RUFORUM Working Document Series (ISSN 1607-9345), 2018, No. 17 (1): 755-760.
Available from *http://repository.ruforum.org*

Research Application Summary

Incorporating genomic markers and weather variables in genotype-by-environment interaction analyses for cassava brown streak disease

Ozimati, A.[1*] Kawuki, R.,[1] Esuma, W.,[1] Akdemir, D.,[2] Wolfe, M., [1] & Jannink, Jean-Luc[1]

National Crops Resources Research Institute (NaCRRI), P.O.Box 7084, Kampala, Uganda
[1]School of Integrative Plant Science, Plant Breeding and Genetics Section, Cornell University, Ithaca, NY 14853, USA
[2]Department of Statistical Consultancy, Cornell University, Ithaca, NY 14853, USA
*Corresponding author: ozimatialfred@gmail.com.

Abstract

Genotype-by-environment (GxE) interactions is a reality that scientists deal with when developing new and better varieties. With increase in scale of both phenotypic and genetic data, coupled with environmental data, prediction of environment-specific multi-environment trials (MET) is gaining importance. We leveraged phenotypic and genotypic data for ~150 clones and five checks evaluated in 31 environments (location-season-year combination) to define our mega environments. Further, we used this dataset to counter different prediction problems faced in cassava breeding such as (i) predicting for unobserved genotypes across environments, (ii) predicting for unobserved genotype in never evaluated environments, and (iii) making predictions for unobserved environments. No clear grouping of the environments was observed, based on the planting seasons or proximity of the trials from the phenotypic data of the five-checks. Our prediction accuracies for the three prediction strategies ranged from 0.47 in CV2 to 0.91 in CV3 for CBSDRs. From this study, we established that CBSD (assessments undertaken at three, six and at harvest) can be predicted with reasonable accuracies under different scenarios that mimic real problems encountered in cassava breeding.

Key words: Multi-environment trials, genomic predictions, cassava, cassava brown streak diseases

Résumé

Les interactions génotype - environnement (GxE) sont une réalité à laquelle les scientifiques font face lorsqu'ils développent de nouvelles et meilleures variétés. Avec l'augmentation de l'échelle des données phénotypiques et génétiques, couplée aux données environnementales, la prédiction des essais multi-environnements spécifiques à l'environnement (MET) est en train de gagner de l'importance. Nous avons exploité les données phénotypiques et génotypiques pour environ 150 clones et cinq contrôles évalués dans 31 environnements (combinaison emplacement-saison-année) pour définir nos méga-environnements. De plus, nous avons utilisé cet ensemble de données pour contrer les différents problèmes de prédiction rencontrés dans la sélection du manioc tels que (i) la prédiction de génotypes non observés à travers les environnements, (ii) la prédiction de génotype non observé dans des environnements jamais évalués, et (iii) la réalisation de prédictions pour les environnements non observés. Aucun regroupement clair des environnements n'a été observé, basé sur des saisons de plantation ou la proximité des essais à partir des données phénotypiques des cinq contrôles. Nos précisions de prédiction pour les trois stratégies de prédiction variaient de 0,47 en CV2 à 0,91 en CV3 pour les CBSDRs. De cette étude, nous avons établi que la CBSD (évaluations effectuées à trois, six et à la récolte) peut être prédite avec une précision raisonnable dans différents scénarios qui imitent les problèmes réels rencontrés dans la sélection du manioc.

Mots clés: Essais multi-environnementaux, prédictions génomiques, manioc, maladies des striures brunes du manioc

Introduction

The target of a plant breeding programme is to release varieties that perform consistently better than the existing varieties grown by farmers (Bernardo, 2003). This ambition is often frustrated by variation in the responses of genotypes to the diversity of conditions in farmers' fields, a phenomenon known as genotype-by-environment (G x E) interaction. To circumvent the impact of G x E and develop varieties with wide adaptation, plant breeders conduct extensive Multi-Environment Trials (MET).

