IJBPAS, May, 2015, 4(5): 2591-2603

ISSN: 2277–4998

International Journal of Biology, Pharmacy and Allied Sciences (IJBPAS)

'A Bridge Between Laboratory and Reader'

www.ijbpas.com

INVESTIGATE THE ACCURACY OF IMPUTATION OF LOW –DENSITY GENOTYPED AND NON-GENOTYPED INDIVIDUALS USING GENOTYPE INFORMATION FROM RELATIVES

KAMAEI M[1], HONARVAR M[2], AMINAFSHAR M[1*], HOSSEINPOUR MASHHADI M[3] AND EMAM JOMEKASHAN N[1]

1: Department of Animal Science, Science and Research Branch, Islamic Azad University, Tehran, Iran

2: Department of Animal Science, Shahr-e-Qods Branch, Islamic Azad University, Tehran, Iran

3: Department of Animal Science, Mashhad Branch, Islamic Azad University,Mashhad, Iran

*** Corresponding Author: E Mail: aminafshar@gmail.com; Tel: 00989203224852**

ABSTRACT

The population structures observed in many livestock species are often characterized by large full- and half-sib families, and by the presence of animals (especially males) with a very large number of progeny. These conditions make it possible to infer the genotype of anon-genotyped individual using genomic information from its family members. The objectives of this study were to investigate the accuracy of imputation of low-density genotyped offspring with low-density genotyped and non-genotyped dams using genotype information from relatives. A population consists of 510 individuals and 85 families was formed. Each family contains Maternal Grand Sire (MGS), dam, Sire 1, Sire 2, Offspring 1 and Offspring 2.To assess the accuracy of imputation of non-genotyped and low-density genotyped individuals, 6 different scenarios were considered on the basis of their relatives genotypes. In low-density genotyped individuals (50, 90 and 95 percent) genotypes with evenly space were missed. The software Beagle was used for imputation. When dam is completely genotyped average of imputation accuracy in offspring with 50 percent determined genotype was 0.96 and in conditions that dam has 50 percent determined genotype, the imputation accuracy decreased to 0.87.in the first scenario that dam and one of the offspring are completely genotyped there are the most amount of imputation accuracy.

Keywords: Imputation, Low-Density, Accuracy, Scenario

INTRODUCTION

Prediction of breeding values of animals using genomic information was proposed by Meuwissen *et al*., 2001 **[1]** and since then the way breeding programs of livestock are conducted has changed considerably. Development of low-density single nucleotide polymorphism (SNP) panels will allow the extension of genomic selection to a larger portion of the population **[2]**. Prediction of un-genotyped markers, called imputation, is a strategy that allows using the same low-density chips for all traits (and for different breeds) **[3]**. Due to recent advances in genotyping technologies, the amount of genomic information available for genomic selection (GS) has increased from a few thousand **[4]** to 50k **[5]** and 800k **[6]** single nucleotide polymorphism (SNP) markers and today tends towards whole-genome sequence **[7]**. Genomic selection combines information on genotypes, phenotypes and pedigree to increase the accuracy of the estimatedbreeding values (EVBs) **[8]**.

In the SNP genotype data obtained from the SNP chip technique, missing genotype information is a common phenomenon that leads to a low call rate for some SNPs and for some animals.Imputation can be used to deduce the missing genotypes and could be helpful in increasing the accuracy of genomic selection **[9]**.

If a relevant genotyping strategy can be chosen such that imputation accuracy is sufficiently high, imputation of un-genotyped animals might also be of interest for breeding programs to reduce genotyping costs **[10]**. The population structures observed in many livestock species are often characterized by large full- and half-sib families, and by the presence of animals (especially males) with a very large number of progeny. These conditions make it possible to infer the genotype of anon-genotyped individual using genomic information from its family members. Several software programs for imputation are available some programs were designed for human populations and others for livestock populations. Comparisons of imputation programs have been mostly carried out for situations in which low-density genotyped individuals are imputed to high-density genotypes **[11]**. The performance of different imputation programs depends mostly on the data structure, density of single nucleotide polymorphism (SNP) panels, size of the reference population, and whether related or unrelated individuals were genotyped **[12-13]**. The two primary categories of imputation algorithms are

population-based and family-based. Pedigree-based imputation programs use mainly family information for imputation such as FImpute.

