation (SD), correlation between the true and imputed genotyped (r) allele concordance rate (CRa) and genotype concordance rate (CRg).

The results show that increase in several progenies raises the imputation accuracy, which is consistent with observations of some authors who have used real data and various software. In a study on seven dairy and beef cattle in Ireland, the genotype of parents were predicted via the information on the genotype of 4, 5 and 12. In this study, the genotype concordance rate was 0.942, 0.962, and 0.976 for these individuals. (Berry et al. 2014). Another authors showed that one, two and four genotyped progenies increased the accuracy of imputation across all animals from 0.57 (no progeny) to 0.73, 0.82, and 0.92, respectively (Bouwman et al. 2014). As in another study result, FImpute showed that the average accuracy of imputation for ungenotyped individuals using the genotype of 4-9, 10-19, 20-29 and 30-39 progenies, were 0.924, 0.941, and 961, respectively (Boison et al. 2014). Also, Piccoli et al. 2014. observed the accuracy of imputation for ungenotyped individuals using on the genotyped of 5 and 10 offspring using BEAGLE software, 0.90 to 0.92 and 0.96 to 0.98, respectively. Nicolazzi et al. 2013. reported that imputation of ungenotyped individuals with the genotype of 10 progenies and above10 using PedImpute were 0.93 and 0.97, respectively. Grossi et al. 2018. found that concordance rates increased greatly when the number of genotyped half-sib progeny increased from four to five, raised CRa of 0.94 and 0.96, respectively. Therefore, where biological material or genotypes are not available on individual animals, at least five progeny can be genotyped the parental with accuracy of imputation 96%, as described by Grossi et al. 2018. Mean imputation accuracy per individual was increased with raising in the number of progenies. Moreover, the genotype of the mate of the sire was also used, high mean genotype concordance rate for the individual of 0.988, with little additional benefit and was therefore obtained with seven genotyped progeny (Berry et al. 2018). A significant problem is if, for example, in genomic prediction, the application of phenotypes from imputed individuals is beneficial. In the simulation analysis discussed here, this problem is not directly addressed but has received some focus in the literature. For instance, the use of imputation genotypes for phenotyped individuals might improve the accuracy of GWAS in human GWAS studies if close relatives are genotyped (Wang et al. 2017; Datta et al. 2019; Liu et al. 2019; Ullah et al. 2019). A definite correlation between imputation accuracy and power improvement was not found by Ullah et al. 2019. But there appeared to be a tradeoff between imputation accuracy and sample size. While imputation was relatively unreliable, the use of phenotyped individuals with predicted genotypes increased power compared to the failure to use these individuals in the study (Willer et al. 2010; Ionita- Laza et al. 2013), which indicated that GWAS is worth with imputing ungenotyped individuals with phenotypes. In various situations, this research applied and compared the precision of two machine learning approaches using simulated data sets. 100 percent imputation accuracy was never observed on all data sets by both methods. However, due to its high performance in the imputation accuracy, RF was a more precise tool. A great alternative to RF might be the SVM, albeit less precise than RF. However, compared to SVM, RF takes more processing time. This research, however, did not suggest tested methods of machine learning with a high missing data rate for imputation. Under different conditions, we recommend other machine learning strategies for imputation accuracy in trios. Also, in population-based research, the approaches tested in this study can be used.

Content Ready

4. Conclusion

The results of each scenario indicated that imputation of non-genotyped individuals based on parent-offspring trios and close relatives paired. However, the use of child-one parent genotyped, both parents genotyped, and sires and maternal grandsire genotyped, average imputation accuracy would not exceed 85%. Also, in all scenarios, 100% accuracy of imputation was not observed, while genotyped progenies were the best source of predicted genotypes for ungenotyped individuals. If the number of progeny is more than 4, the imputation accuracy using both machine learning methods is increased more than 95%. In most scenarios of this study, the accuracies of RF were slightly higher than that of the SVM. Furthermore, a comparison of accuracy showed that the average of a gene that was not modified significantly reduced the correct accuracy, and imputation of ungenotyped individuals could help include valuable phenotypes for genome-wide association studies or genomic prediction, especially when ungenotyped individuals have genotyped progenies. The current data indicated that the performance of machine learning methods in a family of trios has a good accuracy and computational speed, which can be used in estimated breeding value. Therefore, using machine learning methods can be a great way to avoid increasing costs.

