Using Genetic Distance to Infer the Accuracy of Genomic Prediction

(for Quantitative Traits)

Marco Scutari

scutari@stats.ox.ac.uk
Department of Statistics
University of Oxford

September 7, 2015

The Problem

The extent to which predictive models generalise from the populations used to train them to distantly related target populations is an open question.

- The accuracy of such models is typically evaluated in the context of the training population using cross-validation, implicitly assuming that any new individual will have a similar general genetic layout [5, 7, 9].
- Strong focus on models' ability to correctly estimate heritability, but it is not clear how increases in explained genetic variance in the training sample translate to the prediction of unobserved phenotypes; while heritability provides an upper bound to predictive accuracy, it is rarely attained [9].
- Causal variants with both large and small effects are often different between different ethnic groups (in humans) or subspecies/families (in plants and animals). This can dramatically reduce the performance of a genomic prediction model because of the mismatch between the effect sizes or the allele frequencies in the training and the target population, even when population structure is taken into account [6, 7].

Background

Here we concentrate on how to extrapolate a decay curve for predictive accuracy as a function of a measure of genetic distance.

- We assume the training population is available and that the target population for prediction is not.
- We concentrate on quantitative traits, and use predictive correlation as a measure of predictive accuracy.
- We consider a maximum likelihood estimate of F_{ST} [2] to measure the genetic distance between the training and target samples. Average allelic correlation kinship [1] works just as well for this purpose.
- We also implicitly assume that the training population has enough genetic variability for the extrapolation to work, and that relevant causal variants have reasonably high MAF.

Extrapolating the Decay Curve

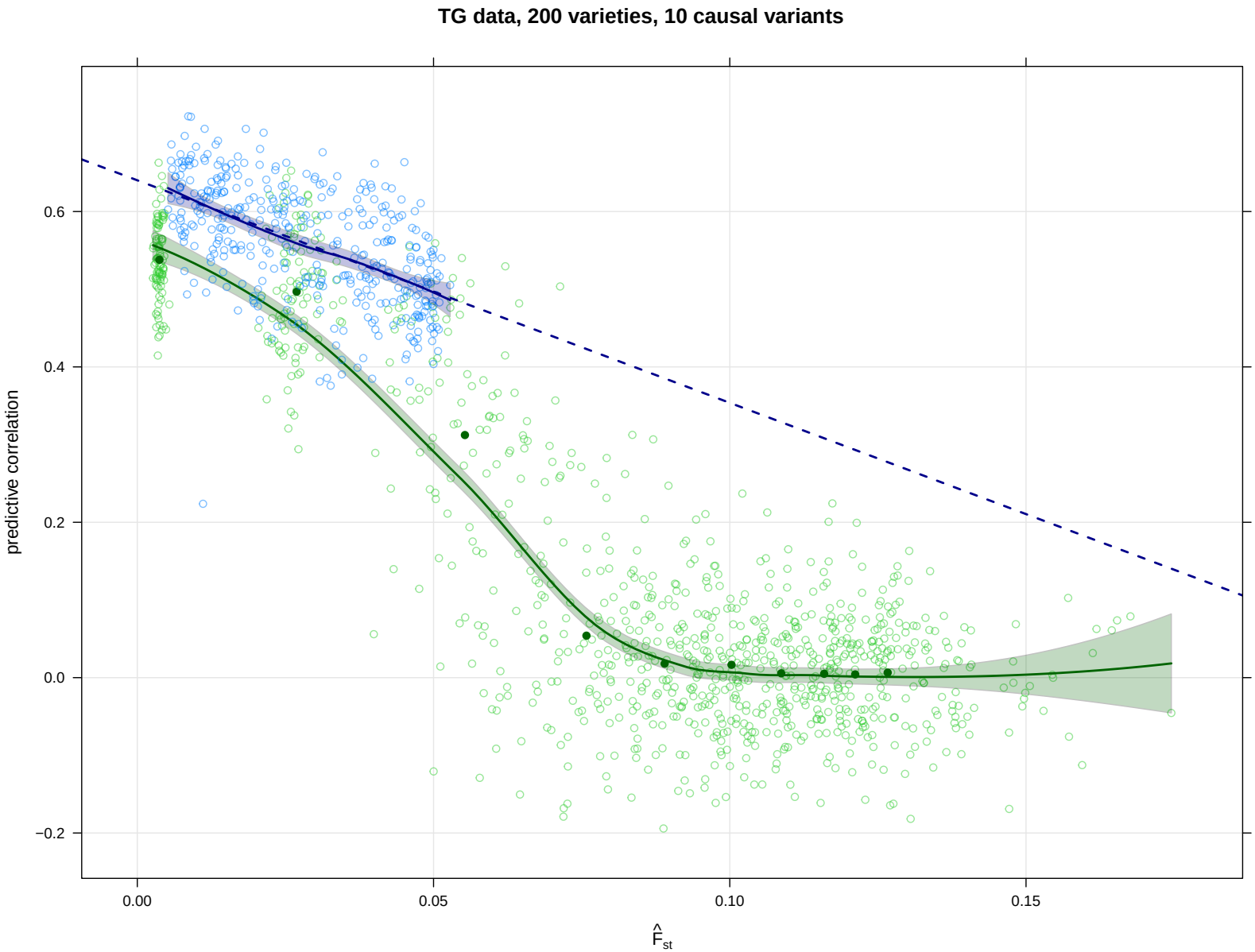
1. Produce a pair of minimally related subsets (i.e., with maximum F_{ST}) from the training population using *k*-means, $k = 2$. The largest of these two subsets will be used to train the genomic prediction model, and will be considered the ancestral population for the purposes of computing F_{ST}; the smallest will be the target used for prediction.
2. Compute $(\hat{F}_{\mathrm{ST}}^{(0)}, \hat{\rho}_{\mathrm{D}}^{(0)})$ for the pair subsets, which will act as the far end of the decay curve (in terms of genetic distance), using the elastic net.
3. For increasing values of m:
   - 3.1 create a new pair of subsamples by swapping m varieties at random between the training and the test subsamples from step 1;
   - 3.2 fit a genomic prediction model on the new training subsample and use it to predict the new target subsample, thus obtaining $(\hat{F}_{\mathrm{ST}}^{(m)}, \hat{\rho}_{\mathrm{D}}^{(m)})$.
4. Estimate the decay curve from the set of $(\hat{F}_{\mathrm{ST}}^{(m)}, \hat{\rho}_{\mathrm{D}}^{(m)})$ points using LOESS [4] or a simple linear regression.