BEAGLE uses population information for imputation. Therefore it is expected that these program will be able to impute genotypes of animals with incomplete pedigree, which are the animals that are not imputed by FImpute [14].

BEAGLE imputed missing genotypes of both animals with and without complete pedigree with high accuracy **[15]**.

Erbe *et al*., 2012 **[16]** used the software BEAGLE without pedigree information to impute genotypes at 800 k SNPs from dairy bulls genotyped at 50 k and reported accuracies of imputation (defined as the proportion of correctly imputed genotypes) ranging from 0.96 to 0.98 in Jersey and Hosltein cattle, respectively. Meuwissen and Goddard., 2010 **[17]** applied a method for imputing whole sequence genotypes on individuals genotyped at a low density panel and reported that 10% of the missing genotypes were erroneously imputed.

The objective of this study was to investigate the accuracy of imputation of low-density genotyped offspring with low-density genotyped and non-genotyped dams using genotype information from relatives.

MATERIAL AND METHODS

Genome Simulation

A genome which consisted of 5 chromosomes, each had 100 cM length and the number of markers 1000, 5000 and 10000 and the number of QTLs 25, 50 and 100 and heritability 25 percent were simulated, mutation rate was considered 2.5×10^{-8}.

Population Simulation

A small population that consisted individuals (50 male and 50 female) were simulated. To make linkage disequilibrium, Villumsen *et al*., 2009 **[18]** method was used. In this method, 2 factors of small population and drift were made to make linkage disequilibrium. Such population structure continues for 50 generations with random mating to make required linkage disequilibrium. Maternal and Paternal haplotypes of each individual formed randomly and on the basis of probability of recombination and regarding markers distances. After 50 generations of random mating, population size developed to 1000 individuals (500 male and 500 female) in 51 generation.

Regarding that the aim is to determine individuals genotype on the base genotyped relatives a population consist of 510 individuals and 85 families was formed. Each family contains Maternal Grand Sire (MGS),

dam, Sire 1, Sire 2, Offspring 1 and Offspring 2. Totally, 1020 haplotypes were simulated, which 340 of its haplotypes are related to 2 Offspring.

Scenarios

To assess the accuracy of imputation of non-genotyped and low-density genotyped individuals, 6 different scenarios were considered (**Figure 1**). In low-density genotyped individuals (50, 90 and 95 percent) genotypes with evenly spaces were missed. In the first scenario, dam in reference population is genotyped for all SNPs, one of the 2 half-sib offspring of this dam is completely genotyped and the second offspring is genotyped with low density. In the second scenario, dam is completely genotyped, and 2 half-sib offspring of this dam are low-density genotyped. In first and second scenarios, initially, family structure which consists of one dam, 2 sires and 2 offspring was formed, and then genotype of second offspring in the first scenario and genotypes of both offspring in the second scenario (50, 90 and 95 percent) were missed. In the third scenario, dam is low-density genotyped, first offspring is completely genotyped and the second offspring is low-density genotyped. But dam is low-density genotyped, first dam's genotype should be determined, and then using of genotype information from dam, second sire and first half-sibs offspring, genotypes of second offspring will be determined. Therefore, to determine dam's genotype, genotype information of dam close relatives is needed, here only information of offspring's maternal grand sire (MGS) and first offspring is available.

The fourth scenario is similar to the third scenario with this difference that both half-sib offspring are low-density genotyped. As a result to determine dam's genotype just genotype information of MGS is completely available, first dam's genotype were imputed by information of her father, and then by completing dam's genotype and having genotype information of two sires, genotypes of two offspring are determined. In fifth and sixth scenarios, dam is non-genotyped. In fifth scenario the first offspring of this dam completely genotyped and the second offspring was low-density genotyped, but in sixth scenario both offspring were low-density genotyped.

Imputation

Beagle is able to impute missing genotypes of both groups of animals with and without completely pedigree. In this study that aim is imputing of low-density genotyped individuals with low-density and non-genotyped dams, beagle is proper software.

Assessing imputation accuracy

Individuals in test group were 2 dam's half-sib offspring who were low-density genotyped. After imputation of genotypes in each scenario, in order to assess the accuracy of imputation, a comparison between imputed genotypes and real genotypes were carried out and the amount of genotypic correspondence for the means of all individuals was calculated. For each scenario, means of imputation accuracy, standard deviation, percentage of correct and incorrect imputed SNP genotypes of each individual with 10 replicates were calculated.

