Effect of genotype imputation on integrated model for genomic selection

S. GUHA MAJUMDAR, D. C. MISHRA AND A. RAI
Centre for Agricultural Bioinformatics,
ICAR-IASRI, New Delhi-110012

Received : 26.11.2019 ; Revised : 16.03.2020 ; Accepted : 29.03.2020
DOI: https://doi.org/10.22271/09746315.2020.v16.i1.1283

ABSTRACT
Genomic selection is a very recent area of study in case of molecular breeding of livestock or crop species. There are various statistical models available for genomic selection. The performances of these models depend on several factors like sampling population, genetic architecture of target species, statistical models as well as missing genotypes. Missing genotype is very common problem in high throughput sequencing data. These missing genotypes are necessary to be imputed in order to implement the genomic selection models. Different statistical models of genomic selection behave differently in imputed data. So, it is highly imperative to evaluate the performances of statistical models under different levels of imputations to know the behavior of the models. In this article, performance of three statistical models viz. Sparse Additive Models (SpAM), Hilbert-Schmidt Independence Criterion Least Absolute Shrinkage and Selection Operator (HSIC LASSO) and Integrated Model for genomic selection are compared after incorporating the various degree of imputation (0, 2, 5 and 10%) in the real data. Results indicate that integrated model is found to be more robust against the level of imputation of the genotypic data.

Keywords: Genomic selection, HSIC LASSO, imputation, integrated model, missing genotype, SpAM

Genomic selection (GS), a new improved variant of marker-assisted selection (MAS) for breeding of livestock and crop species, has been used globally for increasing agricultural production and productivity. GS was first introduced by Meuwissen et al. (2001). High density markers covering whole genome were used to estimate the Genomic Estimated Breeding Values (GEBV) in GS. Thousands of Single Nucleotide Polymorphisms (SNPs) are chosen to represent the whole genome by assuming that must be at least one SNP in close proximity to the particular gene or QTL which is in linkage disequilibrium with this particular gene or QTL of interest. Individual searching for significant QTL–marker loci associations is not required in GS. Instead, GS considers simultaneously all the markers as predictor variables. The primary advantage of GS is it can accelerate breeding cycles so that the rate of annual genetic gain in terms of time and cost can be enhanced as GS can be implemented in very early life of an individual. Furthermore, it can be extended to any trait as long as the record of reference population for that trait is present. The implementation of GS is beneficial for plants, although it has been applied mainly for animals, especially for dairy cattle. There are several evidences of the increase in accuracy of selection for crop species, like wheat (Triticum aestivum) (Daetwyler et al., 2014), maize (Zea mays) (Zhao et al., 2012), rice hybrids (Oryza sativa L.) (Xu et al., 2014), barley (Hordeum vulgare) (Lorenz et al., 2012) etc. Many statistical models are available in literature for GS, such as Best Linear Unbiased Prediction (BLUP) (Henderson, 1975), Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996), ridge regression (Hoerl and Kennard, 1970), Linear Least Squared Regression (Meuwissen et al., 2001), Sparse Additive Models (SpAM) (Ravikumar et al., 2005 and 2009), HSIC LASSO (Gretton et al., 2005 and Yamada et al., 2014), Bayes A, Bayes B (Meuwissen et al., 2001) etc. Most of them are used to capture additive genetic effects. However, some models are also available which can be used for modeling non-additive genetic effects i.e. epistasis.