The Data

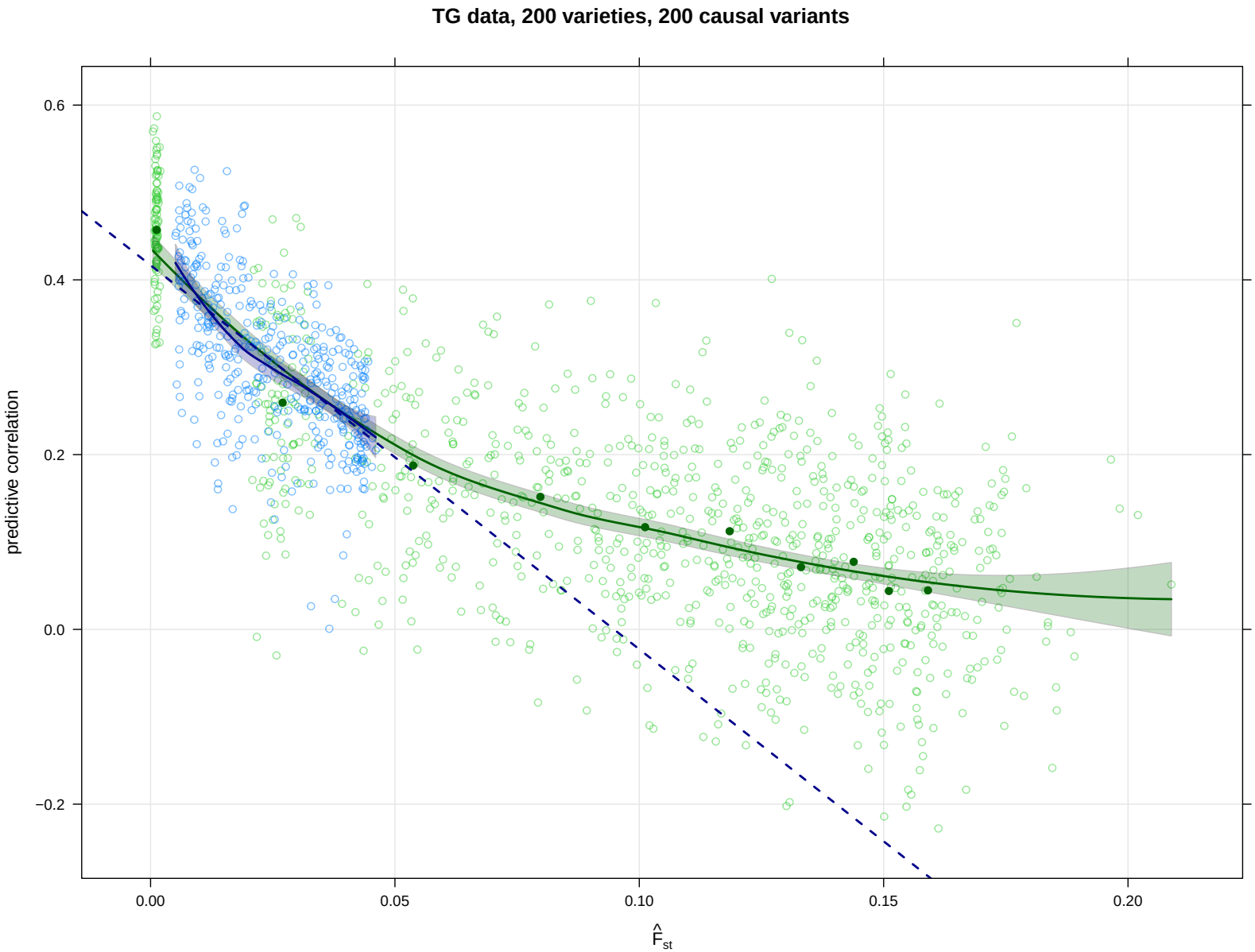
We consider 3 data sets both with their original phenotypes and with synthetic phenotypes (in the simulation studies).

- The TriticeaeGenome (TG) data [3], 376 registered wheat varieties from France (210), Germany (90) and the UK (75), genotyped using 2.7k DArT markers and known genes assays. Among the recorded traits we consider grain yield, height, flowering time, and grain protein content.
- The heterogeneous mouse population [11], 1940 mice genotyped with 12k SNPs; among the recorded traits, we consider growth rate and weight. The data include a number of inbred families, the largest being F005 $(287$ mice), F008 (293), F010 (332) and F016 (309).
- The Human Genetic Diversity Panel (HGDP) [8], 1043 individuals from Africa (151), America (108), Asia (435), Europe (167), the Middle East (146) and Oceania (36) genotyped with 650k SNPs. No phenotypes are available, so we only use chromosomes 1 and 2 $(90$k SNPs) for simulations.

