Combining genomic prediction with citizen science for crop improvement and early selection in durum wheat

Kauê de Sousa1,2, Jacob van Etten2, Svein Ø. Solberg1, Mario Enrico Pè3, Carlo Fadda4, Yosef Gebrehawaryat4, Jean-Luc Jannink5,6, Basazen Fantahun Lakew3, Dejene Mengistu7, Jesse Poland8, [… add others], Matteo Dell’Acqua3[\*]  
1 Department of Agricultural Sciences, Inland Norway University of Applied Sciences, 2318 Hamar, Norway  
2 Bioversity International, Costa Rica Office, 30501 Turrialba, Costa Rica  
3 Scuola Superiore Sant’Anna, Institute of Life Sciences, 56124 Pisa, Italy  
4 Bioversity International, Ethiopia Office, 1000 Addis Ababa, Ethiopia  
5 College of Agriculture and Life Sciences, Cornell University, 14853 Ithaca, USA  
6 Agricultural Research Service, United States Department of Agriculture, 148532901, Ithaca, USA  
7 Bioversity International, Nepal Office, 3055 Kathmandu, Nepal  
8 Kansas State University, Kansas, USA  
[\*]Correspondence should be addressed to MDA (email: [m.dellacqua@santannapisa.it](mailto:m.dellacqua@santannapisa.it))

# Research questions:

* Does farmers’ evaluations associated with genes provide better predictions on varietal performance then station-only data?
* Does farmer’s crowdsourced evaluations associated to SNP markers improve GxE estimates?
* Can we bring a new breeding paradigm where farmers actively participate in plant selection?

# Methods

## Genomic prediction

Methods from RBlup to genomic prediction with SNP markers. Farmers’ evaluation from citizen science trials are connect to this approach by adding the probability of winning from Plackett-Luce as explanatory variables.

Other explanatory variables include the metric traits from station experiments.

## GxE interactions

Split varieties from citizen science trails by groups from their origin (cold tolerants and low regions). This cal also be used in the genomic prediction.

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(?)

**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 for discussion (or keep in mind)

* GxE driven by under-representative samples or farm management.
* Is data missingness non-random? Can data missingness be predicted from farmer preferences?
* Can we say something about which genomic regions play a role?

# References