IMPUTING MISSING GENOTYPES USING NUMERATOR RELATIONSHIP MATRIX

Funda Ogut,[1] Fikret Isik, Ross Whetten, and Steve McKeand

[1]Department of Forestry and Environmental Resources, North Carolina State University
Raleigh, NC

Genotyping does not work for all samples for all markers, so genetic data from a lab might have many missing genotypes. Yet, predictions of genetic merit of trees across markers require complete genotyping information or gene content. Excluding individuals with missing genotypes is not desired, since it will reduce the number of individuals in the population considerably, thus reducing the power of association of markers and traits. It is therefore important to use efficient statistical methods to accurately impute missing genotypes. Human geneticists rely on genetic maps and linkage disequilibrium (LD) information from nearby markers to replace missing genotypes. Their algorithms rely on known map positions for the SNPs. Since completely sequenced reference genomes are available for only two forest tree species, methods developed by human geneticists do not work well for most forest trees.

Gengler et al. (2007) described a method to impute missing genotypes using mixed linear models and BLUP. We determined the effect on accuracy of BLUP estimated breeding values of imputation with different levels (10%, 20%, 40%, 60% and 80%) of missing genotypes. Analyses were conducted both with empirical data (3461 SNP markers in a cloned loblolly pine population of 178 genotypes) and simulated data using missing data created by random sampling (some loci missing in all individuals) or by structured sampling (all loci missing in some individuals). Simulations were used to examine the effect of family and progeny size, mating design, proportion of missing genotypes, genotyping strategy and the method for imputation on the accuracy of breeding values. Imputed genotypes were obtained using the numerator relationship matrix (the A matrix) and solving the mixed model equations of $y = Xb + Mu + e$, where **y** is the vector of gene content predictions, X is the design matrix (vector of 1s) for the mean, M is the design matrix connecting trees to the gene content vector **y**, u is the individual tree effect and e is the error variance. The solutions of mixed model equations produce predicted SNP genotypes for trees with missing genotypes. The solutions are continuous, centered on 1 because the gene content values are 0, 1 or 2.

Imputation of missing genotypes in empirical data from an unbalanced mating design with family sizes ranging from 1 to 35 was more powerful for data with structured missing genotypes at all levels of missing data than for data with random missing genotypes with same proportions of missing data. The accuracy of imputation for 10% and 80% missing genotypes ranged between 0.96 to 0.23 and 0.96 to 0.16 for structured and random missing genotypes in the data, respectively. As the proportion of missing genotypes increased in the data, the power of imputation decreased. With simulation, we found that the imputation was less affected by the distribution of missing genotypes in a balanced mating design with families of equal size. The accuracy of imputation ranged between 0.97 to 0.75 for the 10% and 80% missing genotypes in the data, respectively.