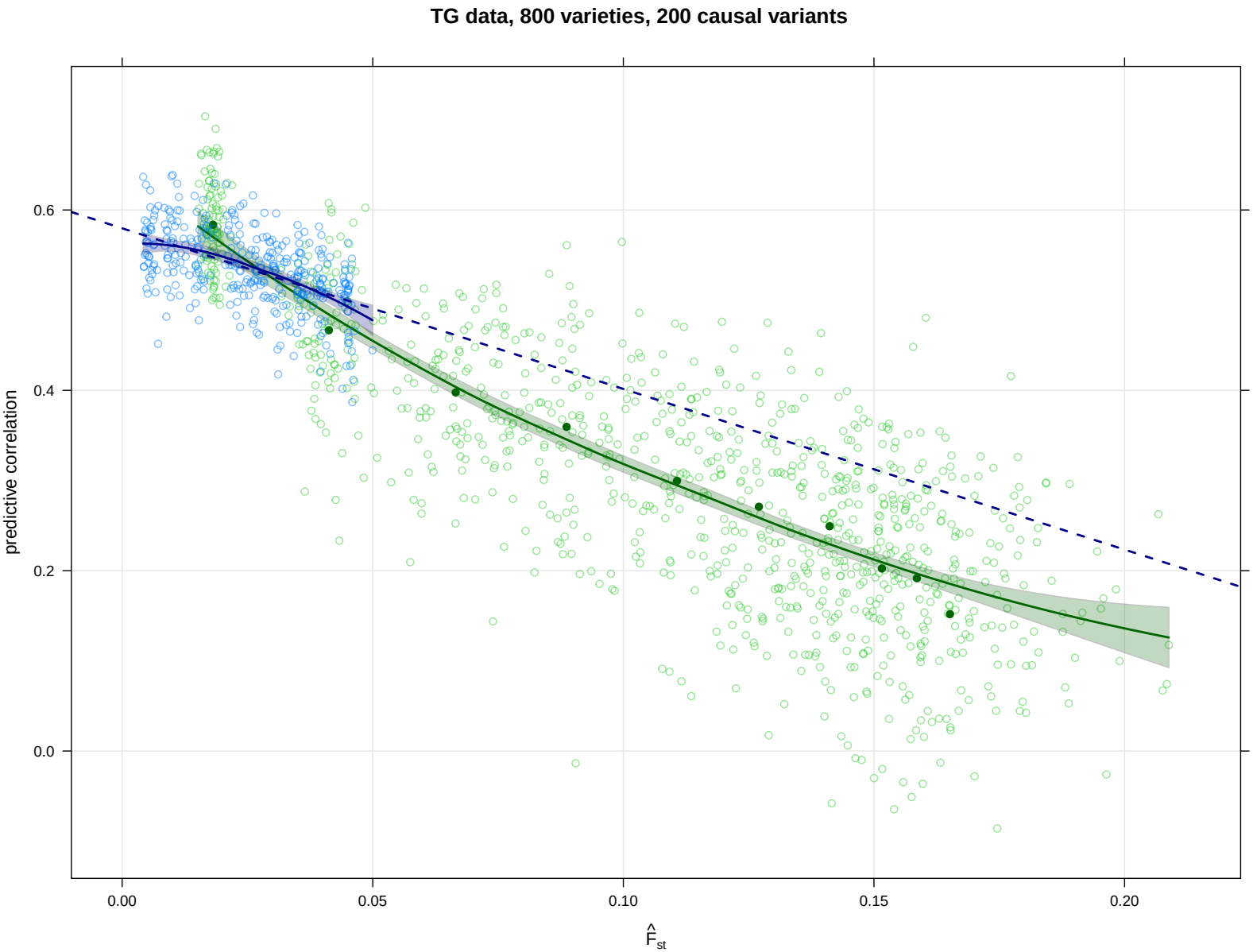
Simulation: Genomic Selection (Few Causal Variants)



Simulation: Genomic Selection (More Causal Variants)



Simulation: Genomic Selection (More Training Samples)



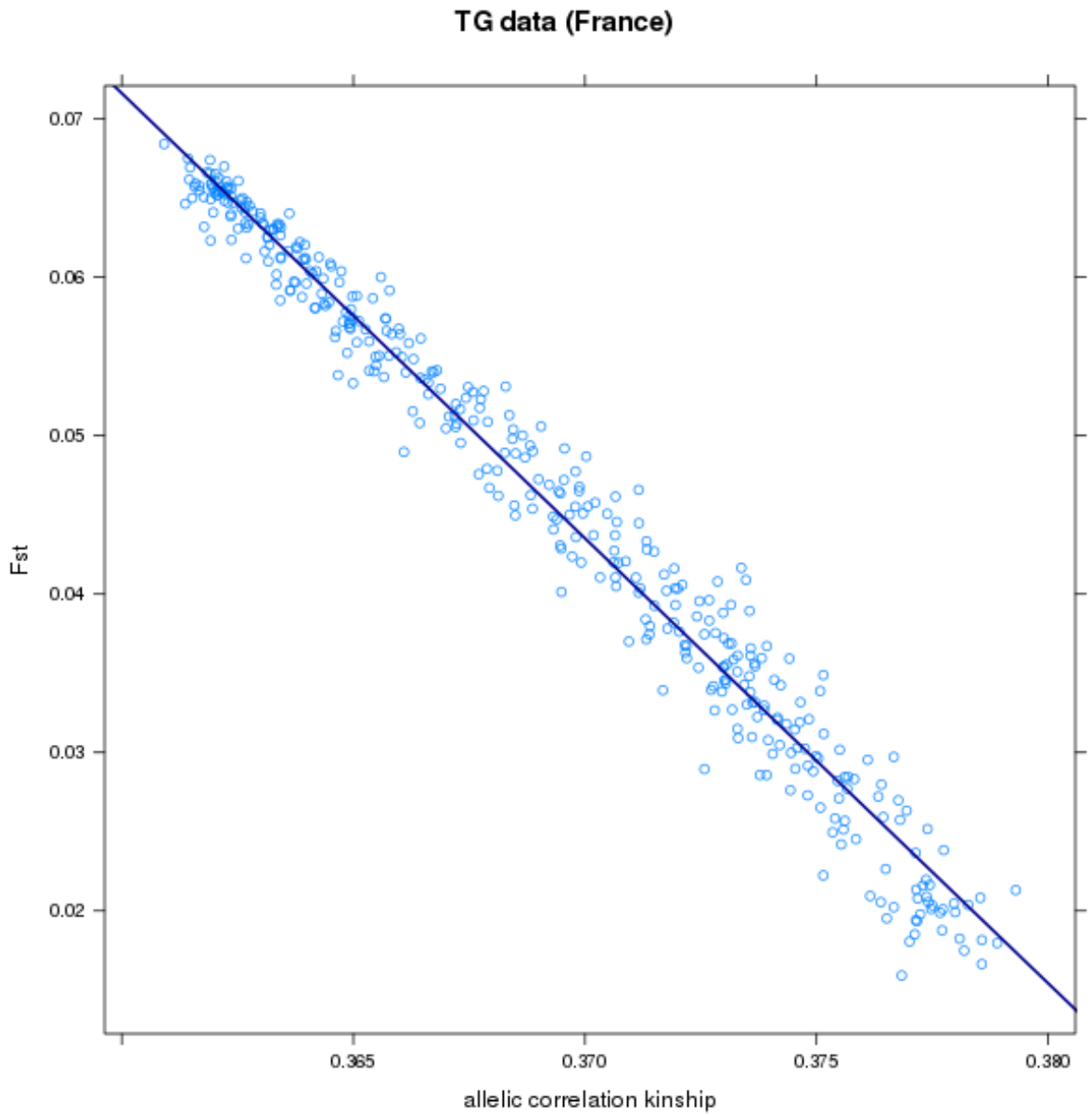
Why is That Useful for Genomic Selection?