RESULTS

The results of each scenario indicated that imputation of genotyped individuals was possible on the basis of their family relations.

Imputation accuracy of non-genotyped and low-density genotyped dams

Table 1 shows the imputation accuracy and calculated SD of non-genotyped and low-density genotyped dams. The range of imputation accuracy varied from 0.76 to 0.91. In a condition with non-genotyped dam, the least amount of imputation accuracy and most SD were achieved. The most amount of imputation accuracy was related to genotyped dams with 50 percent determined genotype, as missing SNPs in dams increased, the imputation accuracy amount decreased. Table 2 provides a total comparison between different scenarios and presents percentage of correct and incorrect imputed genotypes in each scenario. In the fifth scenario which dam is non-genotyped and one of the offspring is low-density genotyped, the percentage of correct imputed genotypes in conditions that 50, 10 and 5 percent of offspring's genotypes are determined, were reported as 80.42, 76.8 and 75.16 percent. whereas, in the third scenario in a situation that 50 percent of dam's genotype are determined, the percentage of correct imputed genotypes in 50, 10 and 5 percent of determined genotypes were 87.5, 83.12 and 77.61 percent and in 5 percent of determined genotypes were 79.9, 76.9 and 76.8 percent orderly. In addition, in **Table 2**, the difference between imputation accuracy in sixth and fourth scenarios showed that in fourth scenario that dam is low-density genotype, the average of imputation accuracy of offspring in conditions that dam had 50 percent determined genotype, was 0.83±0.03 and in the sixth scenario which dam is non-genotype, it was 0.73. But in the fourth scenario in condition that dam and offspring have 5 percent determined genotype this amount was 0.72±0.07. In the sixth scenario, in condition that offspring has 5 percent determined genotype, this amount reached to 0.67±0.13. Average of imputation accuracy and percentage of correct imputed genotypes

increased in all scenarios when dam and one of the offspring are genotyped. For example, in the first scenario that dam and one of the offspring are completely genotyped there are the most amount of imputation accuracy and percentage of correct imputed genotype. As missing SNPs in offspring increased, imputation accuracy reduces clearly. In the second scenario that both offspring have missing genotypes, the accuracy of imputation and percentage of correct imputed genotypes decreased in comparison with the first scenario. As missing SNPs in dam increased, average of imputation accuracy in offspring decreases. When dam is completely genotyped, average of imputation accuracy in offspring with 50 percent determined genotype was 0.96±0.004 and in conditions that dam has 50 percent determined genotype, the imputation accuracy decreased to 0.87±0.01. Such reduction of imputation accuracy in offspring is observed in by increasing missing genotypes in dam.

Table 2 indicates that there is a different between imputations of those who are low-density genotyped and is dependent on the amount of missing SNPs. As the numbers of missing SNPs are lower, the imputation accuracy will be higher.

Figure 2 shows imputation accuracy for each animal in conditions that dam is completely genotyped, low-density genotyped and non-genotyped and animal is non-genotyped.

The average of animal imputation accuracy ranged from 0.70 to 0.85.The least of animal imputation accuracy was related to non-genotyped dams**.**

The effect of MAF on the imputation accuracy

The SNP imputation accuracy increased when the number of genotyped offspring increased (**Figure 3**). When dam and one of the offspring are genotyped, the accuracy of imputation is lesser dependent to MAF (the first scenario).

Figure 3 shows that SNP imputation accuracy increased significantly when dam is completely genotyped than conditions that dam low-density genotyped or non-genotyped. In this study the imputation accuracy only a few SNP had an accuracy imputation equal to 1.

DISCUSSION

The aim of this study is considering the imputation accuracy in both low-density genotyped offspring with completely genotyped, non-genotyped and low-density genotyped dams. The results show that higher accuracy achieved when dam and one of the two offspring are genotyped.

