Acta fytotechn zootechn, 19, 2016(Special Issue), 27-30
http://www.acta.fapz.uniag.sk

Short Communication

Imputation of missing genetic markers SNP using linear regression models

Anita Kranjčevičová*, Josef Přibyl

Institute of Animal Science, Praha–Uhříněves, Czech Republic

Article Details: Received: 2016-04-28 | Accepted: 2016-06-01 | Available online: 2016-09-01

http://dx.doi.org/10.15414/afz.2016.19.si.27-30

For imputation of missing SNP are used software products which require known relationship between genotyped individuals. In common breeding business the genotypes of parents are not always known. That is why our own methodological process was used. The aim of this study is to map the current research of genetic chips and to verify the calculation process. The testing was processed at chosen loci in two datasets and in 8 models with different amount of SNPs. For the dataset A was prediction of missing values almost accurate with model reliability 100 % with the exception of one homozygous locus where the reliability reached only 55 %. In the dataset B the most extensive model reached the reliability of 80–90 % even in case of homozygous loci. The prediction error value was higher than in the first case. It was proven that missing values prediction is possible to calculate using the neighbouring SNPs.

Keywords: GLM, imputation, linear model, SNP chips, SNP markers

1 Introduction

Working with genomic information in cattle breeding has become a standard procedure. These polymorfisms are used for evaluation of genomic relationship, prediction of genomic breeding values and for the evaluation of tested animals. The most common chips used for genotyping are Illumina and Affymetrix. Each company develops its own techniques of genotype obtaining. Affymetrix has unified coding type of SNPs among chips of different generations and thus even older data can be used. Illumina uses many coding types between different generations of chips. Thus, direct comparison of SNPs is not possible. Illumina has chips of different density and financial costingness. Illumina chips have become a standard all over the world and it is used by all breeding companies. The most used software programs for imputations are Beagle (Browning et Browning, 2007), AlphaImpute (Hickey et al., 2012), Impute 2 (Howie and Marchini, 2009), DAGPHASE (Druet and Georges, 2010), FImpute (Sargolzaei et al., 2008), PedImpute (Nicolazzi et al., 2013) and MaCH (Li et al., 2010).This study is focused on completion of missing genetic markers – SNPs (single nucleotide polymorphisms) – on genetic chips. More specifically completion of missing values in datasets which contain pieces of information about SNP occurence in cattle genome. It was developed our own methodology because the genotypes of parents were missing and also allele coding was incomplete. The aim of this study was to map the current research of genetic chips and to verify the calculation process.

2 Material and methods

2.1 Data

Dataset A contained 260 bull genotypes of different dairy breeds from the Czech Republic. Dataset B contained 3982 genotypes of pure Holstein bulls from nine countries.

2.2 Dataset preparation

For the marking of the tested SNPs were used three numbers according to the genotype (0 = BB, 1 = AB, 2 = AA). Three loci (located on chromosome 1) from each dataset were chosen for testing according to the percentage rate of allele A.

* **Corresponding Author:** Anita Kranjčevičová. Institute of Animal Science, Přátelství 815, 104 00, Praha–Uhříněves, Czech Republic. E-mail: kranjcevicova.anita@vuzv.cz

Dataset A:
Locus 201: 50 % of allele A and average value of the locus 1.05 (heterozygous locus)
Locus 716: 75 % of allele A and average value of the locus 1.5
Locus 133: 95 % of allele A and average value of the locus 1.9 (almost homozygous locus)
Dataset B
Locus 760: 50 % of allele A and average value of the locus 1.04 (heterozygous locus)
Locus 893: 75 % of allele A and average value of the locus 1.5
Locus 201: 95 % of allele A and average value of the locus 1.9 (almost homozygous locus)

2.3 Statistical methods

In total, 8 models was used for the testing of both datasets. Each model was different in number of used neighbouring loci (10–100 loci). The number of neighbouring loci was determined on the basis of assumption that the loci are all inherited together and there is no crossing-over in the particular area. The largest model obtained 100 loci which means 50 loci from the left side and 50 from the right side of the tested locus. These loci were used for calculation of regression coefficients, that were used for backward prediction of tested loci. Testing was processed in SAS analytical software using GLM procedure. The following model equation was used:

$$li = \mu + ll_{(i-j)} + ll_{(i-j+1)} + \ldots + ll_{(i-1)} + l_{(i+1)} + \ldots + l_{(I+j-1)} + l_{(i+j)} + \varepsilon$$

Where li is tested locus; μ is mean; ll is locus on the left side of the tested locus; l is locus on the right side of the tested locus; i is number of tested locus; j is number of used neighbouring loci; ε is error. With more loci better results could be obtained but the bigger amount of data causes higher costingness of the calculations.