The main application of genomic prediction models to plants and animals is to help in selecting individuals with desired phenotypes of commercial interest in the context of breeding programs.

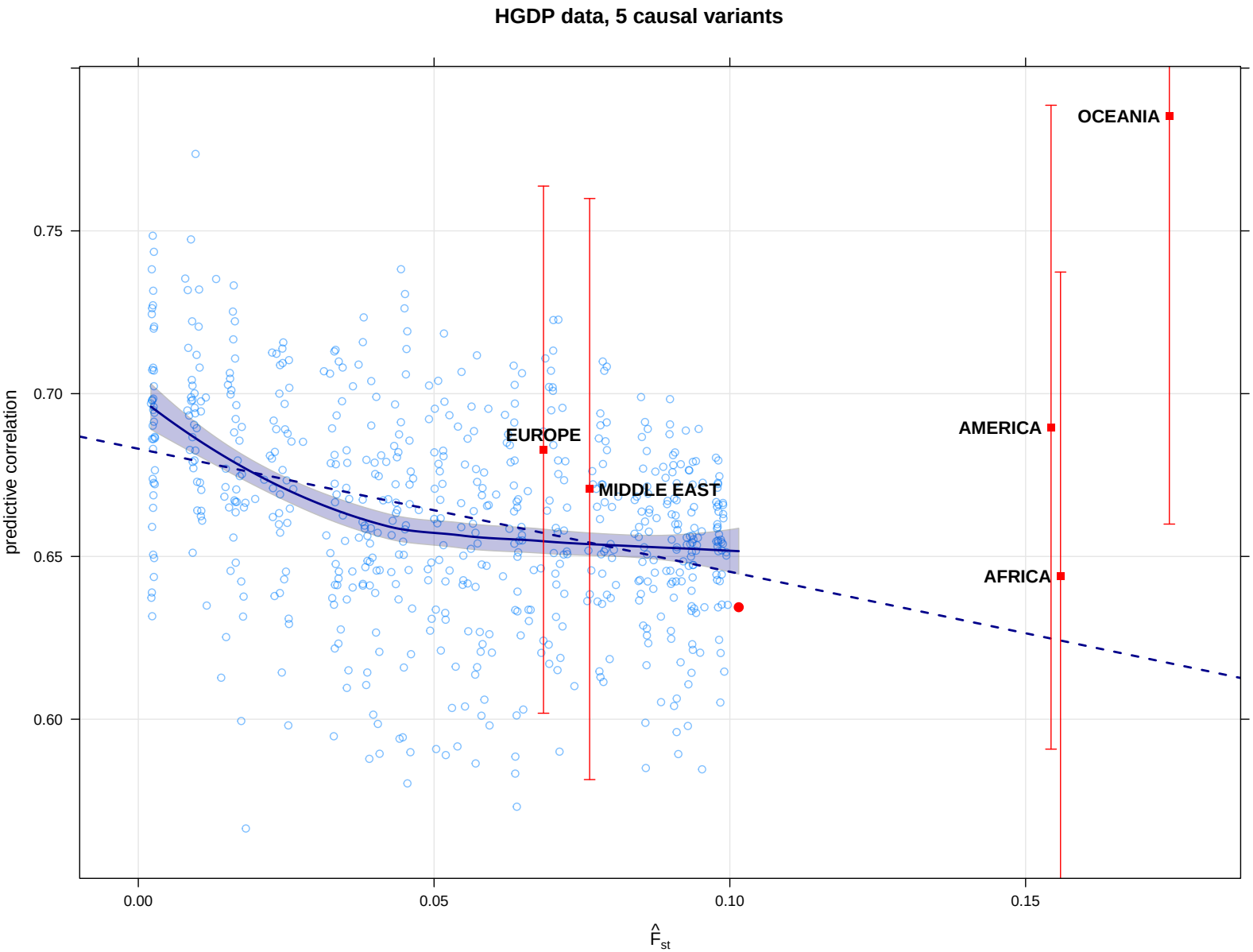
- Systematic selection to fix favourable variants in a pool of inbred individuals results in target populations that are always different from the training (e.g. future generations for later rounds of selection).
- Individuals from other populations are periodically included in the program to maintain a suitable level of genetic variability; but they must be evaluated first.
- Genomic selection models must be retrained every few generations to maintain accuracy, but not too often for cost reasons.

Since it is often possible to gauge genetic distances in terms of F_{ST}, we can read the expected predictive correlation from the curve for that $\hat{F}_{\mathrm{ST}}$ and take informed decisions, e.g., is the model still accurate enough or is it time to retrain it?

