Journal of Crop and Weed, 16(1): 133-137 (2020)
ISSN- O : 2349 9400 ; P : 0974 6315

http://cwssbckv.org
www.cropandweed.com

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 there 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 earlier 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 (*Triticumaestivum*) (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*., 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

Email : sayanti23gm@gmail.com

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 asMENDEL (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 (Nicolae *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*. haveinvestigated 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. Weigel*et al*. (2010) also studied the effect of imputation in dairy cattles 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 (GuhaMajumdar *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 (Crossa *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

$$y_{Int} = w y_{sp} + (1 - w) y_{HL}$$

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

$$\sigma^2_{Int} = \left(\frac{\sigma^2_{HL}}{\sigma^2_{Sp} + \sigma^2_{HL}}\right)^2 \sigma^2_{Sp} + \left(\frac{\sigma^2_{Sp}}{\sigma^2_{Sp} + \sigma^2_{HL}}\right)^2 \sigma^2_{HL}$$
$$= \frac{\sigma^2_{Sp}\sigma^2_{HL}}{\sigma^2_{Sp} + \sigma^2_{HL}}$$

The estimation of 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_{pred}) (values (Howard *et al*., 2014).

Prediction accuracy (PA) = correlation (y_{actual}, y_{pred})

The RED score (Zhao *et al*., 2010) can be obtained by

$$RED = \frac{1}{m(m-1)} \sum_{u_k, u_j, k>l} |\rho_{k,l}|$$

where, $\rho_{k,l}$ is the correlation coefficient between the *k*-th and *l*-th markers, is the number of selected markers. A high RED score indicatesstrong correlation between selected markers, which is not desirable because it means many redundant markers are selected. In order to estimate theprediction 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 wereremoved 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 werefollowed 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 comparingestimated 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.

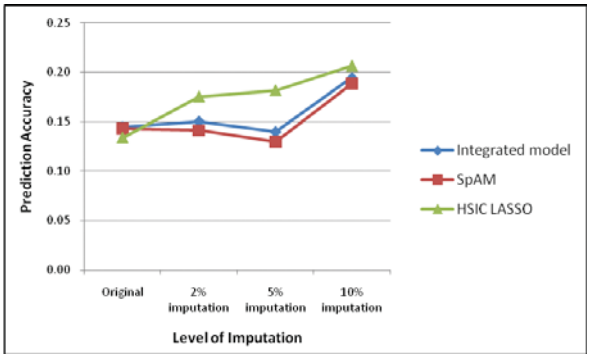

Fig. 1: Comparison of prediction accuracy (PA) for integrated model, SpAM and HSIC LASSO at different levels of imputation

LASSO)the prediction accuracy for imputed datasets deviates from the prediction accuracy of original data and with the increase in the imputation level the difference is increasing. This is due to obvious reasons that with the increase in the extent of imputation on the dataset, the original dataset become more distant from the imputed dataset. Another observation from Fig. 1 is that fluctuation of the value of prediction accuracy is highest in case of HSIC LASSO and it is minimum in case of integrated model. This clearly indicates that the integrated model is robust against the missing genotypes. Also, it is observed that prediction accuracy increases with the increase in the level of imputation. This is due to the method of imputation that has been applied to the dataset. The method of imputation assumes that the data is in Hardy-Weinberg equilibrium, but in case of real datasets it is not true always. So, the imputation method makes the dataset close to the Hardy-Weinberg equilibrium. That is why at the 10% level of imputation the prediction accuracy is highest.

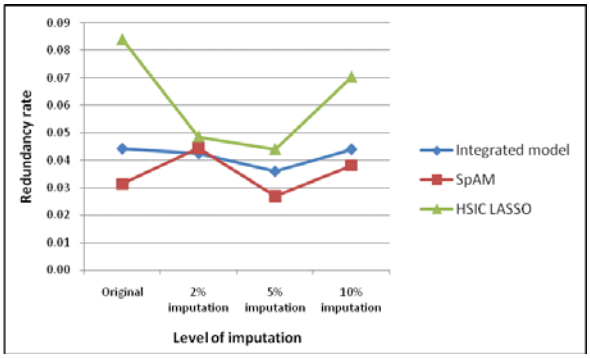

Fig. 2: Comparison of redundancy rate (RED) for integrated model, SpAM and HSIC LASSO at different levels of imputation

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.

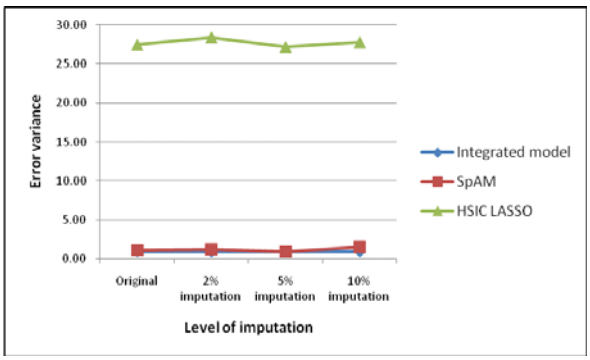

Fig. 3: Comparison of error variance for integrated model, SpAM and HSIC LASSO at different levels of imputation

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

Abecasis, G.R., Cherny, S.S., Cookson, W.O. and Cardon, L.R. 2002. Merlin — rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genet.*, **30(1)** : 97-101.

Browning, B.L., Zhou, Y. and Browning, S. R. 2018. A one-penny imputed genome from next generation reference panels. *American J. Human Genet.*, **103(3)**:338-48.doi:10.1016/j.ajhg.2018.07.015

Burdick, J.T., Chen, W.M., Abecasis, G. R., Cheung, V.G. 2006. In silico method for inferring genotypes in pedigrees. *Nature Genet.*,**38(9)**:1002-04.