3 Results and discussion

The testing of every model indicated that the prediction of SNPs was the most successful at heterozygous locus with 50% rate of allele A. Only 50 neighbouring loci was enough for almost precise prediction of SNP (locus 201). In locus with 75% rate of allele A (locus 716) were obtained the same results when 100 loci was used. At almost homozygous locus (locus 133) with 95% rate of allele A was achieved only 56 % of reliability in the largest model (100 loci).

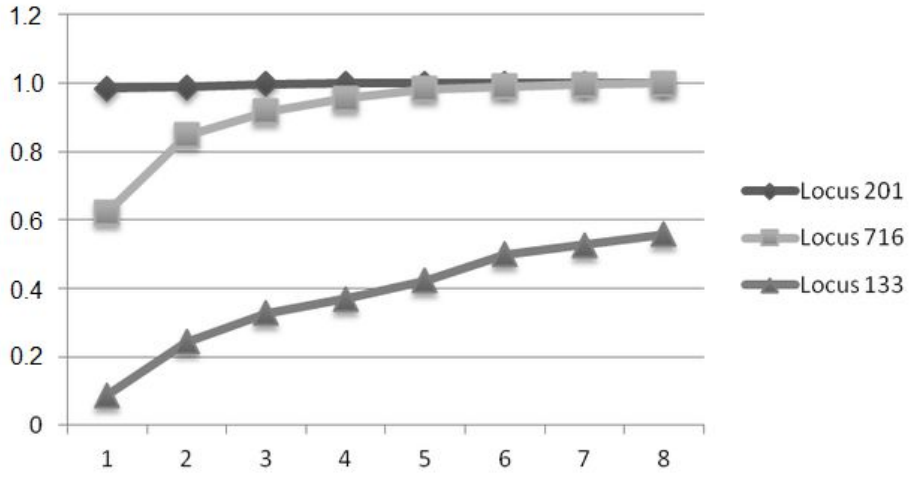

Figure 1 Increasing reliability (R^2)of every model for dataset A

In locus 760 with 50% rate of allele A the reliability reached 47–88%. The best rate of reliability was obtained in locus 893 (90–96 %). In almost homozygous locus 201 the reliability reached 41–88 %.

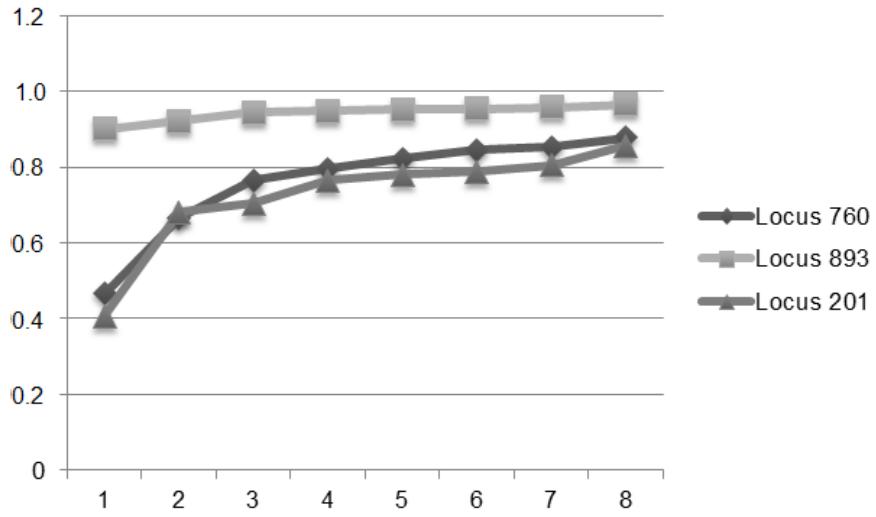

Figure 2 Increasing reliability (R^2) of every model for dataset B

Table 1 Values of maximal absolute error for every model

	Model 1 (10 loci)	Model 2 (20 loci)	Model 3 (30 loci)	Model 4 (40 loci)	Model 5 (50 loci)	Model 6 (60 loci)	Model 7 (70 loci)	Model 8 (100 loci)
Dataset A								
Locus 201	0.786	0.589	0.196	0.129	3.5E-13	1.4E-12	1.4E-12	4.0E-12
Locus 716	1.453	0.880	0.801	0.568	0.325	0.254	0.197	0.064
Locus 133	0.992	0.942	0.855	0.819	0.797	0.740	0.754	0.729
Dataset B								
Locus 760	1.742	1.834	2.067	1.904	1.659	1.765	1.543	1.407
Locus 893	1.033	1.059	1.129	1.174	1.069	1.050	0.940	0.995
Locus 201	1.415	1.347	1.392	1.389	1.228	1.141	1.106	0.964

In value 0 was obtained conformity of prediction with real value approximately 10 %, for value 1 – 50 % and for value 2 – 65 %.