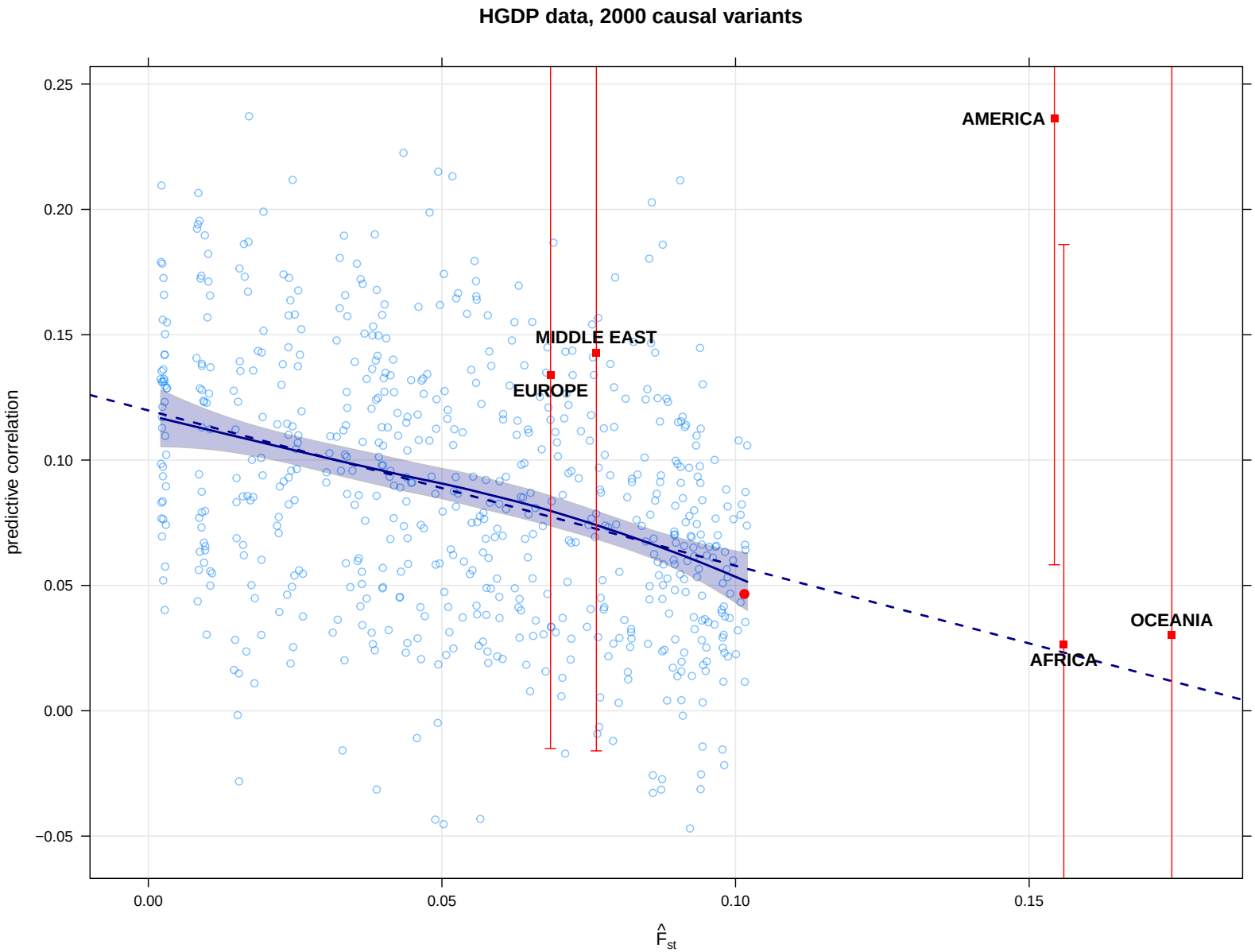
Mean Kinship and F_{ST} Really are Interchangeable



Simulation: Human Populations (Few Causal Variants)



Simulation: Human Populations (More Causal Variants)



Why is That Useful in Human Genetics?

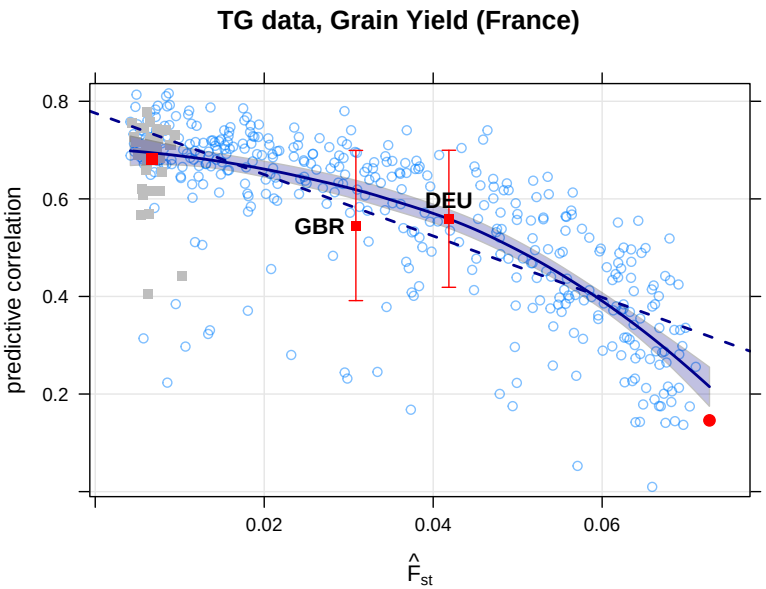
- Association mapping and trait prediction are often based on samples collected from a single ethnic group – such as Caucasians – but then results are referenced in more general contexts.

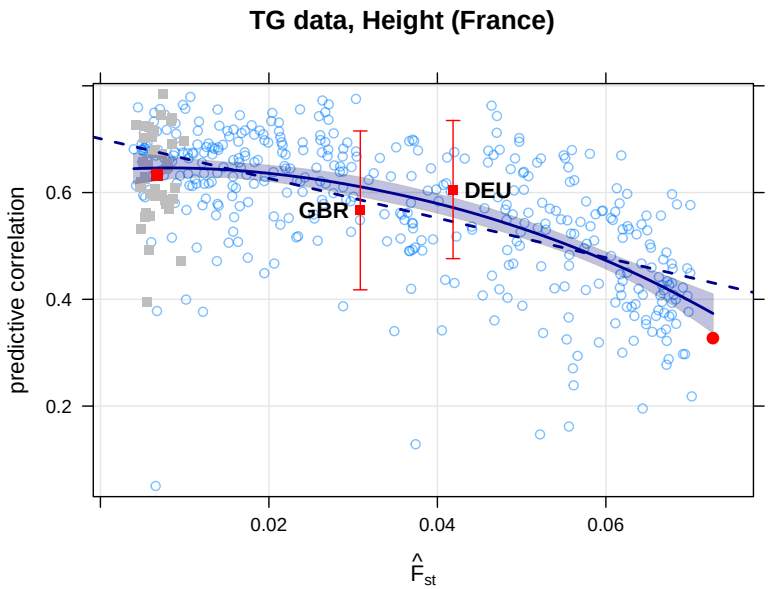
- Even assuming two populations are closely related, causal variants differ in both frequency and effect size [6]. Lactose persistence is a known example, it is driven by different variants in various way in different human populations [10].