Recent advances in genotyping technology have facilitated the availability of high density genotyping data, which makes it easy to implement genomic selection in breeding. But missing genotype is quite common in genomic data. There are possibilities to underestimate GEBV due to these missing genotypes. To avoid such situations, we need to impute the missing genotypes. Several types of imputation methods are available for genotype imputation. Burdick et al. (2006) extended the idea of imputation of missing genotypes through computational analyses. Burdick was the first to coin the term “in silico genotyping” to explain the idea that genotype imputation could be performed through computational analyses by replacing laboratory based procedures. For related individuals, family members share long stretches of haplotype that are identical-by-descent. In such cases, imputation of missing genotypes can be performed by considering the
distribution of potential genotypes of each individual jointly with that of other individuals in the same pedigree. For unrelated individuals, the haplotypes of the individuals over short regions of sequence will be related to each other by being identical by descent (IBD). Imputation can be performed by identifying the similarity between the haplotypes of the study individuals and the haplotypes in the reference population. Then this sharing is used to impute the missing alleles in individuals under study (Marchini and Howie, 2010). Genotype imputation can be performed in silico with packages such as MENDEL (Lange et al., 2005) and MERLIN (Abecasis et al., 2002). The Lander-Green (Lander et al., 1987) or Elston-Stewart (Elston et al., 1971) algorithms are used to implement these computational tools. Some tools are also based on Monte Carlo sampling (Heath, 1997). There are various imputation tools to analyse genotype imputation. These tools include IMPUTE (Marchini et al., 2007), MaCH (Li et al., 2010) and fastPHASE/BIMBAM (Scheet et al., 2006, Servin et al., 2007), TUNA (Nicolaie et al., 2006), BEAGLE (Browning et al., 2018), PLINK (Purcell et al., 2007) and WHAP (Zaitlen et al., 2007) etc. Performance of different statistical models of GS may be affected due to extent of the imputation on missing genotypic data. Therefore, there is a need to evaluate the performance of the GS models at different level of imputation of the genotypic data. Chen et al. in 2014 have compared the performance of GBLUP and Bayesian methods for genomic prediction in case of milk yield, fat percentage, protein percentage and somatic cell score. They have shown that Bayesian methods are more prone to imputation error than GBLUP. Also, it is found that imputations from lower density SNP panels have lower accuracy of genomic prediction than higher density SNP panels. Earlier in 2012, Mulder et al. have investigated the accuracy of imputation in case of low-density chip in animals and found that imputation error rate is higher in low-density SNP than high density SNP. Weigela et al. (2010) also studied the effect of imputation in dairy cattle using Bayesian least absolute selection and shrinkage operator (LASSO) method and found that if a suitable reference population with high-density genotypes is available, the imputation error reduced in a low-density chip comprising 3,000 equally spaced SNP. In our previous work we have developed an integrated model framework by combining two efficient additive and non-additive methods i.e. SpAM and HSIC LASSO, for GS to estimate GEBV (Guha Majumdar et al., 2019). In this article, an attempt has been made to evaluate the performance of the integrated model framework for GS under the influence of different levels of imputation in genotypic data.

**MATERIALS AND METHODS**

Integrated GS model has been implemented on the real dataset of wheat available at https://www.genetics.org/content/186/2/713.supplemental. This dataset includes trait grain yield (GY) of 599 lines for four mega environments. However, for our convenience we have just considered GY for first mega environment. The genotyping of wheat lines was done using 1447 Diversity Array Technology markers generated by Triticarte Pty. Ltd. (Canberra, Australia; http://www.triticarte.com.au). Total 1279 number of DArT markers has been used in this study (Crosa et al., 2010).

At first the integrated model for GS has been applied to the real dataset of wheat in R statistical computing platform. For that purpose, in-house R package “GSelection” has been developed and used. The package “GSelection” (Guha Majumdar et al., 2019) is now available at CRAN (https://cran.r-project.org/web/packages/GSelection/index.html). The integrated model(Guha Majumdar et al., 2019) for estimation of GEBV can be expressed as

\[ y_{\text{pred}} = w y_{w} + (1-w) y_{HL}. \]

where, \( y_{\text{pred}} \) is the predicted phenotype (GEBV) of the integrated model, \( w = \frac{\sigma_{\text{HL}}^2}{\sigma_{\text{HL}}^2 + \sigma_{\text{sp}}^2} \), where \( \sigma_{\text{HL}}^2 \) and \( \sigma_{\text{sp}}^2 \) are the error variances of models HSIC LASSO and SpAM respectively, \( y_{w} \) is the predicted GEBV from Sparse Additive Models and \( y_{HL} \) is the GEBV from HSIC LASSO. The performance of this model is evaluated based on three criteria, viz. error variance of estimating GEBV, prediction accuracy (PA) of GEBV, redundancy rate (RED) among the selected markers. Let the error variance is denoted by. The can be expressed as

