Predictive Accuracy: a Function of Genetic Distance

Marco Scutari[1], Anne-Marie Bochard[2], Ian Mackay[3], David Balding[1]

[1] Genetics Institute, University College London (UCL), [2] Limagrain Europe, [3] National Institute of Agricultural Botany (NIAB)

How Far Can We Predict Reliably?

How well conclusions generalise to different populations than that used in a statistical analysis is a key question in any genome-wide association study (GWAS) or genomic selection (GS) study. In addition to the influence of exogenous factors, an important consideration is how far a target population is from the training population the genome-wide model is estimated from. Furthermore, the former may not be available along with the latter, either because the data have yet to be collected (*e.g.* multi-stage studies) or because the individuals do not exist yet (*e.g.* future generations in a breeding program).
Naturally, we expect the predictive ability of the model to decay as the two populations are increasingly unrelated — that is, when their genetic distance increases. An interesting question then is, how far a target population can we reliably predict for a certain phenotype given a training population and an associated genome-wide model?

Clustering, Genetic Distance and Relatedness

The genetic distance between the two populations is known as kinship or relatedness, and can be estimated from marker profiles in several ways including allele sharing [4] and allelic correlation [1]. Allelic correlation in particular is interesting because it measures relatedness as a linear relationship, which can be easily handled by classic statistical procedures.
For example, if the target population is unknown we can:

- compute the kinship matrix of all samples from their marker profiles using allelic correlation;
- use the kinship matrix as a distance matrix and split the population in two subsets using nearest neighbour (k-nn) with $k = 2$;
- take the largest subset as the new training population to estimate the genome-wide model and the smallest as the new target population to estimate predictive power reliably [5].

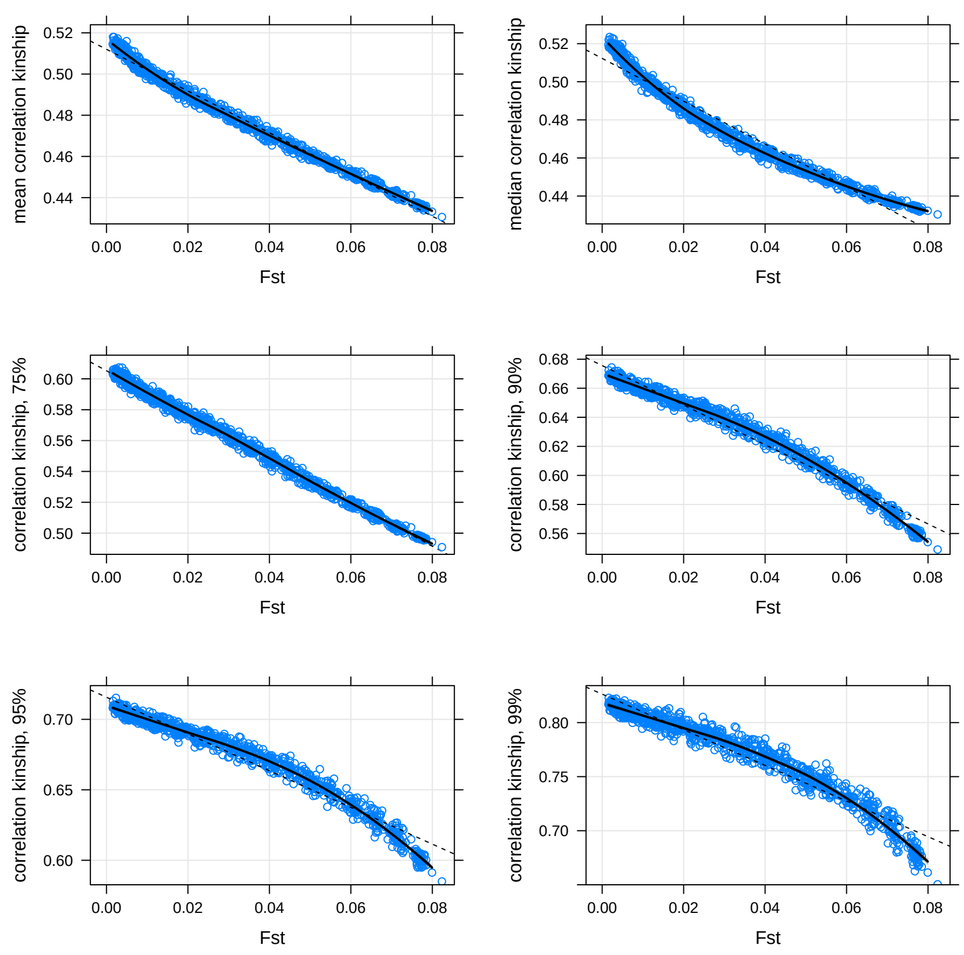
The two subsamples are guaranteed to minimise the average kinship coefficient between their elements, because k-nn maximises the average Euclidean distance and thus minimises the average allelic correlation. This is in contrast with cross-validation, which produces pairs of homogeneous training and test sets.

