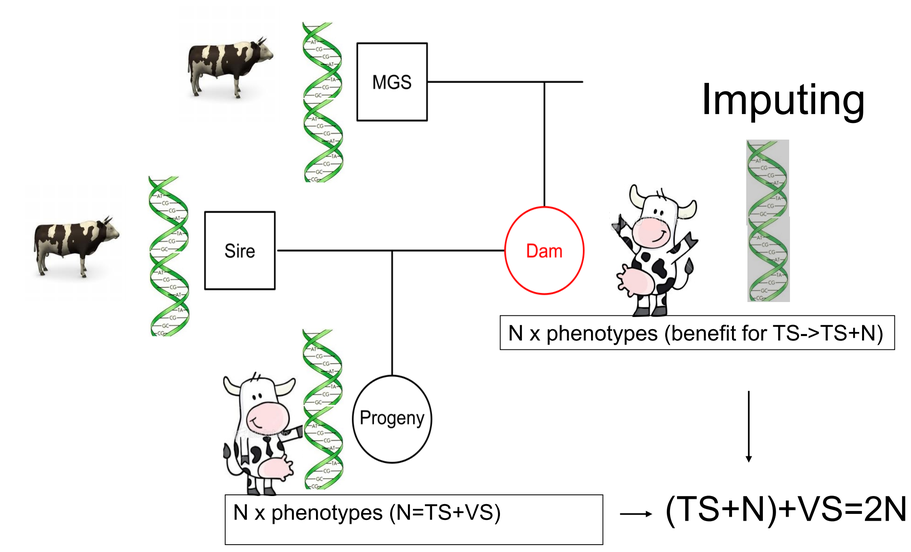
ENLARGING A TRAINING SET FOR GENOMIC SELECTION BY IMPUTATION OF UN-GENOTYPED ANIMALS IN POPULATIONS OF VARYING GENETIC ARCHITECTURE

64th Annual Meeting of the EAAP in Nantes – France, August 26th – 30th, 2013, Session 16.27

M. Wensch-Dorendorf[1], E. Pimentel[2], S. König[2], H. H. Swalve[1]

[1]Institute of Agricultural and Nutritional Sciences, University of Halle, 06099 Halle, Germany
[2]Department of Animal Breeding, University of Kassel, 37213 Witzenhausen, Germany

Introduction



Assumed family members with available genotypic information (black) (MGS=M, Sire=S, Progeny=P) used for imputing an un-genotyped dam (Dam=D, red).

The most common application of imputation is to infer genotypes (GT) of a high-density panel of markers on animals that are genotyped for a low-density panel.

Another application of imputation is to increase the size of the training set with un-genotyped animals. This strategy can be particularly successful when a set of closely related individuals are genotyped.

The posterior probabilities of the dam's three possible GT can be calculated following Bayes' theorem and the allele frequency (use the given GT for MGS ($M=G_j$) and Sire ($S=G_k$) -> $P(D=G_i|M=G_j \cap S=G_k \cap P=G_l)=P(D=G_i|P=G_l)$, $P(P=G_l|S=G_k \cap D=G_i)=P(P=G_l|D=G_i)$, $P(D=G_i|M=G_j)=P(D=G_i)$)

$$P(D=G_i \mid M=G_j \cap S=G_k \cap P=G_l)=\frac{P(P=G_l \mid S=G_k \cap D=G_i)P(D=G_i \mid M=G_j)}{\sum_{m=1}^{3} P(P=G_l \mid S=G_k \cap D=G_m)P(D=G_m \mid M=G_j)} \quad = \quad P(D=G_i \mid P=G_l)=\frac{P(P=G_l \mid D=G_i)P(D=G_i)}{\sum_{m=1}^{3} P(P=G_l \mid D=G_m)P(D=G_m)}$$

Data

QMSim (Sargolzaei et al., 2009) simulated data were used with the following characteristics to analyze **imputing methods**:

- One chromosome of 100cM, 2000 bi-allelic randomly allocated markers, $h^2=0.2$
- Two levels of LD x 2 levels of selection (LowLD_Sel, HighLD_Sel, LowLD_NoSel, HighLD_NoSel)
- 20 generations (for HighLD with bottleneck)
- 2000 genotyped female progeny from the last generation with known genotype for sire and mgs

The **impact** of enlarging a training set (TS) with imputed dams (TSA) **on the accuracy of genomic predictions** was evaluated for:

- different populations
- varying levels of heritability
- different sizes of genotyped females (TS)

Variants of differing heritability were generated by adding a residual term to the simulated true breeding values (h^2 = 0.05, 0.10,…,0.50).