\[ \sigma_{\text{int}}^2 = \frac{\sigma_{\text{sp}}^2 \sigma_{\text{HL}}^2}{\sigma_{\text{sp}}^2 + \sigma_{\text{HL}}^2} + \left( \frac{\sigma_{\text{sp}}^2 + \sigma_{\text{HL}}^2}{\sigma_{\text{sp}}^2 + \sigma_{\text{HL}}^2} \right)^2 \sigma_{\text{HL}}^2. \]

The estimation of \( y_{\text{int}} \) and can be performed by following refitted cross validation approach of Fan et al., 2012. PA can be defined as the correlation between the actual phenotypic \( y_{\text{real}} \) (values (Howard et al., 2014). Prediction accuracy (PA) = correlation \((y_{\text{real}}, y_{\text{pred}})\). The RED score (Zhao et al., 2010) can be obtained by

\[ \text{RED} = \frac{1}{m(m-1)} \sum_{u_k \neq u_j, y_k > y_j} |\rho_{k1}| \]
where, $\rho_{kj}$ is the correlation coefficient between the $k$-th and $l$-th markers, is the number of selected markers. A high RED score indicates strong correlation between selected markers, which is not desirable because it means many redundant markers are selected. In order to estimate the prediction accuracy (PA), redundancy rate (RED) and error variance, the original dataset has been divided into training and testing datasets. 80% of the individuals were selected randomly as training data and remaining 20% data were kept for testing purpose.

In next step, 2% of the marker data were removed randomly from the dataset. Then the `sing.im` function from R package "linkim" (Xu et al., 2014) has been used for imputation of those missing marker data. Here, we assume Hardy-Weinberg equilibrium for all loci. So, for standard case of 3 genotypes, i.e. heterozygous genotypes, values are sampled from distribution $P(x=AA=0)=(1-p)^2$, $P(x=Aa=1)=2p(1-p)$ and $P(x=aa=2)=p^2$, where, AA is homozygous dominant allele, Aa is heterozygous allele and aa is the homozygous recessive allele. In other words, the missing values of the datasets are imputed based on observed data proportions. Then the integrated model has been applied on the imputed (2%) dataset using R package “GSelection” following the same procedure as in case of original dataset to estimate the PA, RED and error variances. Likewise 5 and 10% of the original marker data were eliminated respectively in the subsequent steps and the same procedures were followed to compare the parameters.

RESULTS AND DISCUSSION

The impact of different extent of imputation (i.e. 2%, 5 and 10%) has been studied on developed integrated model by comparing estimated parameters with the actual dataset. The comparison of the performance of actual dataset and the imputed datasets are shown in the Fig. 1, 2 and 3.

In Fig. 2, it is observed that the RED score of imputed datasets in all the three models (i.e. integrated model, SpAM and HSIC LASSO) fluctuates from the RED score of original data (data without imputation). Fluctuation in case of SpAM and HSIC LASSO is very large in comparison with the integrated model. Thus the integrated model shows robustness against missing genotypes in this case also by showing less deviation of the RED score from the original. Considering the error variance of estimating GEBV, Fig. 3 shows that the error variance in SpAM, HSIC LASSO and integrated model does not deviate much from the actual error variance, although the error variance in HSIC LASSO is much higher than the other two models.
From the above study we can conclude that the developed integrated model is more robust against the imputation of missing genotypes in comparison with the two models (i.e. SpAM and HSIC LASSO).

**ACKNOWLEDGEMENT**

The first author acknowledges the fellowship received from ICAR-IASRI for the PhD programme. The facilities provided by ICAR-IARI and ICAR-IASRI are duly acknowledged.

**REFERENCES**