In papers, many definitions of imputation accuracy were used. Hickey *et al*., 2012 [19]

and Ma *et al*., 2013 [20] showed that percentage of correctly imputed SNPs are widely dependent to MAF, and correlation between real genotypes and imputed genotypes is more useful measurement of the quality of imputation.Beagle is software for imputing genotypes and non-genotyped individuals. This software is able to impute genotypes of animals with incomplete pedigree and makes complete outputs. Bouwman *et al*., 2014 **[21]** carried out imputation of non-genotyped individuals by Alphalmput software, their results showed that as genotyped offspring increased, the accuracy of imputation increases which a range of 0.57 in a condition without offspring reached to 0.92 with 4 offspring. Pimentel *et al*., 2013 **[22]** imputed non-genotyped individuals with one genotyped offspring, and reported that imputation accuracy ranged from 0.52 to 0.93 which is related to population structure and used method. Cleveland *et al*., **[23]** investigated use of phenotypes from imputed animals in a simulated population. Their method used segregation analysis and information on haplotype frequencies, and they reported a success rate of 69% when dams were completely un-genotyped.

CONCLUSION

Non-genotyped individuals of a population can be imputed and indicate its accuracy as the correlation between real genotypes and imputed genotypes, if genotypes of sire and maternal grand sire were available. Imputation accuracy increases with genotyped offspring, in conditions that dam is genotyped, imputation accuracy in offspring reached to 0.96. The imputation accuracy with non-genotyped dam decrease to 0.67, in conditions that genotype information of offspring is not completely available. As a result, imputation of non-genotyped individuals could help valuable phenotypes in genomic prediction or GWAS, specifically when genotyped offspring is available.

Table 1: Average imputation accuracy (r) for low-density genotyped and non-genotyped dam

low-density genotyped and non-genotyped dam	r[1]	SD[2]
low-density 50%	0.91	0.009
low-density 10%	0.82	0.04
low-density 5%	0.79	0.06
non-genotyped	0.76	0.07

[1]mean of imputation accuracy calculated as the correlation between true genotypes and Imputed genotype dosages, [2]Standard deviation

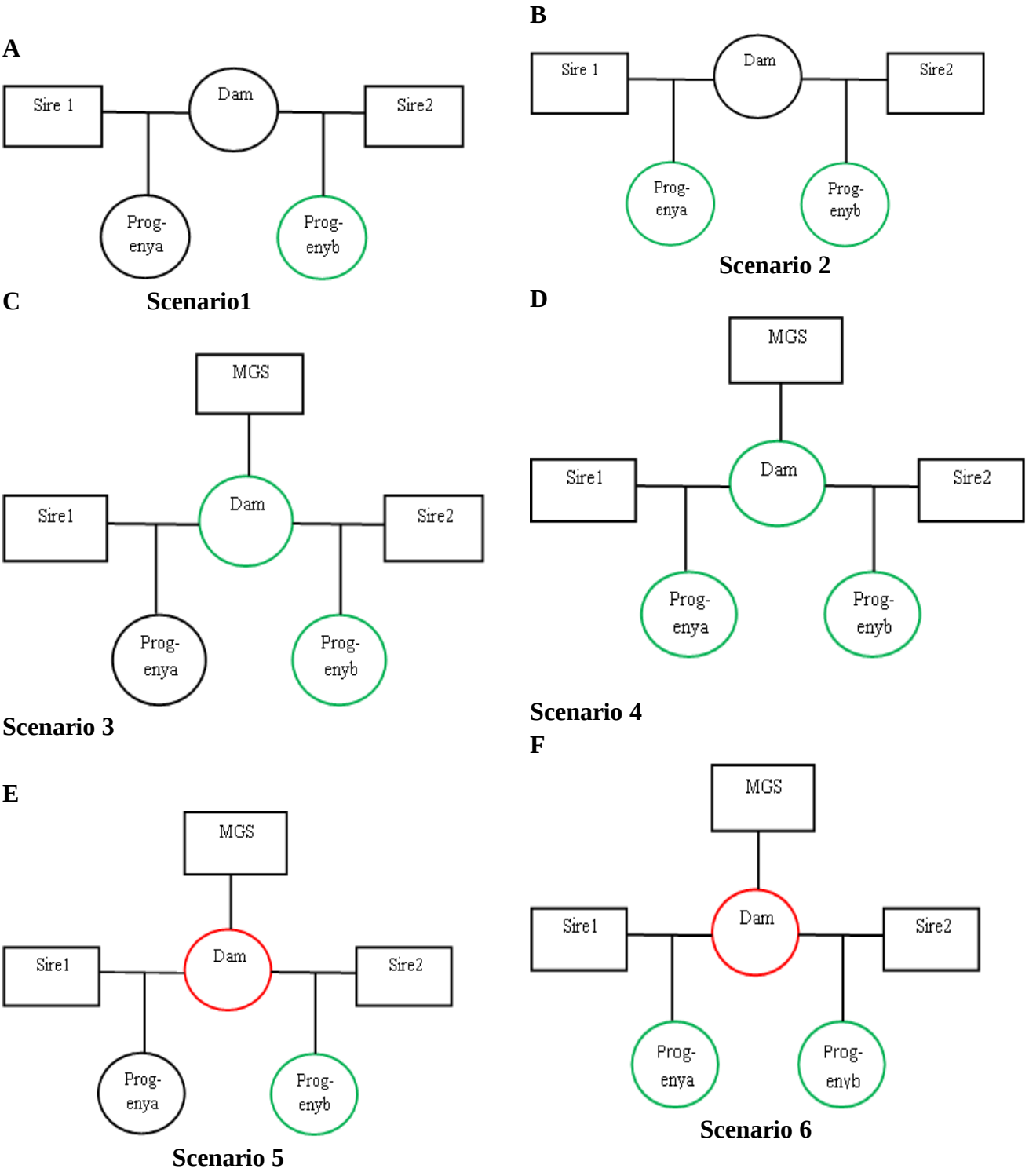
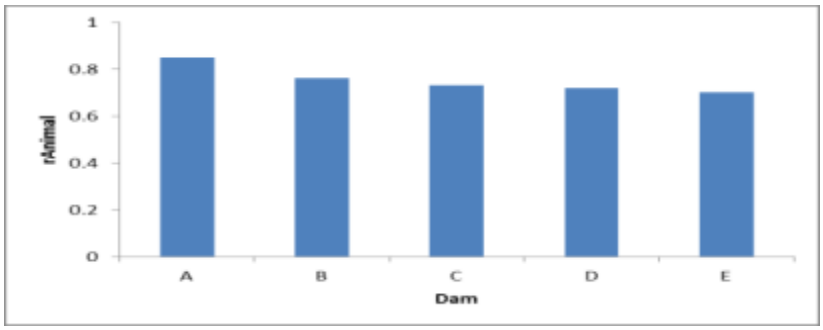


