# An Imputation Strategy which Results in an Alternative Parameterization of the Single Stage Genomic Evaluation

***J.M. Hickey[1], G. Gorjanc[2], B.P. Kinghorn[1], B. Tier[3], J.H.J. van der Werf[1,4] and M.A. Cleveland[5]***
*[1]School of Environment and Rural Science, University of New England, Armidale, Australia*
*[2]Univ. of Ljubljana, Biotechnical Fac., Dept. of Animal Science, Domžale, Slovenia*
*[3]Animal Breeding and Genetics Unit, University of New England, Armidale, Australia*
*[4]Cooperative Research Centre for Sheep Industry Innovation, Armidale, Australia*
*[5]Genus plc., 100 Bluegrass Commons Blvd., Suite 2200, Hendersonville, TN, 37075, USA*
John.Hickey@une.edu.au

## Abstract

A method is presented for imputing genotypes in pedigreed populations based on long-range phasing, haplotype libraries, recombination modelling and segregation analysis. In two very different data sets, one a pig data set comprising animals from a single line and the other a multiple breed cattle data set, imputation accuracy was high and was always higher than that of Impute2 a widely used alternative. Accuracy was highest for animals which had both parents genotyped at high-density, however some animals with neither parent genotyped at high-density also had high imputation accuracy. The method imputes genotypes or sum of the allele probabilities for all animals in the pedigree and thus facilitates single stage genomic evaluations combining all available pedigree, genomic, and phenotypic information in a single step. This was explored using both simulated and real data with favourable results.

## Introduction

Genomic information is now widely used in many breeding programs. To be successful it requires large numbers of animals that are both intensively genotyped as well as phenotyped. Densely genotyping the numbers of animals required is prohibitively expensive. Therefore much research effort has been focussed on developing methods and strategies for the imputation of genotypes. Imputation in livestock scenarios can access information on low-density genotypes, pedigree, linkage, and linkage disequilibrium. The first objective of this paper was to briefly describe an imputation method that accesses pedigree, linkage, and low-density genotype information to perform imputation in livestock scenarios. The method, which is implemented in a new software package called AlphaImpute combines simple phasing rules, long-range phasing, haplotype libraries, segregation analysis, and recombination modelling to generate imputed genotypes or sum of the allele probabilities for all animals in a pedigree at all loci. The second objective was to illustrate how these imputed genotypes or sum of the allele probabilities can be used to implement a single stage genomic evaluation that combines all pedigree, phenotypic, and genotypic information available in a single step (SSAI). This may overcome the problem of the pedigree and genomic relationship matrices having different bases in the single stage genomic evaluation methodology of Misztal *et al.* (2009) (SSMi) and easily facilitates variable weighting on individual SNP using an approach such as BayesB.

## Material and Methods

AlphaImpute combines simple phasing rules, long-range phasing, haplotype libraries, segregation analysis, and recombination modelling to impute genotypes or sum of the allele probabilities for all loci of all animals in a pedigree. It proceeds by firstly separating out a set of high-density animals. These animals have their SNP phased using long-range phasing and haplotype library imputation (Hickey *et al.*, 2011). Allele probabilities (Kerr and Kinghorn, 1996) are calculated for all SNP of all animals in the pedigree and where these probabilities are >0.99 alleles are imputed. This can be thought of as single locus phasing. Next the haplotypes identified from the long-range phasing step are matched to alleles phased via the single locus phasing step. This matching step begins with parental and other

ancestral haplotypes by processing the data from the top of the pedigree downward. The second part of the matching step involves processing the haplotype libraries to see if the haplotypes that an animal carries exist in the library in animals that are not ancestors. This process is iterated a number of times before the sum of the allele probabilities are re-calculated. Finally recombination locations are identified and modelled assuming that SNP and recombination hotspots are evenly distributed across the genome.