Acknowledgements

The results presented in this paper were from a Ph.D. student thesis. This study has been performed in the Science and Research of Azad University, Tehran, Iran.

Reference

Bai, W.Y., Zhu, X.W., Cong, P.K., Zhang, X.J., Richards, J.B., Zheng, H.F. (2019). Genotype imputation and reference panel: a systematic evaluation on haplotype size and diversity. Briefings in Bioinformatics. 6, bbz108. Doi: 10.1093/bib/bbz108.

Bernardes, P.A., do Nascimento, G.B., Savegnago, R.P., Buzanskas, M.E., Watanabe, RN., de Almeida Regitano, L.C., and et al. (2019). Evaluation of imputation accuracy using the combination of two high-density panels in Nelore beef cattle. Scientific Reports. 9 (17920). https://doi.org/10.1038/s41598-019-54382-w.

Berry, D.P., McHugh, N., Randles, S., Wall E, McDermott K, Sargolzaei M, and et al. (2018). Imputation of non-genotyped sheep from the genotypes of their mates and resulting progeny. Animal. 12(2), 191-198. https://doi.org/10.1017/S1751731117001653.

Berry, D.P., McParland, S., Kearney, J.F., Sargolzaei, M., Mullen, M.P. (2014). Imputation of ungenotyped parental genotypes in dairy and beef cattle from progeny genotypes. Animal. 8(6), 895-903. https://doi.org/10.1017/S1751731114000883.

Berry, D.P., O'Brien, A., Wall, E., McDermott, K., Randles, S., Flynn, P., and et al. (2016). Inter- and intra-reproducibility of genotypes from sheep technical replicates on Illumina and Affymetrix platforms. Genetics Selection Evolution. 48(1), 86. https://doi.org/10.1186/s12711-016-0267-0

Boichard, D., Chung, H., Dassonneville, R., David, X., Eggen, A., Fritz, S., and et al. (2012). Design of a bovine low-density SNP array optimized for imputation. PLoS One. 7(3). https://doi.org/10.1371/journal.pone.0034130.

Boison, S.A., Neves, H.H.R., Pérez O'Brien, A.M., Utsunomiya, Y.T., Carvalheiro, R., da Silva, M.V.G.B., and et al. (2014). Imputation of nongenotyped individuals using genotyped progeny in Nellore, a Bos indicus cattle breed. Livestock Science. 166,176-189. https://doi.org/10.1016/j.livsci.2014.05.033

Boison, S.A., Santos, D.J.A., sunomiya, A.H.T.U., Carvalheiro, R., Neves, H.H.R., PerezO'Brien, A.M., and et al. (2015). Strategies for single nucleotide polymorphism (SNP) genotyping to enhance genotype imputation in Gyr (Bos indicus) dairy cattle: Comparison of commercially available SNP chips. Journal of Dairy Science. 98 (7), 4969-4989. https://doi.org/10.3168/jds.2014-9213.

Bouwman, A.C., Hickey, J. M., Calus M.P.L., Veerkamp R.F. (2014). Imputation of non-genotyped individuals based on genotyped relatives: assessing the imputation accuracy of a real case scenario in dairy cattle. Genetics Selection Evolution. 46 (1), 6. https://doi.org/10.1186/1297-9686-46-6.

Butty, A.M., Sargolzaei, M., Miglior, F., Stothard, P., Schenkel, F.S., Gredler-Grandl, B., and et al. (2019). Optimizing selection of the reference population for genotype imputation from array to sequence variants. Frontiers in Genetics. 10(510). https://doi.org/10.3389/fgene.2019.00510.