Figure 1: Assumed Family members with available genotypic information (black) used for imputing a low-density genotyped (green) or non-genotyped (red) individual

Table 2 Average imputation accuracy (r) and percentage of correct and incorrect imputed genotypes for offspring in each scenario

		r				correct				incorrect	
Scenarios	**[a]50%**	**[b]10%**	**[c]5%**		**[a]50%**	**[b]10%**	**[c]5%**		**[a]50%**	**[b]10%**	**[c]5%**
Scenario1	**0.96**	**0.92**	**0.9**		**96.17**	**91.76**	**91.47**		**3.82**	**8.23**	**8.52**
Scenario2	**0.88**	**0.84**	**0.8**		**87.64**	**85.29**	**80.29**		**12.58**	**14.70**	**19.70**
Scenario3, [d]dam50%	**0.87**	**0.83**	**0.78**		**87.5**	**83.12**	**77.61**		**12.5**	**16.88**	**22.39**
Scenario3, [e]dam10%	**0.81**	**0.79**	**0.77**		**80.78**	**79.41**	**77.14**		**19.22**	**20.59**	**22.85**
Scenario3, [f]dam5%	**0.8**	**0.77**	**0.77**		**79.9**	**76.9**	**76.8**		**20.1**	**23.1**	**23.2**
Scenario4, dam50%	**0.83**	**0.78**	**0.75**		**83.05**	**77.51**	**74.85**		**16.95**	**22.49**	**25.15**
Scenario4, dam10%	**0.75**	**0.73**	**0.73**		**74.7**	**73.14**	**72.91**		**25.3**	**26.86**	**27.09**
Scenario4, dam5%	**0.74**	**0.72**	**0.72**		**74.21**	**72.41**	**72.35**		**25.79**	**27.59**	**27.64**
Scenario5	**0.8**	**0.77**	**0.75**		**80.42**	**76.8**	**75.16**		**19.58**	**23.2**	**24.84**
Scenario 6	**0.73**	**0.69**	**0.67**		**73.10**	**68.88**	**67.34**		**26.9**	**31.11**	**32.66**

a- progeny with low-density 50%, b- progeny with low-density 10%, c- progeny with low-density 5%,d-dam with low-density 50%, e- dam with low-density 10%, f- dam with low-density 5%