The performance of AlphaImpute was tested in two data sets. The first a PIC pig data set, comprising a pedigree of 6,473 animals in which 3,709 animals were genotyped at high-density. The second was a multiple breed LIC cattle data set, comprising a pedigree of 24,017 animals of which 5,047 were genotyped at high-density. Both data sets were divided into training and testing sets. Testing sets had proportions of their high-density genotypes masked and then imputed. Low-density genotyping platforms roughly equivalent to 0.5k, 2.5k, 5k, and 7.5k per genome were created. In the pig data set the testing set comprised 509 animals from the most recent generation. For cattle, the testing set comprised 626 animals randomly selected from the genotyped animals with the restriction that the animals testing animals had to have their parents identified in the pedigree. The accuracy was calculated for 6 different groups of animals, those with both parents genotyped, sire and maternal grandsire genotyped, dam and paternal grandsire genotyped, sire genotyped, dam genotyped, and other animals. Imputation accuracy was measured as the correlation between true and imputed genotypes. Imputation performance was compared to that of Impute2 (B. Howie and J. Marchini, Oxford University).

AlphaImpute generates imputed genotypes or sum of the allele probabilities for all loci of all animals in a pedigree. These can be used in the SSAI to combine all pedigree, genotype, and phenotype information in a single step evaluation. The SSMi of Misztal *et al.* (2009) combines genomic, pedigree and phenotype matrix by modifying the elements in the pedigree derived relationship matrix based on genomic information. SSAI was explored in two steps. Firstly simulated data was used to show that using sum of the allele probabilities (Kerr and Kinghorn, 1996), without any imputation via AlphaImpute, for all ungenotyped animals in the pedigree gave identical results to SSMi. Secondly real data was used to show that the SSAI could increase the accuracy of genomic selection.

Briefly the simulated data were created by first simulating ancestral haplotypes using a coalescent process, these haplotypes were then dropped through a pedigree of ten generations where each generation comprised 1000 offspring from 500 dams and 50 randomly selected sires. Mating was at random. Genotype information was assumed to be available for the sires from generations 1, 2, and 3, for all animals in generations 4 and 5, and for 500 randomly selected individuals in each of generations 6, 8, and 10. Phenotype information was available for generations 2,3,4, and 5. The animals in generations 6, 8, and 10 were candidates for selection representing close, medium, and distant relatives. Four traits with different distributions of QTL effects were simulated. This data was analyzed using SSMi. It was also analyzed by first calculating sum of the allele probabilities for all animals in the pedigree, then using this information to construct a genomic relationship matrix for all animals and then estimate genomic breeding values using Gblup. To clarify, no imputation was done here other than the single locus peeling of Kerr and Kinghorn (1996).

In the real data analysis the pig data set was used. Two scenarios were compared. In scenario 1 (SC1) only 3200 high-density training animals were used to train the prediction equation. In scenario 2 (SC2) the 2764 animals without genotypes were used in addition to the high-density animals to train the prediction equation. To apply SSAI to SC2 AlphaImpute was used to generate imputed genotypes or sum of the allele probabilities for all animals in the pedigree. Genomic breeding values were then predicted using Gblup based on this information. Contrary to the simulated data, explicit imputation was carried out here. The accuracy of the GEBVs were validated in the 509 testing animals by correlating them to

progeny test estimated breeding values calculated using traditional BLUP.

## Results and Discussion

AlphaImpute imputed genotypes with very high accuracy. Across all of the categories of animals in both the pig and cattle data sets it had greater accuracy than IMPUTE2 (Tables 1 and 2) and performed the task in less time. The increase in accuracy of AlphaImpute over IMPUTE2 increased with reducing density of the low-density panels. For AlphaImpute the correlation between true and imputed genotypes/sum of the allele probabilities increased with increasing relatedness between the ancestors who were genotyped and the animal to be imputed and as the density of the low-density genotyping increased.

In the analysis of the simulated data using sum of the allele probabilities to include completely ungenotyped relatives in a single stage genomic evaluation gave results identical to SSMi for all traits in each of the groups of relatives (close, medium, and distant) suggesting that these two models have similar properties. Two problems may exist with SSMi, firstly the genomic relationships and the pedigree relationships may have a different base generation and secondly SSMi does not automatically place differential emphasis on important SNP in a way that a variable selection method such as BayesB would do. SSAI overcomes the base generation problem by using the genotype information to estimate the base generation. Because all animals have genotype information, variable selection methods such as BayesB can be used in SSAI to differentially weight different SNP according to their effect on phenotype for each trait.