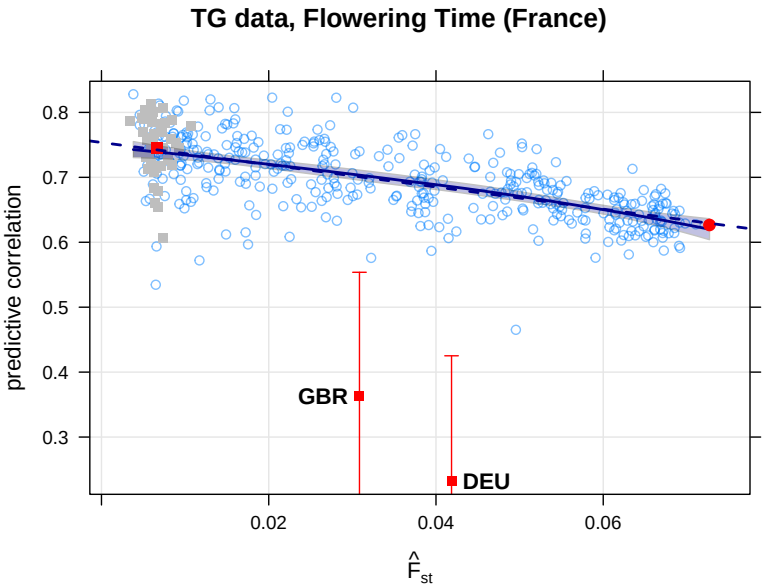
- Even when taking population structure into account, classic cross-validation overestimates predictive accuracy because random splits are at $\hat{F}_{\mathrm{ST}} \approx 0$ from each other.

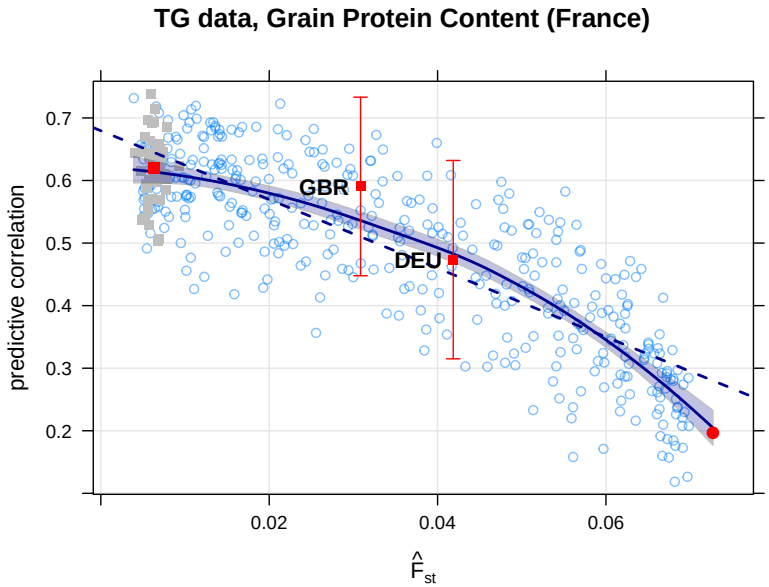
It is important to take this in consideration to develop and to improve the performance of medical diagnostics for general use.