Chen, J., Zhang, J.G., Li, J., Pei, Y.F., Deng, H.W. (2013).On combining reference data to improve imputation accuracy. PLoS One. 8(1). https://doi.org/10.1371/journal.pone.0055600.

Dassonneville, R., Fritz, S., Ducrocq, V., Boichard, D. (2012). Short communication: Imputation performances of 3 low-density marker panels in beef and dairy cattle. Journal of Dairy Science. 95(7), 4136-40. https://doi.org/10.3168/jds.2011-5133.

Datta, A.S., Lin, S., Biswas, S. (2019). A Family-Based Rare Haplotype Association Method for Quantitative Traits. Human Heredity. 83(4), 175-195. https://doi.org/10.1159/000493543.

Erbe, M., Hayes, B.J., Matukumalli, L.K., Goswami, S., Bowman, P.J., Reich, C.M., and et al. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of Dairy Science. 95, 4114–4129. https://doi.org/10.3168/jds.2011-5019

Grossi, D. A., Brito, L. F., Jafarikia, M., Schenkel, F. S., Feng. Z. (2018). Genotype imputation from various low-density SNP panels and its impact on accuracy of genomic breeding values in pigs. Animal. 12, 2235–2245. https://doi.org/10.1017/S175173111800085X.

Heidaritabar, M., Calus, M.P. L., Addie Vereijken, A., Groenen, M.A.M. Bastiaansen, J.W.M. (2015). Accuracy of imputation using the most common sires as reference population in layer chickens. BMC Genetics. 16 (101). https://doi.org/10.1186/s12863-015-0253-5.

Herry, F., Hérault, F., Picard Druet, D., Varenne, A., Burlot, T., Le Roy, P., and et al. (2018). Design of low density SNP chips for genotype imputation in layer chicken. BMC Genetics. 19,108. https://doi.org/10.1186/s12863-018-0695-7

Ionita-Laza, L., Lee, S., Makarov, V., Buxbaum, J. D., Lin, X. (2013). Family-based association tests for sequence data, and comparisons with population-based association tests. European Journal of Human Genetics. 21(10), 1158- 1162. https://doi.org/10.1038/ejhg.2012.308.

Camera Ready

Johnson, N. (2008). Genotype Prediction with SVMs. http://cs229.stanford.edu/ proj2008/Johnson-GenotypePredictionWithSVMs.pdf.

Kamaei, M., Honarvar, M., Aminafshar, M.,Abdollahi-Arpanahi, R. (2017). Imputation of parentoffspring trios and their effect on accuracy of genomic prediction using Bayesian method. Journal of Livestock Science and Technologies. 5(2). 43-52.

Liesbeth van der Waaij, K. O. (2014). Textbook. Animal breeding and genetics for BSc students. Centre for Genetic Resources and Animal Breeding and Genomics Group, Wageningen University and Research Centre. [Updated November 22, 2015; accessed September 1, 2019]. https://wiki.groenkennisnet.nl/display/TAB/.

Liu, C. T., Deng, X., Fisher, V., Heard-Costa, N., Xu, H., Zhou, Y., and et al. (2019). Revisit Populationbased and Family-based Genotype Imputation. Scientific Report. 9 (1800). https://doi.org/10.1038/s41598-018-38469-4

Lu, A.T., Cantor, R.M. (2014). Identifying rare-variant associations in parent-child trios using a Gaussian support vector machine. BMC Proceedings. 8 (Suppl 1), S98. https://doi.org/10.1186/1753-6561-8-S1-S98.

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 (7), 4666-77. https://doi.org/10.3168/jds.2012-6316.