Applying SSAI to the real pig data increased the accuracy of genomic selection. The accuracy of the GEBV for SC1 was 0.41 across all 509 of the testing animals and 0.51 for the 32 testing animals that had an accuracy of their BLUP EBV of greater than 0.90. In SC2 the accuracy increased to 0.49 for all 509 testing animals and 0.62 for the group of animals with highly accurate BLUP EBV.

Imputing through phasing with use of information from relatives, as carried out here, paves the way for inference of IBD status between contributing gametes across the dataset, at the locus level and above. This will result in genomic relationship matrices based on IBD rather than IBS, which remove scaling issue and likely increases in accuracy of GEBVs.

## Conclusions

The method to impute genotypes implemented in AlphaImpute is robust and accurate. The output of the program results in an alternative parameterization of the single stage genomic evaluation that uses all pedigree, genotype, and phenotype information available.

## Availability

AlphaImpute is available at: http://sites.google.com/site/hickeyjohn/

## Acknowledgements

We thank LIC and PIC for providing genotypes and pedigrees.

## References

Hickey, J.M., Kinghorn, B.P., Tier, B., Wilson, J.F., Dunstan, N. & van der Werf, J.H.J. 2011b. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genet. Sel. Evol. 43*,12.

Kerr, R.J. & Kinghorn, B.P. 1996. An efficient algorithm for segregation analysis in large populations. *J. Anim. Breed. Genet. 113,* 457-469.

Misztal, I., Legarra, A. & Aguilar, I. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J. Dairy Sci. 92*, 4648-4655.

**Table 1.** Accuracy of imputation for AlphaImpute (AI) and Impute2 (I2) for the pig data sets for the different low density scenarios.

| | 0.5k | | 2.5k | | 5.0k | | 7.5k | |
|---|---|---|---|---|---|---|---|---|
| Category | AI | I2 | AI | I2 | AI | I2 | AI | I2 |
| BothParents | 0.98 | 0.77 | 0.99 | 0.92 | 1.00 | 0.96 | 1.00 | 0.96 |
| SireMGS | 0.93 | 0.80 | 0.98 | 0.92 | 0.99 | 0.94 | 0.99 | 0.96 |
| DamPGS | 0.96 | 0.79 | 0.98 | 0.92 | 0.99 | 0.95 | 0.99 | 0.96 |
| Sire | 0.89 | 0.78 | 0.97 | 0.92 | 0.99 | 0.95 | 0.99 | 0.97 |
| Dam | 0.90 | 0.76 | 0.96 | 0.89 | 0.98 | 0.93 | 0.98 | 0.95 |
| Other | 0.86 | 0.79 | 0.94 | 0.91 | 0.97 | 0.95 | 0.97 | 0.96 |

**Table 2.** Accuracy of imputation for AlphaImpute (AI) and Impute2 (I2) for the cattle data set for the different low-density scenarios.

| | 0.5k | | 2.5k | | 5.0k | | 7.5k | |
|---|---|---|---|---|---|---|---|---|
| Category | AI | I2 | AI | I2 | AI | I2 | AI | I2 |
| BothParents | 0.97 | 0.64 | 0.99 | 0.92 | 0.99 | 0.94 | 1.00 | 0.95 |
| SireMGS | 0.87 | 0.60 | 0.97 | 0.91 | 0.98 | 0.95 | 0.99 | 0.96 |
| DamPGS | 0.92 | 0.63 | 0.97 | 0.87 | 0.98 | 0.91 | 0.98 | 0.95 |
| Sire | 0.86 | 0.60 | 0.96 | 0.90 | 0.98 | 0.95 | 0.98 | 0.96 |
| Dam | 0.95 | 0.63 | 0.98 | 0.90 | 0.99 | 0.97 | 0.99 | 0.95 |
| Other | 0.84 | 0.58 | 0.94 | 0.90 | 0.96 | 0.95 | 0.97 | 0.96 |