Real Data: Four Traits from the TG Data

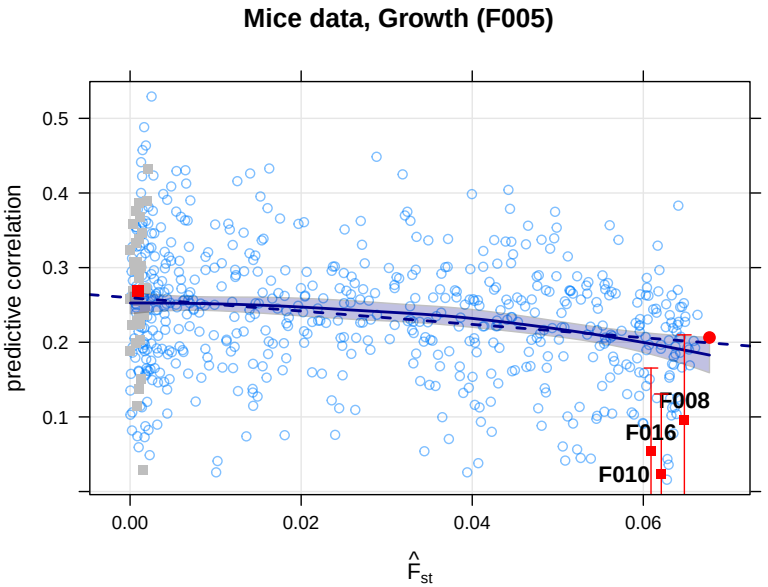


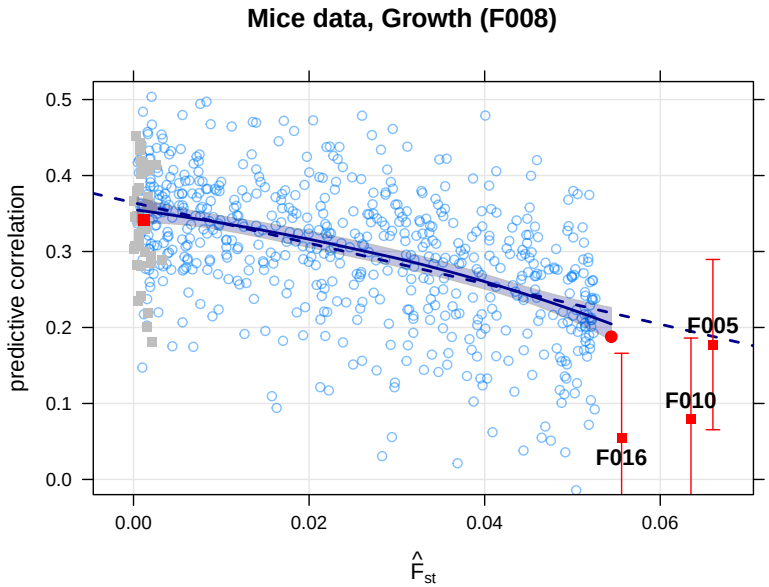


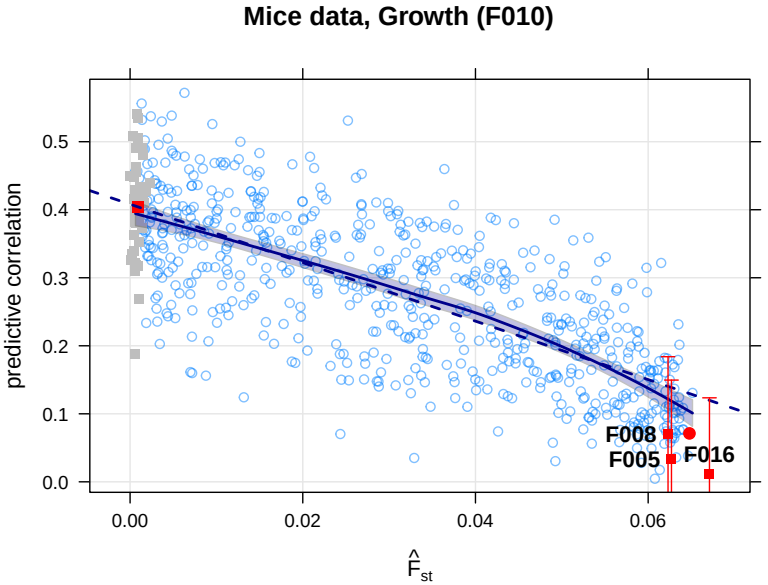


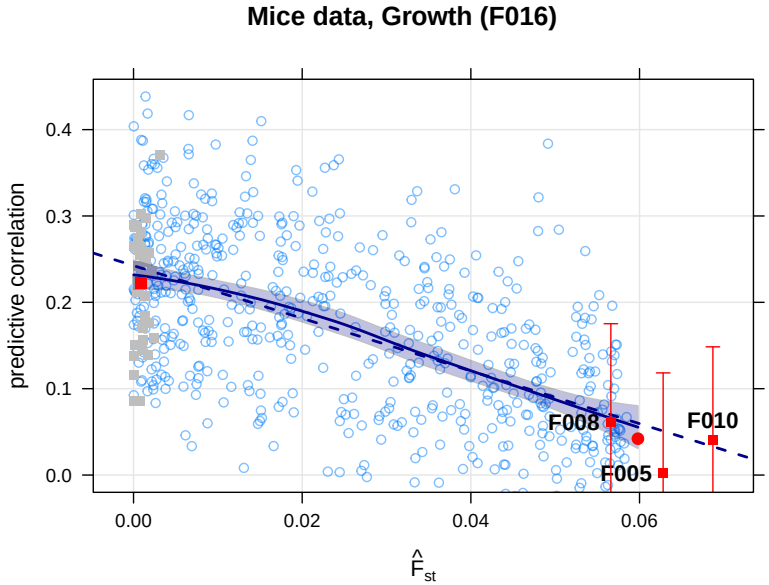


Real Data: Growth from the WTCCC Mice Data

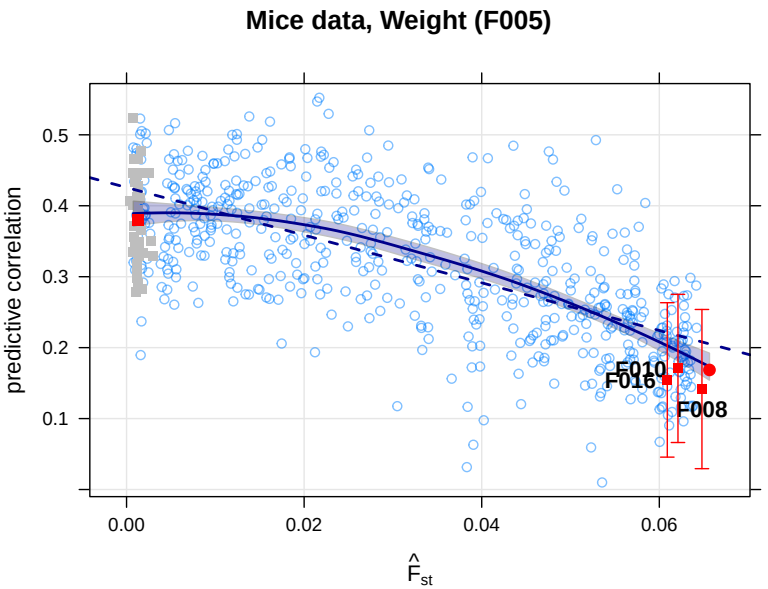


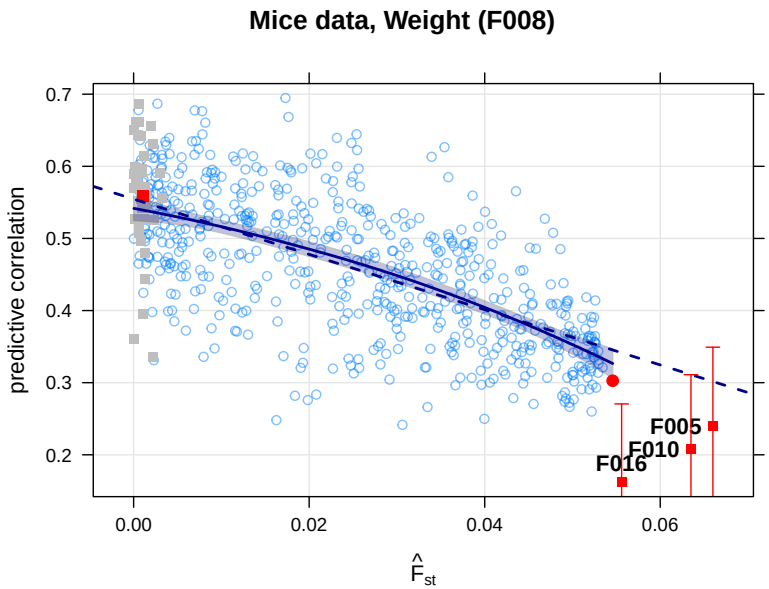


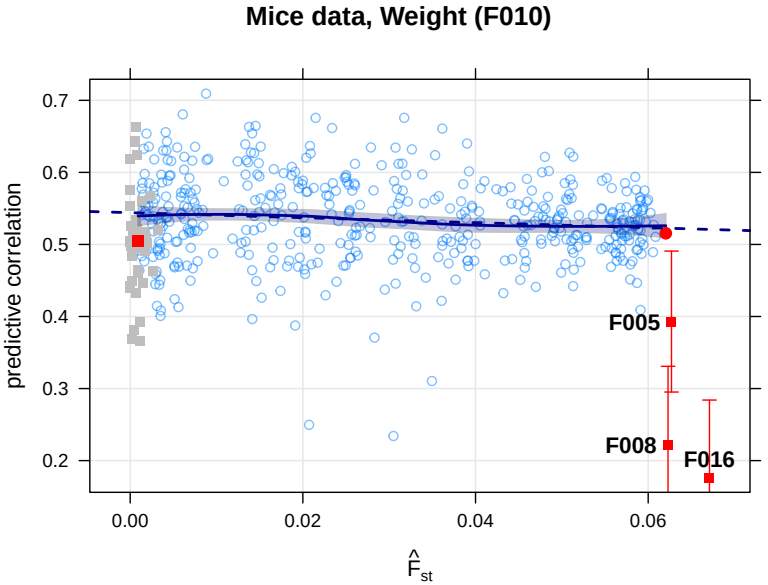


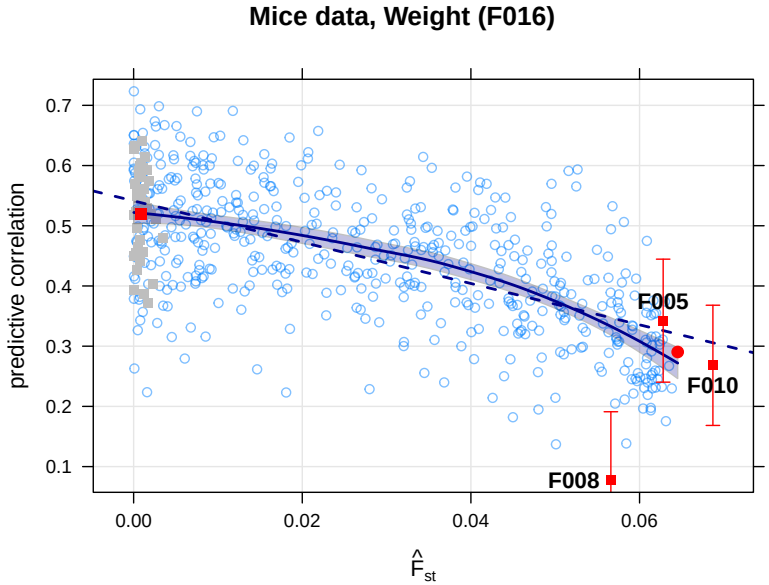


Real Data: Weight from the WTCCC Mice Data









Conclusions

- The target population is not necessarily available or even known when training the model, we are often limited to extrapolating a decay curve from the training population. A possible way to do that is through clustering and random swapping; any genomic prediction model can be plugged in.
- In the context of a breeding program the reliability of the decay curve depends on the polygenic nature of the trait being predicted in genomic selection; this is less true in human genetics.
- Being an extrapolation, the reliability of the curve decreases as F_{ST} increases. Its linear interpolation has the same problem for $\hat{\rho} \approx 0$.
- With different interpretations, such a decay curve has applications in all of plant, animal and human genetics.

References I

W. Astle and D. J. Balding.
Population Structure and Cryptic Relatedness in Genetic Association Studies.
Statistical Science, 24(4):451–471, 2009.

D. J. Balding.
Likelihood-based inference for genetic correlation coefficients.
Theoretical Population Biology, 63(3):221–230, 2003.

A. R. Bentley, M. Scutari, N. Gosman, S. Faure, F. Bedford, P. Howell, J. Cockram, G. A. Rose, T. Barber, R. Horsnell, C. Pumfrey, E. Winnie, J. Shacht, K. Beauchêne, S. Praud, A. Greenland, D. J. Balding, and I. Mackay.
Applying Association Mapping and Genomic Selection to the Dissection of Key Traits in Elite European Wheat.