The values of maximal absolute error were bigger in dataset B but on the other side the values of reliability were more balanced in comparison with dataset A. The reason of these differences could be caused by using of different animals and different tested loci in each dataset.

Our results are not comparable with other studies because we developed our own methodology. We could not use any program commonly used for imputations because our database was not tailored to these softwares. If we had all pieces of information needed for the programs the best option for us would be Beagle and Impute 2 (Browning and Browning, 2007; Howie and Marchini, 2009) because these programs do not need genotypes connected with pedigree data for correct calculation.

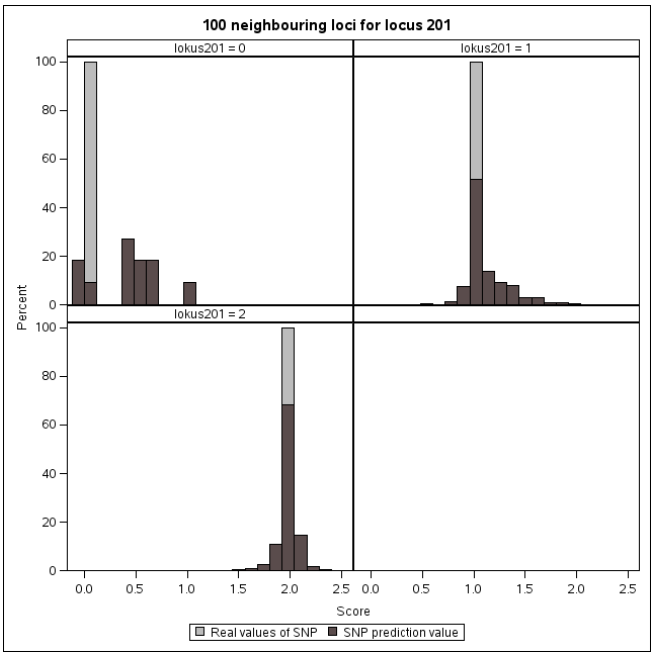

Figure 3 Conformity of prediction and real value for locus 201 in 8th model (100 loci)

4 Conclusions

It was proven that missing values prediction is possible to calculate using the neighbouring SNPs. For the calculations were excluded loci with more than 5 % of missing data values and individuals with more than 10 % of missing data values.

Acknowledgments

The research was supported by the project NAZV Q1 1510139.

References

BROWNING, S. R. and BROWNING, B. L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics,* vol. 81, no.5, pp.1084–1097. doi: http://dx.doi.org/10.1086/521987

DRUET, T. and GEORGES, M. (2010) A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics*, vol. 184, no. 3, pp. 789–798. doi: http://dx.doi.org/10.1534/genetics.109.108431

HICKEY, J. M. et al. (2012) A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. *Genetics Selection Evolutiom* vol. 44, no. 1/9, 11 p. doi: http://dx.doi.org/10.1186/1297-9686-44-9

HOWIE, B. and MARCHINI, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, vol. 5, no. 6, e1000529. doi: http://dx.doi.org/10.1371/journal.pgen.1000529

LI, Y. et al. (2010) MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes*. Genetics Epidemiology, vol.* 34, no. 8, pp. 816–834. doi: http://dx.doi.org/10.1002/gepi.20533

NICOLAZZI, E. L., BIFFANI, S. and JANSEN, G. (2013) Short communication: Imputing genotypes using PedImpute fast algorithm combining pedigree and population information. *Journal of Dairy Science*, vol. 96, no. 4, pp. 2649–2653. doi: http://dx.doi.org/10.3168/jds.2012-6062

SARGOLZAEI, M. et al. (2008) Extent of linkage disequilibrium in Holstein cattle in North America. *Journal of Dairy Science*, vol. 91, no. 5, pp. 2106–2117. doi: http://dx.doi.org/10.3168/jds.2007-0553