

**Training to validation population relatedness affects genomic prediction accuracy: a cassava case study**

**Delphine Ly<sup>1,2</sup>, Martha Hamblin<sup>2</sup>, Ismail Rabbi<sup>4</sup>, Gedil Melaku<sup>4</sup>, Bakare Moshood<sup>4</sup>, Richardson Okechukwu<sup>4</sup>, Alfred G.O. Dixon<sup>5</sup>, Peter Kulakow<sup>4</sup>, and Jean-Luc Jannink<sup>2,3</sup>**

1. Montpellier Supagro, Montpellier, France

2. Cornell University, Ithaca, NY, USA

3. USDA-ARS, R.W. Holley Center for Agriculture and Health, Ithaca, NY, USA

4. International Institute for Tropical Agriculture, Ibadan, Nigeria

5. Sierra Leone Agricultural Research Institute (SLARI), Tower Hill Free Town

lyd@supagro.inra.fr

**Abstract**

Prior to implementation of genomic selection, an evaluation of the potential accuracy of prediction can be obtained by cross validation. In this procedure, a population with both phenotypes and genotypes is split into training and validation sets. The prediction model is fitted using the training set, and its accuracy is calculated on the validation set. The degree of genetic relatedness between the training and validation sets, however, may strongly influence the expected accuracy. We developed a method to assess the importance of this effect and tested it in cassava. We used historical phenotypic data available from the International Institute of Tropical Agriculture Genetic Gain trial. DNA from clones in this trial was extracted and SNPs scored using genotyping by sequencing. A cross validation sampling scheme that prevented the training and validation sets from sharing close relatives was compared to one that forced this condition. Over 19 traits encompassing important agronomic, disease resistance, and morphological traits, plot basis heritabilities ranged from 0.04 to 0.66 and the correlation between predicted and observed phenotypes ranged from 0.15 to 0.47. Across traits, including close relatives in the training population increased accuracy between 0.00 and 0.07, a small but consistent effect over traits. We conclude that genomic selection has potential to accelerate gains in cassava and the existing training population should give a reasonable estimate of future prediction accuracies.