A= full genotyped dam

B= low-density genotyped dam (50%)

C= low-density genotyped dam (10%)

D= low-density genotyped dam (5%)

E= non-genotyped dam

Figure 2: Animal imputation accuracy with full genotype, low-density genotype and non-genotype dam

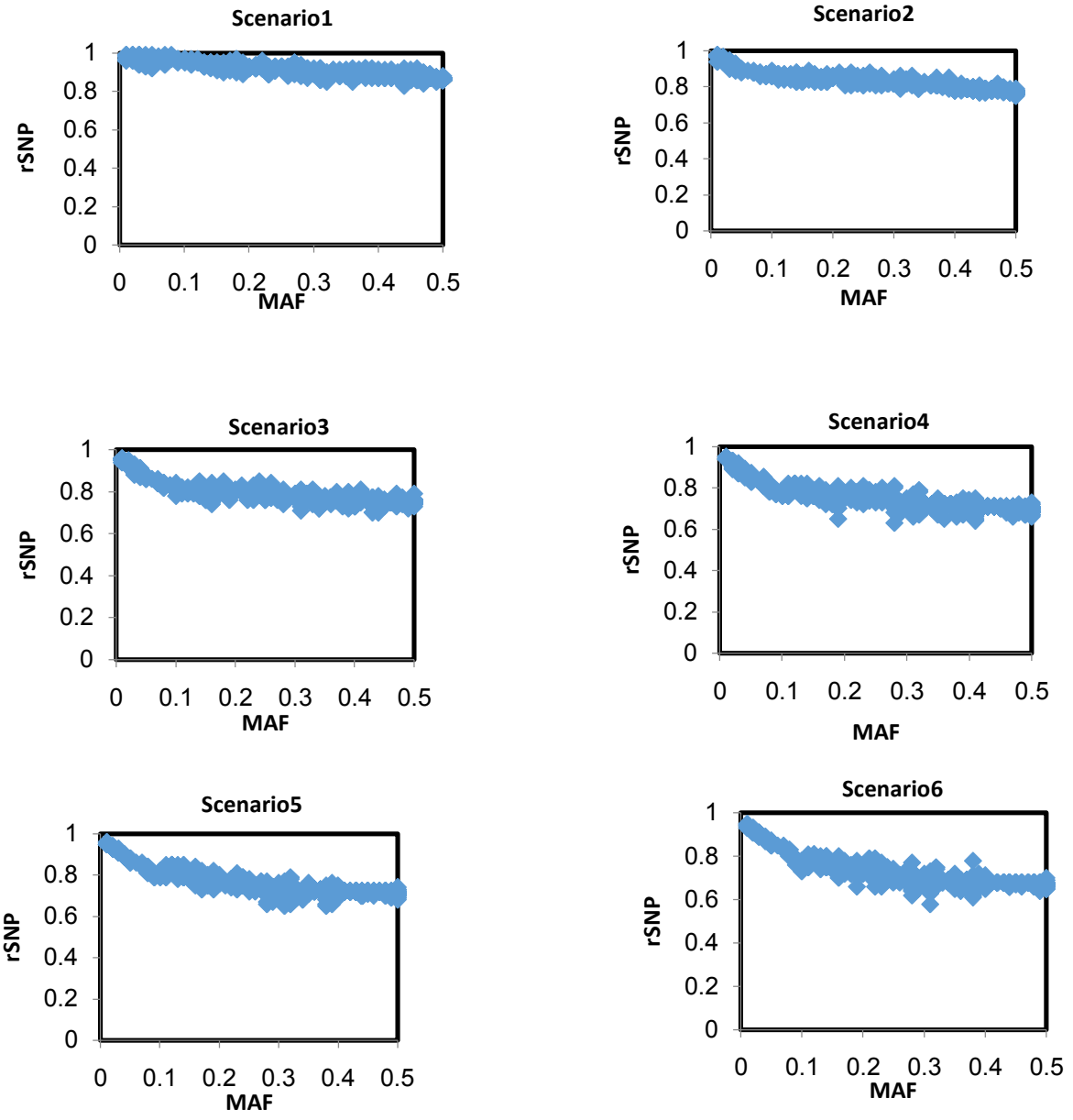


Figure 3: Imputation accuracy by SNP (rSNP) plotted against the minor allele frequency (MAF) in conditions that offspring in all scenarios and dam in third and fourth scenarios are low-density genotyped 5%

REFERENCES

[1] Meuwissen THE, Hayes BJ, Goddard ME, Prediction of total genetic value using genome-wide dense marker maps, Genetics, 157, 2001,1819–1829.

