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

Kauê de Sousa1,2, Jacob van Etten2, Carlo Fadda3, Yosef Gebrehawaryat3, Jean-Luc Jannink4,5, Basazen Fantahun Lakew6, Dejene Mengistu7, Svein Ø. Solberg1, Mario Enrico Pè6, Jesse Poland8, [… add others], Melfa and Workaye Farmer Community, Matteo Dell’Acqua6[\*]  
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 Bioversity International, Ethiopia Office, 1000 Addis Ababa, Ethiopia  
4 College of Agriculture and Life Sciences, Cornell University, 14853 Ithaca, USA  
5 Agricultural Research Service, United States Department of Agriculture, 148532901, Ithaca, USA  
6 Scuola Superiore Sant’Anna, Institute of Life Sciences, 56124 Pisa, Italy  
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:**

* Can farmers’ evaluations better predict varietal performance associated with genes?
* Can we use crowdsourcing to improve GxE estimates?
* What is the ability of local varieties to outperform varieties from other zones?

# Introduction

* Crowdsourcing citizen science showed the ability to register the climate effects as they occur in farms, with potential to be used in plant breeding programs and risk assessments for farm adaptation1.

# Methods

## 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(?)

**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.

# Results

# Discussion

* 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?

# References

1. van Etten, J. *et al.* Crop variety management for climate adaptation supported by citizen science. *Proceedings of the National Academy of Sciences* **116**, 4194–4199 (2019).