Spring Barley Data from Limagrain

The genotypic data for the MIDRIB (**M**olecular **I**mprovement of **D**isease **R**esistance in **B**arley") project comprise 2767 genotyped varieties with 6138 markers each, spread across 7 chromosomes. Phenotypic data on yield have been collected from 2006 to 2013 in France, Germany and the UK. We prototyped the proposed predictive accuracy estimation using UK varieties, and later validated it on the French and German varieties.

Which Measures of Relatedness When Clustering?

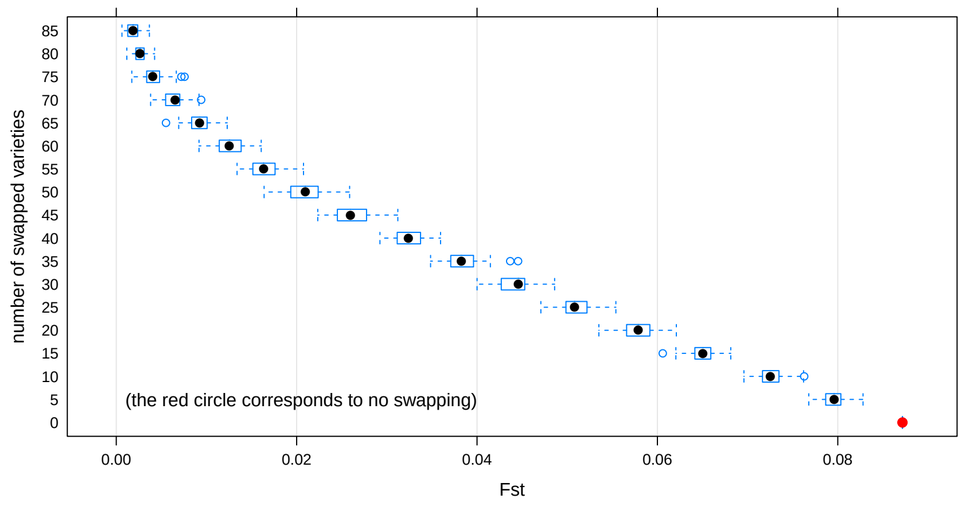
Can we use measures other than the average allelic correlation to estimate relatedness in the context of k-nn clustering? From simulation studies we can see that both the mean and some high quantiles of kinship are almost perfectly linearly related to F_{ST} [2, 3], so we can use any of them with comparable results and with the same interpretation.



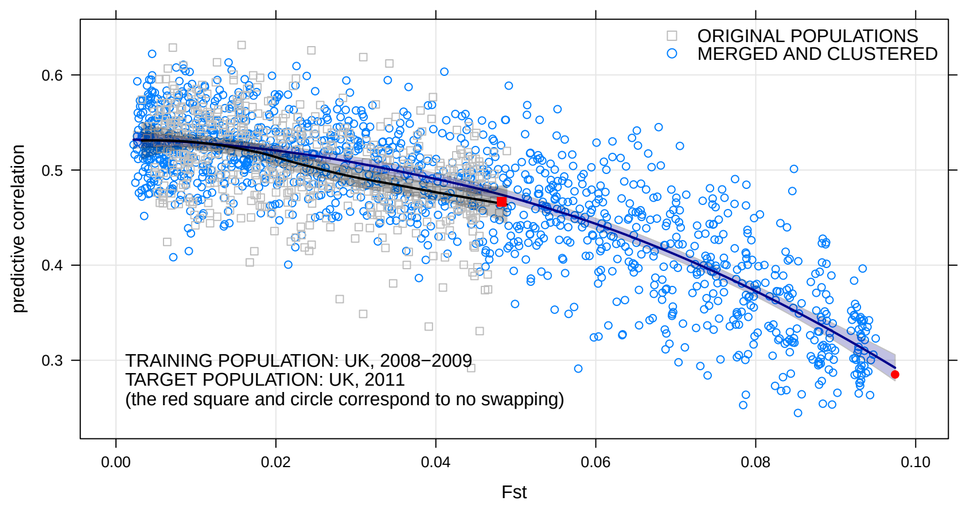
In particular mean allelic correlation and the 75% quantile are linearly related with F_{ST}. The same is true for other kinship estimators such as allele sharing.