Recently, plant breeding has undergone a revolution due to an increase in the scale of both phenotypic and genetic data generation, further boosted by increase in computational efficiency in linear mixed model framework (Poland and Rife, 2012). The emergence of such big data and the techniques necessary to analyze them have enabled the application of a new breeding and selection method known as genomic selection (Meuwissen *et al.,* 2001). In genomic selection (GS), the breeding value of new individuals, not yet observed in the field can be predicted at early stages based on their genetic relationships to a phenotyped and genotyped calibration set known as the training population (TP) (Hayes *et al.,* 2009).

From these massive data generated (phenotypic, genotypic and environmental variables), G x E can be precisely modelled in GS framework. A number of studies have tested prediction models that incorporate environmental covariates, allowing information sharing from environments of interest (Heslot *et al.,* 2014; Lado *et al.,* 2016; Jarquín *et al.,* 2017; Ly *et al.,* 2018). These studies indicated substantial increases in genomic prediction accuracies are possible when G x E is modelled explicitly using environmental covariates because they enable information sharing among environments. Given these benefits, the use of environmental covariates in genomic prediction models could be worthwhile in cassava. Genomic prediction of G x E in cassava is expected to permit more optimal resource allocation to boost genetic gains without significantly increasing costs in cassava breeding.

Materials and Methods

A total of 150 clones and five checks were planted at 10 sites in an augmented design. These sites were chosen to represent the major cassava production and consumption patterns in Uganda. Each check was replicated 5-6 times per block. We collected data on Cassava Brown Streak Disease (CBSD) foliar symptoms scored at three and six months after planting (3 and 6 MAP), using the standard scoring scale of 1-5 (IITA, 1990; Hillocks and Thresh, 2000). At harvest, all roots in a plot were pooled and assessed individually for CBSD necrosis. Each root was cut transversely into pieces, and the cross-sections were scored for necrotic symptoms, also on a scale of 1-5. In addition to the phenotypic and genotypic data, weather data loggers (HOBO onset weather loggers) were installed at each of the 10 experimental sites to record rainfall (mm), temperature (oC), solar radiation (lux) and relative humidity (%) for the two years of the experiment.

To define the mega environments, we constructed a distance matrix based on the normalized phenotypic data for cassava brown streak disease, using the dist function in (**Ref)** Further for visualization, we used the hierarchical clustering function hclust to generate a dendrogram using the "Ward. D" method (Murtagh and Legendre, 2011). For genomic predictions, three prediction models were tested (i) a G-BLUP model with genotypic and environmental main effects, where the environmental main-effect included a covariate-based variance-covariance matrix (Ω) and the genotype main-effect was fitted with the SNP-based realized relationship matrix (G) (ii) a G-BLUP with genotypic main-effect as in (i), but with environment main-effect fitted as independent and identically distributed (IID) and the G x E term was modelled without the environmental covariates, and (iii) G-BLUP with G x E variance-covariance matrix estimated from the realized genomic relationship matrix and environments variance-covariance matrix derived from weather.

Similarly, three cross validations (CV) strategies were evaluated. In the first cross validation (CV1) strategy, genomic predictions were made for genotypes that have not been evaluated in any environments. This prediction scenario mimics predicting the performance of newly developed genotypes (crosses) or introductions into the breeding program. The second cross-validation scenario (CV2) involved predicting the performance of a subset of individuals who were unobserved in a subset of environments as a strategy to assess ability to predict the performances of a clone in an environment where it has not yet been evaluated, but where some training data is available. The last prediction problem (CV3), did not involve random cross-validation per-se. This was a scenario predicting the performance of clones in totally unobserved environments, also referred to as the "leave-one-environment-out" scheme in Jarquín e*t al.* (2017).

