**Working title:**

Combining genomic prediction with citizen science for crop improvement, as case study with “early selection” in durum wheat

**Authors:**

Kauê de Sousa, Matteo Dell’Acqua, Carlo Fadda, Yosef Gebrehawaryat, Jean-Luc Jannink, Dejene Mengistu, Mario Enrico Pè, Svein Ø. Solberg, Jacob van Etten, … add others

**What we have:**

1. Station data for 400 farmer varieties (genebank accession) and a few modern varieties (derived from Mediterranean germplasm) in two regions of Ethiopia
2. Citizen science (tricot) data for > 60 varieties (a subset from the 400) in 3 seasons with >1k plots in total, 3 varieties + check for each plot (Calculate how many times each variety is repeated)

* Each observation has a rank of 4 randomised varieties, with planting dates, geographic location and some other information, such number of seeds per spike and reason for the rank

High-density SNP genomic data for 400 varieties, using Illumina

**Research questions:**

* Is genomic selection with citizen science a viable approach?
* Does the genetic information provide sufficient gain in the predictions for variety selection (given the cost)?
* Can the combination of genomic and citizen science provide new insights in which traits farmers prioritise or which areas of the genome play a role?
* Does genomic prediction using phenotypic data produced on station (agronomic traits, farmer preferences) improve accuracy in the field as measured in the tricot trial?

**Analytical/Model approaches**

*Run a number of scenarios from varieties selection using the citizen science and station data to determine whether the genetic information provides any gain or not*

Scenario 1

Baseline: tricot without genomic prediction, blocked cross-validation with seasons as blocks

* Plackett-Luce model
* Bradley-Terry model

Improvement: same exercise, but adding genomic data as explanatory data

* Plackett-Luce model with covariance matrix derived from genomic data
* Bradley-Terry model with genomic random variables in a GLM framework(?)[[1]](#footnote-1)

Scenario 2

Baseline: farmer preferences from station rating exercise, predicting field performance in tricot perhaps stratifying by environment and using the two station sites separately for their respective most similar environments (to circumvent GxE)

Predict rating data with an appropriate GLM model for ordinal data (the rates should not be assumed to be equidistant, the difference between rates 1 and 2 may be different from the rates between 4 and 5).

Improvement: same farmer preferences + genomic data, letting the ordering be influenced by genomic relatedness

The exercise could use rank correlation, Kendall´s tau as criterion, but could also make a real “selection” (say top 50%) of the accessions and determine the gain as the % that indeed falls in the top 50% in the tricot trial. A p value for this exercise could be determined by doing 10,000 random selections of 50% to get a null distribution.

**Topics to explore:**

* What is the relative costs of these approaches
* Is data missingness non-random? Can data missingness be predicted from farmer preferences?
* Can we say something about which genomic regions play a role?

1. We have a paper under revision (Steinke et al.) that uses a permutation approach to determine the contribution of different co-variates. This approach could be used to correct for the fact that rankbreaking with the BT model produces overoptimistic p values. [↑](#footnote-ref-1)