University of São Paulo
"Luiz de Queiroz" College of Agriculture

Enviromics, nonlinear kernels and optimized training sets for a climate-smart genomic prediction of yield plasticity in maize

Germano Martins Ferreira Costa Neto

Thesis presented to obtain the degree of Doctor in Science. Area: Genetics in Plant Breeding

Piracicaba
2021

Germano Martins Ferreira Costa Neto
Bachelor in Agronomy

Enviromics, nonlinear kernels and optimized training sets for a climate-smart genomic prediction of yield plasticity in maize

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:
Prof. Dr. **ROBERTO FRITSCHE NETO**

Thesis presented to obtain the degree of Doctor in Science. Area: Genetics in Plant Breeding

Piracicaba
2021

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Costa Neto, Germano Martins Ferreira

Enviromics, nonlinear kernels and optimized training sets for a climate-smart genomic prediction of yield plasticity in maize / Germano Martins Ferreira Costa Neto- - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2021.

117 p.

Tese (Doutorado) - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Seleção genômica 2. Interação genótipo x ambiente 3. Tipagem de ambientes 4. Big data I. Título

DEDICATORY

To my wife Mariana and my son Augusto

- You are the love and the strength of my life

To my mothers Dara and Célia (*in memoriam*)

To my father Germano (*in memoriam*)

- You are the pillar of my life

AGRADECIMENTOS

À minha esposa, Mariana, pela compreensão, cumplicidade, incentivo e carinho.
À toda minha família por sempre darem seu apoio incondicional.

Ao Professor Dr. Roberto Fritsche Neto, pela oportunidade de integrar seu Laboratório de Melhoramento de Plantas Alógamas, além de suas orientações e todo seu apoio em momentos dentro e fora da vida acadêmica.

Ao Dr. José Crossa (e ao CIMMYT/BSU) pela receptividade, apoio e bons conselhos sobre a carreira científica

Aos irmãos de coração Luís Alberto, Karoline, Rodrigo e Pedro, pela sua amizade, apoio e iincentivo constantes, desde muito antes da minha vida na pós-graduação. Também é impossível não agradecer aos amigos de longa data, que pude conviver novamente neste periodo, além dos novos amigos que tive o prazer de conhecer, dos quais foram imprescindíveis com seu apoio, entre eles: Karina, Filipe, Ana Letycia, Pedro Henrique, Miriam e Camila.

À Escola Superior de Agricultura 'Luiz de Queiroz' (ESALQ/USP), por fornecer uma nova base e instrumentos que me permitram alçar voos que nunca havia pensado que conseguiria. À todos os professores e membros do corpo técnico por seu apoio e dedicação para com o trato dos alunos e para com a construção e transmissão do conhecimento. Agradeço também a paciência de dedicação de todos os membros das Secretarias de Pós-Graduação.

Ao CNPq pela concessão da bolsa doutorado.

Ao International Maize and Wheat Improvement Center (CIMMYT) e à Bill & Melinda Gates Foundation pelo suporte financeiro para publicação dos artigos da tese, também e no auxilio em minha estadia no México, sobretudo em um momento tão dificil quanto a da pandemia.

.

CONTENTS

RESUMO

Envirômica, kernels não-lineares e otimização de populações de treinamento na predição genômica inteligente para o clima com foco na plasticidade fenotípica em milho