Mikhchi, A., Honarvar, M., Emam Jomeh Kashan, N., Zerehdaran, S., Aminafshar, M. (2016). Comparison of three boosting methods in parent-offspring trios for genotype imputation using simulation study. Journal of Animal Science and Technology. 58 (1). https://doi.org/10.1186/s40781-015-0081-1.

Mrode, R., Ojango, J.M.K., Okeyo, A.M., Mwacharo, J.M. (2019). Genomic Selection and Use of Molecular Tools in Breeding Programs for Indigenous and Crossbred Cattle in Developing Countries: Current Status and Future Prospects. Frontiers in Genetics. 9 (694). https://doi.org/10.3389/fgene.2018.00694

Nicolazzi, E.L., Biffani, S., Jansen, G. (2013). Short communication: imputing genotypes using PedImpute fast algorithm combining pedigree and population information. Journal of Dairy Science. 96(4), 2649-2653. https://doi.org/10.3168/jds.2012-6062.

Ogutu, J.O., Piepho, H.P., Streeck, T.S. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proceedings. 5(Suppl3), S11. https://doi.org/10.1186/1753-6561-5-S3-S11.

Piccoli, M.L., Braccini, J., Cardoso, F.F., Sargolzaei, M., Larmer, S.G., Schenkel, F.S. (2014). Accuracy of genome-wide imputation in Braford and Hereford beef cattle. BMC Genetics. 29, 15:157. https://doi.org/10.1186/s12863-014-0157-9.

R Development Core Team. (2014). R: a language and environment for statistical computing, Vienna. http://www.r-project.org/.

Roberts, A., Millan, L.M.C., Wang, W., Parker, J., Rusyn, I., Threadgill, D. (2007). Inferring missing genotypes in large SNP panels using fast nearestneighbor searches over sliding windows. 23 i401–i407 Bioinformatics. https://doi.org/10.1093/bioinformatics/btm220.

Samorè, A. B., Fontanesi. L. (2016). Genomic selection in pigs: state of the art and perspectives. Italian Journal of Animal Science. 15, 211–232. https://doi.org/10.1080/1828051X.2016.1172034.

Sargolzaei, M., Jansen, G.B., Schenkel, F.S. (2014). A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 15(1), 478. https://doi.org/10.1186/1471-2164-15-478.

Shashkova, T.I., Martynova, E.U., Ayupova, A.F., Shumskiy, A.A., Ogurtsova, P.A., Kostyunina, O.V., and et al. (2020). Development of a low-density panel for genomic selection of pigs in Russia. Translational Animal Science. 4(1), 264–274. https://doi.org/10.1093/tas/txz182.

Shi, S., Yuan, N., Yang, M., Du, Z., Wang, J., Sheng, X., and et al. (2018). Comprehensive Assessment of Genotype Imputation Performance. Human Heredity. 83(3), 107-116. https://doi.org/10.1159/000489758.

Taylor, J.F. Schnabel, R.D. Sutovsky P. (2018). Genomics of Bull Fertility. Animal. 12 (Suppl1), s172–s183. https://doi.org/10.1017/S1751731118000599

Technow, A.F. (2015). Hypred: Simulation of Genomic Data in Applied Genetics. R package version 0.5. Available at: ⟨http://CRAN.Rproject. org//web/packages/ hypred⟩.

Ullah, E., Mall, R., Abbas, M.M., Kunji, K., Nato Jr, A. Q., Bensmail, H., and et al. (2019). Comparison and assessment of family- and population-based genotype imputation methods in large pedigrees. Genome Research. 29 (1), 125-134. https://doi.org/10.1101/gr.236315.118.

Wang, X., Zhang, Z., Morris, N., Cai, T., Lee, S., Wang, C., and et al. (2017). Rare variant association test in family-based sequencing studies. Briefings in Bioinformatics. 18 (6), 954-961. https://doi.org/10.1093/bib/bbw083.

Willer, C.J., Li, Y., Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 26 (17), 2190–2191. https://doi.org/10.1093/bioinformatics/btq340.

Camera Ready