Results and discussions

Defining mega environments. From the phenotypic data of the five checks, no clear distinction of the 31 environments (location-season-year combination) was observed at a tree height of 4, based on the growing season, which we refer to as season "2015A and 2016A" indexed by location (Figure 1). However, in a few cases, trials established within a location clustered together e.g. Lira 2015A and 2015B as well as Kamuli 2015A and Kamuli 2016A. Cassava brown streak disease caused by cassava brown streak virus (CBSV) and Uganda cassava brown streak virus (UCBSV) was initially endemic to East African coastal region and re-emerged in Uganda about 20 years ago (Alicai *et al.,* 2007). More recently, cassava brown streak disease has been observed in all major cassava production areas in Uganda. Therefore, lack of grouping of environments based the disease assessment of the five checks, suggests continuous need for multi-locational evaluations across major cassava production zones in the process of developing clones for variety release.

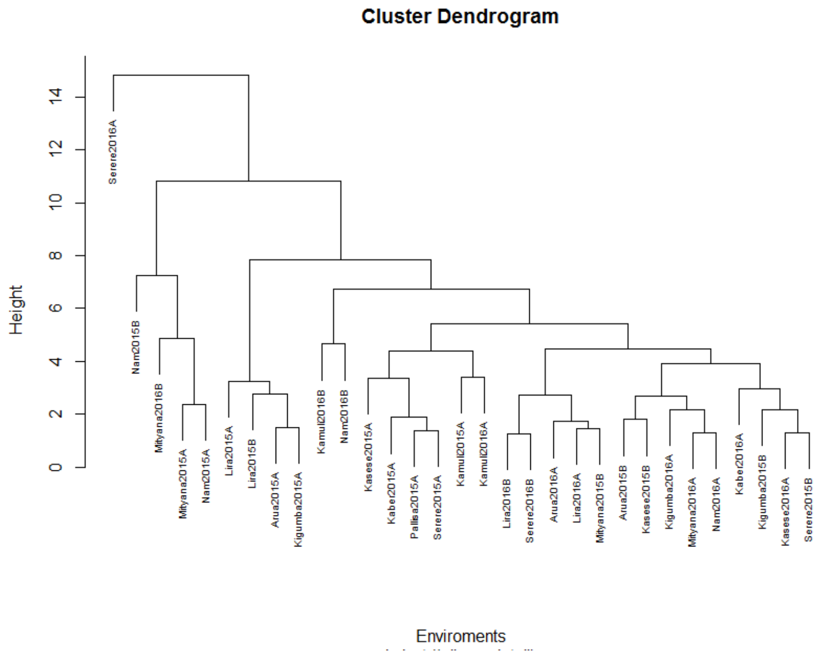

Figure 1. Clustering of the environments using CBSD-related traits for the five-checks, evaluated across 31 environments (Location-season-year combination

Partitioning of total phenotypic variance in a model with inclusion of G x E term. The proportion of the total phenotypic variance attributed to genotypic main effect were 15.8%, 25.4% and 32.1% for CBSD3s, CBSD6s, and CBSDRs (location-season-year combination), respectively. The observed genotypic variances were greater than for the G x E and environment main effect variances, except CBSD3s where the variance explained by the environment main effect was 19.8% (Table 2), indicating that selection of clones for variety release can be made on performance assessment in major cassava production zone**s.**

Table 2. Partitioning of the variances for cassava brown streak disease assessed at three time points

Traits	Proportion of variance explained by the model predictors (%)		
	Genotype-by-Environments	Environments	Genotypes
CBSD3s	12.0	19.8	15.8
CBSD6s	13.5	8.8	25.4
CBSDRs	15.8	1.8	32.1

Table 3. Mean prediction accuracies across folds and models tested

Traits	Prediction accuracies for the three cross validations (CV) strategies		
	CV1	CV2	CV3
CBSD3s	0.58±0.02	0.58±0.01	0.83±0.02
CBSD6s	0.60±0.02	0.59±0.01	0.88±0.01
CBSDRs	0.48±0.04	0.47±0.02	0.91±0.0**2**