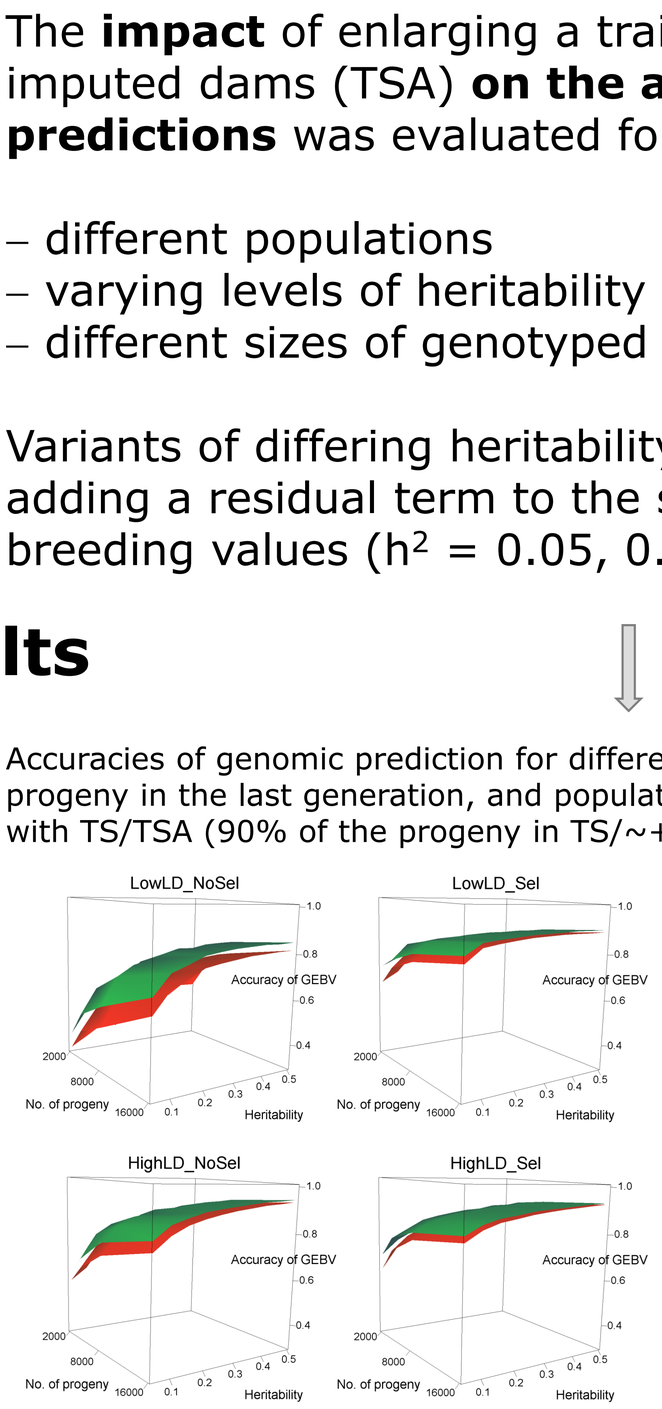
Results

Correlation between true and imputed genotypes from different imputation methods and programs

Imputing Method	Scenario			
	LowLD_NoSel	LowLD_Sel	HighLD_NoSel	HighLD_Sel
Single_Step[ig]	0.76 ± 0.003	0.83 ± 0.038	0.88 ± 0.004	0.90 ± 0.013
Single_Step[gc]	0.81 ± 0.003	0.86 ± 0.028	0.90 ± 0.003	0.93 ± 0.009
Two_Steps	0.57 ± 0.008	0.74 ± 0.066	0.80 ± 0.006	0.85 ± 0.021
findhap.f90	0.52 ± 0.006	0.69 ± 0.065	0.74 ± 0.006	0.82 ± 0.030
AlphaImpute	0.83 ± 0.003	0.87 ± 0.024	0.86 ± 0.004	0.89 ± 0.010

Single_Step=Bayes based methods ([ig] with integer genotypes, [gc] with „gene content"=non integer genotypes);Two_Steps: unambiguously inferred genotypes were used to build low-density panels and then imputed with fastPHASE; findhap.f90, AlphaImpute: Imputing-Software

Accuracies of genomic prediction for different values of h^2, number of female progeny in the last generation, and population structure, red/green surfaces: with TS/TSA (90% of the progeny in TS/~+imputed dams[Single_Step[gc]])



(allele substitution effects of every locus on the simulated phenotypes were fitted in a multiple random regression model similar to the GBLUP method of Meuwissen et al. (2001), accuracies were evaluated as correlation between simulated TBV and estimated GEBV)

Conclusion

With the underlying family structure (typical for Holsteins) imputation can be used to achieve an extra increase in accuracy of genomic predictions by enlarging the training set with completely un-genotyped dams. This strategy was shown to be particularly useful for populations with lower levels of linkage disequilibrium, for genomic selection on traits with low heritability, and for species or breeds for which the size of the reference population is limited.