Chen, L., Li, C., Sargolzaei, M. and Schenkel, F. 2014.Impact of Genotype Imputation on the Performance of GBLUP and Bayesian Methods for Genomic Prediction.*PLoS ONE*, **9(7)**: e101544. https://doi.org/10.1371/journal.pone.0101544

Crossa, J., de los Campos, G., Perez, P., Gianola, D. and Burgueño, J. 2010. Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers.*Genetics*,**186** :713-24.https://doi.org/10.1534/genetics.110.118521

Daetwyler, H.D., Bansal, U.K., Bariana, H.S., Hayden, M.J. and Hayes, B.J. 2014.Genomic prediction for rust resistance in diverse wheat landraces. *Theor. App. Genet.*,**127** :1795-1803.

Elston, R.C. and Stewart, J. 1971. A general model for the genetic analysis of pedigree data. *Human Heredity*, **21** : 523-42.

Gretton, A., Bousquet, O., Smola, A. and Scholkopf, B. 2005.Measuring statistical dependence with Hilbert-Schmidt norms. *Algorithmic Learning Theory (ALT)*. pp 63-77. Springer.

GuhaMajumdar, S., Rai, A. and Mishra, D.C. 2019.Integrated Framework for Selection of Additive and Nonadditive Genetic Markers for Genomic Selection. *J. Comput. Biol*. http://doi.org/10.1089/cmb.2019.0223

GuhaMajumdar, S., Rai, A. and Mishra, D.C. 2019.GSelection: Genomic Selection. R package version 0.1.0.https://CRAN.R-project.org/package=GSelection

Heath, S.C. 1997. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *American J. Human Genet.*,**61** :748-60.

Henderson, C.R. 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics*,**31(2)**: 423-47.

Hoerl, A.E. and Kennard, R.W. 1970. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*,**12**: 55-67.

Howard, R., Carriquiry, A. L. and Beavis, W. D. 2014.Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures.*G3 (Bethesda)*,**4(6):**1027-46.

Lander, E.S. and Green, P. 1987. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America*,**84(8)**: 2363-67.

Lange, K., Sinsheimer, J.S. and Sobel, E. 2005.Association testing with Mendel.*Genetic Epidemiology*, **29**: 36-50.

Li, Y., Willer, C.J., Ding, J., Scheet, P. and Abecasis, G. R. 2010.MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, **34(8)**: 816-34. doi: 10.1002/gepi.20533.

Lorenz, A.J., Smith, K.P. and Jannink, J. L. 2012. Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. *Crop Sci.*, **52(4)** : 1609-21.https://doi.org/10.2135/cropsci2011.09.0503

Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genet.*,**39** : 906-13.

Marchini, J. and Howie, B. 2010.Genotype imputation for genome-wide association studies.*Nature Rev. Genet.*,**11(7)**:499-511. doi: 10.1038/nrg2796

Meuwissen, T.H.E., Hayes, B.J. and Goddard, M.E. 2001.Prediction of total genetic value using genome-wide dense marker maps.*Genetics*,**157**: 1819-29.

Mulder, H.A., Calus, M.P.L., Druet, T. and Schrooten, C. 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in dutchholstein cattle. *J. Dairy Sci.*, **95** : 876-89.https://doi.org/10.3168/jds.2011-4490

Nicolae, D. L. 2006. Testing untyped alleles (TUNA)-applications to genome-wide association studies. *Genet. Epidemiology*, **30**: 718-27.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R. 2007. PLINK: a toolset for whole genome association and population-based linkage analyses. *American J. Human Genet.*, **81**: 559-75.

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. 2009. Sparse additive models. *J. Royal Stat..Soc.: Series B (Statistical Methodology)*,**71(5)**:1009-30.

Scheet, P. and Stephens, M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American J. Human Genet.*, **78** : 629-44.

Servin, B. and Stephens, M. 2007. Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. *PLoS Genet.*, **3(7)**:e114. https://doi.org/10.1371/journal.pgen.0030114

Tibshirani, R. 1996. Regression shrinkage and selection via the Lasso. *J. Royal Stat. Soc*,**58**: 267-88.

Weigel, K.A., de los Campos, G., Vazquez, A. I., Rosa, G.J.M. and Gianola, D. 2010. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in jersey cattle. *J. Dairy Sci.*,**9**B: 5423-35.https://doi.org/10.3168/jds.2010-3149

Xu, S. Z., Zhu, D. and Zhang, Q. F. 2014.Predicting hybrid performance in rice using genomic best linear unbiased prediction. *Proceedings of the National Academy of Sciences of the United States of America*,**111**:12456- 461.

Xu, Y. and Wu, J. 2014. linkim: Linkage information based genotype imputation method. R package version 0.1. https://CRAN.R-project.org/package=linkim

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P. and Sugiyama, M. 2014. High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso. *Neural Computation*, **26** : 185-207.

Zaitlen, N., Kang, H. M., Eskin, E. and Halperin, E. 2007.Leveraging the HapMap correlation structure in association studies. *American J. Human Genet.*,**80** : 683-91.

Zhao, Y, Gowda, M., Liu, W., Würschum, T., Maurer, H.P., Longin, F.H., Ranc, N. and Reif, J.C. 2012.Accuracy of genomic selection in European maize elite breeding populations. *Theor. App. Genet*, **124**: 769-76. doi: 10.1007/s00122-011-1745-y

Zhao, Z., Wang, L. and Li, H. 2010. Efficient spectral feature selection with minimum redundancy. *AAAI Conference on Artificial Intelligence (AAAI)*, pp 673-678.