Genomic prediction accuracies for three prediction strategies. For cassava brown streak measured at the three time points (CBSD3s, CBSD6s and CBSDRs), similar mean prediction accuracies were recorded for CV1 and CV2, ranging from 0.47 for CBSDRs in CV2 to 0.60 in CV1 prediction strategy for CBSD6s (Table 3). In general, the highest mean prediction accuracies were recorded for leave-one-environment-out, also termed as CV3 predictions strategy, varying from 0.83 for CBSD3s to 0.91 for CBSDRs. Our cross-validation prediction accuracies were higher than previously reported cross-validation prediction accuracies for CBSD3s,CBSD6s and CBSDRs in NaCRRI training population (Kayondo *et al.,* 2018), suggesting more phenotypic data on genotypes across target population of environments (TPOE) in the present study and inclusion environmental variates enhanced genomic prediction accuracies.

Conclusion

This study provides insight into the incorporation of environmental variables into genomic prediction models used in cassava breeding to assess cassava brown streak disease performance in light of G x E. Based on the results of the study, CBSD3s and CBSD6s, CBSDRs can reasonable be predicted to counter the different prediction problems in cassava breeding such as, predicting the performance of newly generated seedlings (crosses) as well as unobserved environments, hence cutting the cost of field evaluations.

Acknowledgement

This paper is a contribution to the 2018 Sixth African Higher Education Week and RUFORUM Biennial Conference.

References

Alicai, T., Omongo, C.A., Maruthi, M.N., Hillocks, R.J.,. Baguma, Y., Kawuki, R., Bua, A., Otim-Nape, G.W. and Colvin, J. 2007. Re-emergence of Cassava Brown Streak Disease in Uganda. *Plant Dis.* 91 (1): 24–29.

Bernardo, R. 2003. Breeding for Quantitative Traits in Plants. Stemma Press, Woodbury.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J. and Goddard, M.E. 2009. Invited review: Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92: 433–443.

Heslot, N., Akdemir, D., Sorrells, M.E. and Jannink, J.L. 2014. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theor. Appl. Genet.* 127 (2): 463–480.

Hillocks, R.J. and J.M. Thresh. 2000. Cassava mosaic and cassava brown streak virus diseases in Africa. *Root* 7: 1–8.

IITA. 1990. Cassava in Tropical Africa: A reference manual. 3: 1–176.

Jarquín, D., C. Lemes da Silva, C. Gaynor, R.C., Poland, J., Fritz, A., Howard, R., Battenfield, S. and Crossa, J. 2017. Increasing genomic-enabled prediction accuracy by modeling genotype × environment interactions in Kansas wheat. *The Plant Genome* 10 (2). DOI: 10.3835/plantgenome2016.12.0130

Kayondo, S.I., Del Carpio, D.P., Lozano, R.,. Ozimati, A., Wolfe, M., Baguma,Y., Gracen, V., . Offei, S., Ferguson, M., Kawuki, R. and Jannink, J.L. 2018. Genome-wide association mapping and genomic prediction for CBSD resistance in *Manihot esculenta*. *Sci. Rep.* 8: 1–11.

Lado, B., Barrios, P.G., Quincke, M., Silva, P. and Gutiérrez, L. 2016. Modeling genotype× environment interaction for genomic selection with unbalanced data from a wheat breeding program. *Crop Science* 56 (5): 2165-2179.

Ly, D., Huet, S., Gauffreteau, A., Rincent, R., Touzy, G., Mini, A., Jannink, J.L., Cormier, F., Paux, E., Lafarge, S., Le Gouis, J. and Charmet, G. 2018. Whole-genome prediction of reaction norms to environmental stress in bread wheat (*Triticum aestivum* L.) by genomic random regression. *F. Crop. Res.* 216: 32–41.

Meuwissen, T. H. E. , Hayes, B. J. and Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense markers maps. *Genetics* 157: 1819–1829.

Murtagh, F. and Legendre, P. 2011. Ward's hierarchical clustering method: clustering criterion and agglomerative algorithm. pp.1–20.

Poland, J.A. and Rife, T.W. 2012. Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome J.* 5 (3): 92 - 102.