## Cost-effective genotyping in plant and animal breeding using computational analysis of pooled samples

**Main supervisor:**
**Carl Nettelblad carl.nettelblad@it.uu.se, Associate Senior Lecturer**
**Division of Scientific Computing, Department of Information Technology**
**Science for Life Laboratory**
**Uppsala University**

**Genomic techniques allow fast advances in breeding for production traits, and in animal health. Still, full-scale genomic testing of all individuals is not cost-effective. We will develop a cost-effective way to retrieve exact genotypes at less than a third of current costs.**

Imagine you want to detect the single carrier of a rare mutation in a larger group. You have some 30 odd samples to test. Pooling samples based on binary numbering, or by a row/column structure, could bring it down to as little as 5 tests, rather than the 30 tests when testing each individual separately.

The goal of this project is to use pooling methods for systematic genotyping of samples. A pooled experiment means testing several individuals at the same time, using a single assay, in order to reduce sosts. A high accuracy reconstruction of individual genotypes will be possible by combining partially uncertain information from pooled experiments with known pedigrees. Similar schemes will also be applied to uncertain data from low-coverage sequencing sources.

Within the project we will extend current computational methods for filling in missing genotypes (imputation) to explicitly model the uncertain and incomplete genotypes found in the primary data recorded. These methods are based on Markov models, sometimes realized in sampling Monte Carlo schemes, and a crucial part of the project is to ensure reasonable runtime and convergence properties for these implementations. The main insight underlying all genotype imputation methods is that genetic variations located close to each other tend to be inherited together, unchanged, so missing information can be filled in from close or distant relatives.

Specifically, we will use pooling of samples in nucleotide polymorphism (SNP) arrays. Individual genotypes are partially reconstructed using multiple, overlapping pools. There is a decrease in the total number of assays needed, compared to testing individuals separately. Thus, more individuals can be genotyped at identical cost. When pooling, the identity of carriers of rare variants can be deduced directly, through the detection pattern in the overlapping pools (see figure). *Common* variants, on the other hand, can easily be imputed, based on the presence of rare variants nearby.

By reducing the genotyping cost per individual, larger cohorts are possible, and estimates of breeding values can be made more precise. To date, genomic breeding value estimates have been the most important contribution of genomic techniques to the breeding of livestock. This project is financed by Formas (www.formas.se) due to its potential to have relevant societal impact for Swedish agriculture. Thus, it provides exciting opportunities to tie together mathematical insight, scientific code development, and contacts with industry.

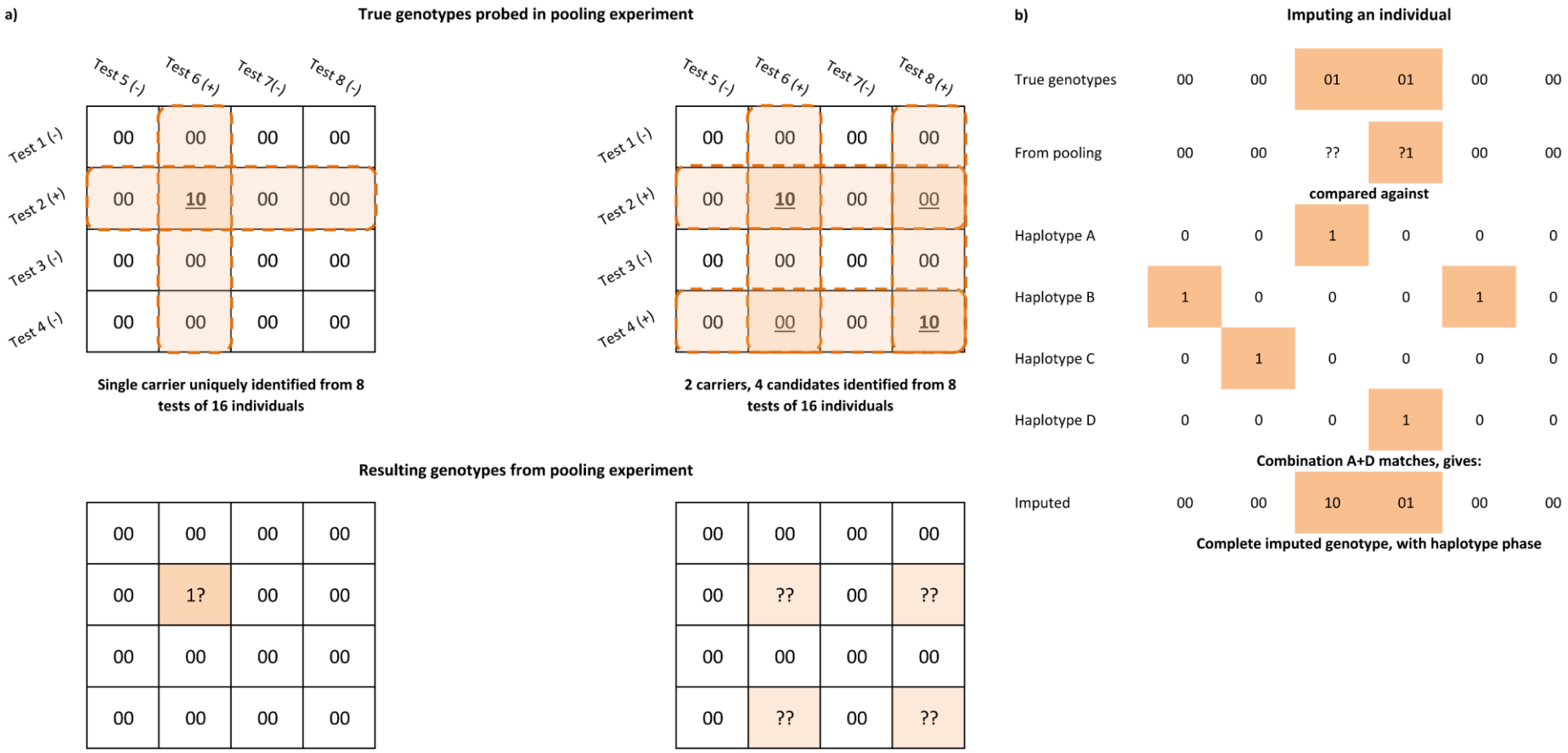


**Figure 1: Imputing genotypes through pooling and linkage. a)** One way of finding a carrier of a comparatively rare genetic variant among 16 samples would be to order them in a 4x4 grid and then test each column and each row as a single pool. This gives 8 pools for 16 samples. To the left, this results in a uniquely identified carrier, although it is unknown whether the genotype is homozygous (11) or heterozygous (?1). To the right, two carriers are present, and in this case the pooling finds 4 possible carriers, while not clearly identifying whether 2, 3, or 4 of them are the actual carriers. 12 individuals are conclusively identified as homozygous 00. **b)** Based on such partial information of genotypes over several linked variant positions in a chromosome, imputation methods can be used to find which reference sequences actually match the observed genotype pattern. In the example above, the partially observed individual genotypes are only consistent with the chromosome pair matching reference haplotypes A and D. The actual statistical models used in the project are mor e complex, since recombination can happen. In the project, I will improve the computational methods for imputation to better handle the patterns of very high levels of uncertainty thatcan result from pooling experiment, and still identify the most likely resolution based on reference haplotypes and the genotypes of relatives.