Swapping Observations between Populations

From a given pair of training and test populations, we can generate new pairs with varying levels of relatedness by swapping samples between them. We can make them closely related by swapping a large proportion of their samples; or we can use k-nn clustering to make them as distantly related as possible. In practice iteratively increasing the proportion by a small number of samples ($m = 5$ to 10 in the figure below) makes it possible to cover the range of observable genetic distances (measured with F_{ST} below).

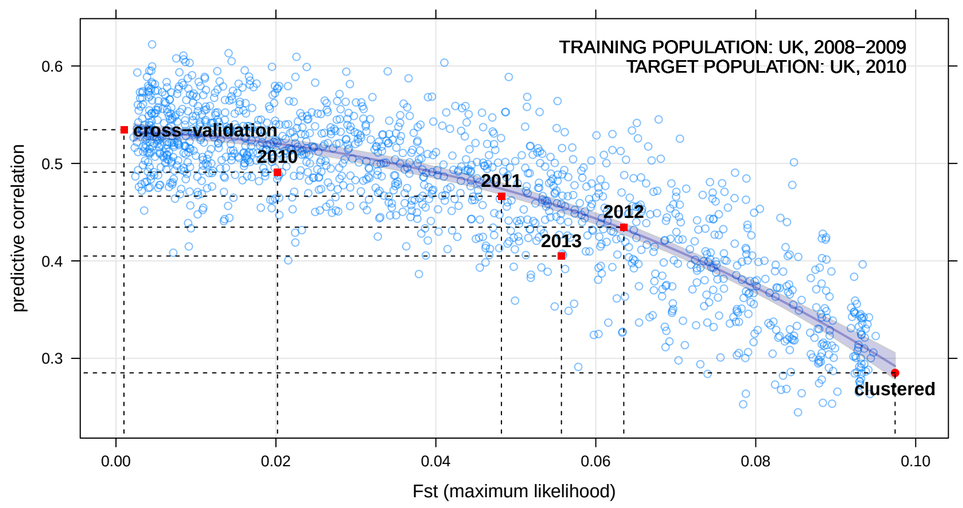


Even if the target population is available we can merge it with the training population and apply k-nn clustering to explore a wider range of genetic distances. If we measure the predictive power of a genome-wide model for every pair of samples we can then draw a curve describing the decay of predictive power as a function of relatedness.



Planning a Breeding Program

From a initial training population collected in the UK during 2008-2009, and using 2010 as an intial traget, we have used k-nn clustering to successfully estimate the predictive power of the genomewide prediction model for subsequent rounds of selection in years 2011, 2012 and 2013.



Given a lower threshold of predictive accuracy, this provides a measure of how many generations ahead we can predict yield reliably and can potentially reduce the number of varieties to be planted in trial fields and phenotyped.

References

[1] W. Astle and D. J. Balding. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471, 2009.

[2] M. A. Beaumont and D. J. Balding. Identifying Adaptive Genetic Divergence Among Populations from Genome Scans. *Molecular Ecology*, 13(4):969–980, 2004.

[3] G. Bhatia, N. Patterson, S. Sankararaman, and A. L. Price. Estimating and interpreting F_{ST}: The impact of rare variants. *Genome Research*, 23(9):1514–1521, 2013.

[4] D. Habier, R. L. Fernando, and J. C. M. Dekkers. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177:2389–2397, 2007.

[5] C. Riedelsheimer and A. E. Melchinger. Optimizing the Allocation of Resources for Genomic Selection in one Breeding Cycle. *Theoretical and Applied Genetics*, 126(11):2835–2848.

Email: m.scutari@ucl.ac.uk