A tipagem de ambientes em larga escala, ou simplesmente a envirômica, é um campo emergente de ciência de dados, tanto na pesquisa agrícola como nas rotinas de programas de melhoramento. Esta "omica" consiste em reunir e processar informações ambientais, respeitando a ecofisiologia do cultivo para, por fim, integrá-las na genômica quantitativa e na seleção baseada em modelos preditivos. No entanto, a maioria das atuais plataformas baseadas em predição aplicáveis ao melhoramento de plantas são baseadas nas relações genótipo-fenótipo, isto é; na modelagem sa variação fenotípica em função da variação genômica caracterizada por marcadores moleculares, na qual o estado da arte é denominado por seleção ou predição genômica (GP). Apesar do sucesso de seu uso em estàgios preliminares de melhoramento, sob condições restritas variações ambientais (p.ex: poucos ambientes ou um único ambiente), baixas acurácias ainda são observadas sob multiplas condições ambientais, na presença de "interação genótipo por ambiente" (G×E). Por outro lado, o conhecimento da ecofisiologia dos cultivos pode ser a alternativa para impulsionar aumentar a acuracia da GP sob G×E. Esta variação ambiental molda respostas fenotipicas específicas de cada genótipo a um dado gradiente de fatores de solo, clima e manejo isto é, a norma de reação. Nesta tese, buscamos estudar esses aspectos, através da realização de três estudos voltados para o uso de envirômica com GP sob cenários de G×E, usando para isso o rendimento de grãos de dois conjuntos de dados de híbridos de milho tropical. O primeiro estudo desta tese envolve o desenvolvimento do primeiro software de código aberto dedicado a ambitipagem (tradução proposta para o termo envirotyping) em predição genômica. Neste estudo, elucidamos o uso de sensoriamento remoto para popularizar o uso da ambitipagem, assim como aspectos de ecofisiologia úteis para compreender e definir os conceitos de 'ambiente', 'envirômica' e 'ambitipagem'. No segundo capítulo, verificamos os ganhos de acurácia adquiridos pela adoção de kernels não lineares (Gaussian Kernel, GK; Deep Kernel, DK) para modelagem de efeitos não-aditivos (p.ex: dominancia e ambitipagem), usando o tradicional GBLUP (genomic best linear unbiased predictor) como método de referência. Nossos resultados sugerem que os kernels não lineares (GK e DK) são a melhor alternativa para modelar efeitos não-aditivos e de norma de reação. A adoção de GK ou DK reduziu o tempo computacional na execução dos modelos, como também aumentou a precisão para prever interações G×E complexas/cruzadas (variações no rank dos genótipos através dos ambientes). Por fim, observamos que o uso de GK ou DK para modelagem de efeitos não-aditivos é fundamental para expandir a resolução da GP em predizer a interação de um hibrido de milho particular através de múltiplos ambientes. Finalmente, no terceiro capítulo propomos o conceito de "marcador qualitativo de ambiente", desenvolvido conciliando conceitos clássicos de ecofisiologia (Lei de Shelford) e caracterização da tipologia ambiental (isto é, frequência de ocorrência de classes qualitativas de fatores ambientais através do tempo e do espaço). A abordagem foi exemplificada com dois estudos de caso abrangendo o uso hipotético de GP sob ensaios de avaliação de em híbridos de milho em diversos ambientes. O uso combinado de envirômica e genômica possibilitou conceber uma plataforma de predição (denominada E-GP) que concilia fenotipagem seletiva (redução das populações de treinamento para GP) e predição de cenários futuros (isto é, G×E desconhecidas). Observamos que o aumento de informações fenotípicas em vários ambientes nem sempre corresponde ao aumento de acurácia da GP. Portanto, a representatividade da rede de avaliação de híbridos (genótipos mais representativos, avaliados nos ambientes "chave") é mais importante que o número de genótipos e ambientes

considerados. Através de E-GP juntamente a algorítimos genéticos, fomos capazes de selecionar as combinações G×E mais representativas, o que refletiu diretamente em uma redução drástica do tamanho da rede experimental, conciliando aumento de acurácia. Por fim, constatamos que o GBLUP sem nenhuma informação de ambitipagem é ineficiente em predizer a plasticidade fenotípica dos híbridos de milho sob múltiplos ambientes e G×E desconhecida. Com E-GP foi possível realizar uma triagem dos melhores híbridos, em termos de plasticidade fenotípica, usando reduzidas informações fenotípicas e suplementadas pelo amplo uso de genômica e envirômica. Tais resultados permitem vislumbrar abordagens inteligentes para o clima, envolvendo a redução drástica dos esforços de testes de campo à medida que aumenta o uso consciente de envirômica (e ambitipagem) combinada com genômica.

Palavras-chave: Seleção genômica, Adaptabilidade, Tipagem de ambientes, Ciência de dados

ABSTRACT

Enviromics, nonlinear kernels and optimized training sets for a climate-smart genomic prediction of yield plasticity in maize

Large-scale envirotyping (environmental + typing) or simply enviromics, is an emerging field of data science, applied both in agronomic research and plant breeding. This "omics" consists of gathering and processing reliable environmental information, respecting the crop-specific ecophysiology aspects, then for further integration of this data into quantitative genetics and prediction-based breeding. However, most of the current prediction-based platforms are based on genotype-phenotype relationships (i.e., the phenotype-genotype association enabled by whole-genome markers), in which the state-of-art of this approach in the context of predictive breeding is so-called genomic selection or prediction (GP). Despite the success of its use in preliminary breeding stages, mostly conducted under restricted environmental variations (e.g., few number of environments or a single environment), the occurrence of low accuracy values are still a reality under multiple environmental conditions, in which is detected the presence of the so-called "genotype by environment interaction" (G×E). On the other hand, knowledge of crop ecophysiology can be the alternative to boost the accuracy of GP under G×E. This environmental variation shapes genotype-specific phenotypic responses to a given gradient of soil, climate and management factors i.e., the reaction norm. In this thesis, we conducted three studies aimed to investigate the use of GP enviromics under G×E scenarios, using for this the grain yield of two datasets of tropical maize hybrids. The first study of this thesis involves the development of the first open-source software dedicated to envirotyping in genomic prediction. In this study, we elucidate the use of remote sensing to popularize the use of envirotyping, as well as aspects of ecophysiology useful to understand and define the concepts of 'environment', 'enviromics' and 'envirotyping'. In the second chapter, we verify the accuracy gains acquired by the adoption of non-linear kernels (Gaussian Kernel, GK; Deep Kernel, DK) for modeling non-additive effects (e.g., dominance and envirotyping-enabled reaction-norms) using the traditional GBLUP (genomic best linear unbiased predictor) as a reference method. Our results suggest that non-linear kernels (GK and DK) are the best alternative to model non-additive and reaction norm effects. The adoption of GK or DK reduced the computational time in running the models, as well as increased the accuracy to predict complex G×E interactions (variations in the rank of genotypes across environments). Finally, we observe that the use of GK or DK for modeling non-additive effects is critical to expand GP's resolve to predict the interaction of a particular maize hybrid across multiple environments. Finally, in the third chapter we propose the concept of 'envirotype marker', developed by reconciling classical concepts of ecophysiology (Shelford's Law) and characterization of the environmental typology (i.e., frequency of occurrence of qualitative classes of environmental factors over time and over time. space). The approach was exemplified with two case studies covering the hypothetical use of GP under evaluation trials in maize hybrids in different environments. The combined use of enviromics and genomics made it possible to design a prediction platform (called E-GP) that reconciles selective phenotyping (reduction of training populations for GP) and prediction of future scenarios (i.e., unknown G×E). We observed that the increase in phenotypic information in various environments does not always correspond to the increase in the accuracy of GP. Therefore, the representativeness of each hybrid under evaluation at the experimental network (most representative genotypes, evaluated in "key" environments) is more important than the number of genotypes and environments considered for training GP. Through E-GP together with genetic algorithms, we were able to select the most representative