Theoretical and Applied Genetics, 127(12):2619–2633, 2014.

W. S. Cleveland, E. Grosse, and W .M. Shyu.
Local Regression Models.
In J. M. Chambers and T. J. Hastie, editors, *Statistical Models in S*. Chapman & Hall, 1993.

References II

H. D. Daetwyler, M. P. L. Calusand R. Pong-Wongand G. de los Campos, and J. M. Hickey.
Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking.
Genetics, 193(2):347–365, 2013.

A. P. W. de Roos, B. J. Hayes, and M. E. Goddard.
Reliability of Genomic Predictions Across Multiple Populations.
Genetics, 183(4):1545–1553, 2009.

J. M. Hickey, S. Dreisigacker, J. Crossaa, S. Hearne, R. Babu, B. M. Prasanna, M. Grondona, A. Zambelli, K. Mathews V. S. Windhausen and, and G. Gorjanc.
Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation.
Crop Science, 54(4):1476–1488, 2015.

J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. M. Myers.
Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation.
Science, 319(5866):1100–1104, 2008.

References III

R. Makowsky, N. M. Pajewski, Y. C. Klimentidis, A. I. Vazquez, C. W. Duarte, D. B. Allison, and G. de los Campos.
Beyond Missing Heritability: Prediction of Complex Traits.
PLoS Genetics, 7(4):e1002051, 2011.

S. A. Tishkoff, F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, J. S. Silverman, K. Powell, H. M. Mortensen, J. B. Hirbo, M. Osman, M. Ibrahim, S. A. Omar, G. Lema, T. B. Nyambo, J. Ghori, S. Bumpstead, J. K. Pritchard, G. A. Wray, and P. Deloukas.
Convergent Adaptation of Human Lactase Persistence in Africa and Europe.
Nature Genetics, 39(1):31–40, 2006.

W. Valdar, L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman, W. O. Cookson, M. S. Taylor, J. N. Rawlins, R. Mott, and J. Flint.
Genome-Wide Genetic Association of Complex Traits in Heterogeneous Stock Mice.
Nature Genetics, 8:879–887, 2006.