**Supplementary Material S3 – Derivation of factors affecting imputation accuracy**

Factors affecting imputation accuracy can be derived theoretically as follows. Below, we consider for convenience that imputation is done using the gene dosage method described by Gengler *et al*. ([2007](#_ENREF_20)), that does not consider observed genotypes of the imputed animals themselves. The imputation model is:

$$x\_{IMP\_{ij}}=μ\_{j}+u\_{ij}+ε\_{ij}$$

where $x\_{IMP\_{ij}}$ is the genotype of imputed animal *i* at locus *j*, $μ\_{j}$ is the average genotype at locus *j*, $u\_{ij}$ is the predicted genotype at locus *j* for animal *i*, and $ε\_{ij}$ is the residual associated with the imputed genotype.

Considering the above model, the distributions of true and imputed genotypes specific for a single locus *j* can be represented as:

 $\left[\begin{matrix}\begin{matrix}X\_{RP}\_{j}\\X\_{IMP\_{j}}\end{matrix}\\\hat{X\_{IMP}\_{j}}\\ε\_{j}\end{matrix}\right]=MVN\left[0,\left(\begin{matrix}A\_{RP}σ\_{g}^{2}\_{j}&A\_{RP,IMP}σ\_{g}^{2}\_{j}&A\_{RP}^{-1}A\_{RP,IMP}σ\_{g}^{2}\_{j}&0\\A\_{IMP,RP}σ\_{g}^{2}\_{j}&A\_{IMP}σ\_{g}^{2}\_{j}+Iσ\_{ε}^{2}\_{j}&A\_{RP}^{-1}A\_{RP,IMP}σ\_{g}^{2}\_{j}&Iσ\_{ε}^{2}\_{j}\\A\_{IMP,RP}A\_{RP}^{-1}σ\_{g}^{2}\_{j}&A\_{IMP,RP}A\_{RP}^{-1}σ\_{g}^{2}\_{j}&A\_{IMP}A\_{IMP,RP}A\_{RP}^{-1}σ\_{g}^{2}\_{j}&0\\0&Iσ\_{ε}^{2}\_{j}&0&Iσ\_{ε}^{2}\_{j}\end{matrix}\right)\right]$

where $X\_{RP}\_{j}$is a matrix of observed genotypes of animals in the reference population, $X\_{IMP}\_{j}$($\hat{X\_{IMP}\_{j}}$)is a matrix of true (imputed) genotypes of imputed animals, $ε\_{j}$ is a vector of residuals of the imputed animals, $A\_{RP}$ is a matrix with additive genetic relationships between all animals in the reference population used for imputation, $A\_{IMP}$ is a matrix with the additive genetic relationships between imputed animals, $A\_{IMP,RP}$ is a matrix with additive genetic relationships between the reference population and imputed animals, $σ\_{g}^{2}\_{j}$ is the variance of the true genotypes (i.e. $σ\_{g}^{2}\_{j}=2p\_{j}\left(1-p\_{j}\right) $under Hardy-Weinberg equilibrium; where $p\_{j}$ is the allele frequency at locus *j*), **I** is an identity matrix, and $σ\_{ε}^{2}\_{j}$ is the variance of the imputation errors.

The presented variances and co-variances from the above equation can be used to compute locus specific imputation accuracies. From this equation, it follows that 1) the variance of the true genotypes of the imputed animals depends on their level of inbreeding (i.e. $A\_{IMP}σ\_{g}^{2}\_{j}+Iσ\_{ε}^{2}\_{j}$), 2) the co-variance of the true and imputed genotypes of the imputed animals depends on the relationship of the imputed animals with the reference animals (i.e. $A\_{IMP,RP}A\_{RP}^{-1}σ\_{g}^{2}\_{j}$), and 3) the variance of the imputed genotypes of the imputed animals depends both on the relationship of the imputed animals with the reference animals and the level of inbreeding of the imputed animals (i.e. $A\_{IMP}A\_{IMP,RP}A\_{RP}^{-1}σ\_{g}^{2}\_{j}$). Assuming that inbreeding levels of imputed animals are similar leads to the assumption that the true genotypes of the imputed animals have (approximately) the same distribution. If imputed animals have varying levels of relationship with the reference animals, however, then imputed genotypes across individuals come from a mixture of distributions, and this does violate the assumption of the Pearson correlation coefficient leading to biased estimates of the imputation accuracy. It also directly shows how this can be solved. Rather than computing the imputation accuracy across all imputed animals, it should be computed within groups of animals that have similar relationships to the reference animals.