[2] Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE,

Genomic selection using different marker types and densities, Journal of Animal Science, 86, 2008, 2447-2454.

[3] Zhang Z, Druet T, Marker imputation with low-density marker panels in Dutch Holstein cattle, Journal of Dairy Science,93, 2010, 5487-5494.

[4] Sargolzaei M, Schenkel FS, Jansen GB, Schaeffer LR, Extent of linkage disequilibrium in Holstein cattle in North America, Journal of Dairy Science, 91, 2008, 2106–2117.

[5] Pimentel ECG, Erbe M, König S, Simianer H, Genome partitioning of genetic variation for milk production and composition traits in Holstein cattle, Front Genet, 2011, 2-19.

[6] Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, Mason BA, Goddard ME, Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels, Journal of Dairy Science, 95, 2012, 4114–4129.

[7] Ober U, Ayroles JF, Stone EA, Richards S, Zhu D, Gibbs RA, Stricker C, Gianola D, Schlather M, Mackay TFC, Simianer H, Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster, PLoS Genet, 2012, 8:e1002685.

[8] Weigel KA, Van Tassell CP, O'Connell JR, VanRaden PM, Wiggans GR. Prediction of unobserved single nucleotide poly-morphism genotypes of Jersey cattle using reference panels and population based imputation algorithms. Journal of Dairy Science, 93, 2010, 2229-2238.

[9] Weng Z, Zhang Z , Ding X, Application of imputation methods to genomic selection in Chinese Holstein cattl,Journal of Animal Science, 2013, 3-6.

[10] Williams AL, Patterson N, Glessner J, Hakonarson H, Reich D, Phasing of many thousands of genotyped samples, Am J Hum Genet, 91, 2012, 238–251.

[11] Pei YF, Li J, Zhang L, Papasian CJ, Deng HW. Analyses and comparison of accuracy of different genotype imputation methods,PLoS ONE, 2008, 3:e3551.

[12] Johnston J, Kistemaker G, Sullivan PG, Comparison of different

imputation methods, Interbull Bull, 44, 2011, 25–33.

[13] Ma P, Brøndum RF, Zhang Q, Lund MS, Su G, Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red cattle, Journal of Dairy Science, 96, 2013, 4666–4677.

[14] Johnston J, Kistemaker G, Success rate of imputation using different imputation approaches, CDN, 2011. http://www.cdn.ca

[15] Browning BL, Browning SR, A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals,Am J Hum Genet, 84, 2009, 210-223.

[16] Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, Mason BA, Goddard ME, Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels, Journal of Dairy Science, 95, 2012, 4114–4129.

[17] Meuwissen THE, Goddard ME, The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to wholegenome sequence density genotypic data. Genetics, 185, 2010, 1441–1449.

[18] Villumsen TM, Janss L, Lund MS. The importance of haplotype length and heritability using genomic selection in dairy cattle, Journal of Animal Breed Genet, 126, 2009, 3-13.

[19] Hickey JM, Crossa J, Babu R, de los Campos G, Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs, Crop Science, 52, 2012, 654–663.

[20] Ma P, Brøndum RF, Zhang Q, Lund MS, Su G, Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red cattle, Journal of Dairy Science, 96, 2013, 4666–4677.

[21] BouwmanAC, Hickey JM, Calus MP, Veerkamp RF, Imputation of non-genotyped individuals based on genotyped relatives: assessing the imputation accuracy of a real case scenario in dairy cattle. Genet SelEvol, 46, 2014, 6.

[22] Pimentel ECG, Wensch-Dorendorf M, Konig S, Swalve HH, Enlarging

a training set for genomic selection by imputation of un-genotyped animals in populations of varying genetic architecture, Genet SelEvol, 45, 2013, 45-12.

[23] Cleveland MA, Hickey JM, Kinghorn BP, Genotype imputation for the prediction of genomic breeding values in non-genotyped and low density genotyped individuals, BMC Proc, 2011, 5-S6.