G×E combinations, which directly reflected in a drastic reduction in the size of the experimental network, reconciling increased accuracy. Finally, we found that GBLUP without any envirotyping information is inefficient in predicting the phenotypic plasticity of maize hybrids under multiple environments and unknown G×E. With E-GP it was possible to screen the best hybrids, in terms of phenotypic plasticity, using reduced phenotypic information and supplemented by the wide use of genomics and enviromics. Such results allow us to envision smart approaches to climate, involving the drastic reduction of field-testing efforts as the conscious use of enviromics (and envirotyping) combined with genomics increases.

Keywords: Genomic selection, Adaptability, Envirotyping, Data Science

ABBREVIATIONS

G×E	Genotype by Environment Interaction
MET	Multi-environment trials
EC	Environmental covariate
CV	Cross-validation
GP	Genomic Prediction
GBLUP	Genomic best-unbiased predictions
GK	Gaussian Kernel
DK	Deep Kernel
A	Additive effects
D	Dominance effects
W	Quantitative environmental covariate
T	Envirotype markers (typology matrix)
E-GP	Enviromic-aided Genomic Prediction using envirotype markers
W-GP	Enviromic-aided Genomic Prediction using quantitative environmental covariables
OTS	Optmized training sets
r	predictive ability given by the average linear correlation between observed and predicted trait values
MSE	Mean squared error
FW	Finlay-Wilkinson adaptability model
b	coefficient of yield adaptability from Finlay-Wilkinson
BD	Block-diagonal genomic by environment matrix
RN	Reaction-norm genomic by enviromic matrix

1. INTRODUCTION

Plant breeding is responsible to deliver novel cultivars, of diverse economically imporant species, according to society's demand. Each novel cultivar is the end result of several years of research, which involves diverse schemes for crossing and selecting the best-evaluated genotypic combinations, according to a certain breeding program goal for some target population of environments (TPE). In order to speed up the breeder's decisions, the use of predictive tools has gained importance in the last 15 years, entering the big data era mainly due to the advance of computational tools and because of the reduction of costs in obtained large-scale omics data. Perhaps one of the oldest yet less explored omics is the large-scale envirotyping, or simply enviromics (Cooper et al., 2014; Xu, 2016; Resende et al., 2020). Its applications starts by the use of the simpler linear regressions of reaction-norm, back in the 1970s (e.g., Freeman and Perkins, 1971; Wood, 1976) and with a higher biological interpretation since the 2000s (e.g., Epinat-Le Signor et al., 2001; Romay et al., 2010 Costa-Neto et al., 2020). Moreover, the adoption of crop growth models and climatic prediction tools for characterizing the TPE of the breeding program (e.g., Messina et al., 2018; Heinemann et al., 2019; Antolin et al., 2021) is also a key analytical tool for envirotyping.

Genomic prediction (GP, Meuwissen et al., 2001) most used and powerful predictive breeding tool. It relies on Fisher's Infinitesimal model, in which the sum of whole-genome markers might be a realization of the genomic variation within a given population (Crossa et al., 2017; Voss-Fels et al., 2019). GP platforms were first designed to model the genotype-to-phenotype relations (G-to-P) under single environment conditions, e.g., in a breeding program nursery (Lorenzana and Bernardo, 2009; Windhausen et al., 2012; Zhao et al., 2012; Zhang et al., 2015). Under these conditions, the micro-environmental variations within breeding trials (e.g., spatial gradients in soil properties) are minimized in the phenotypic correction step by separating useful genetic patterns and experimental noises (non-genetic patterns). Nevertheless, those phenotypic records carry the indissoluble effects of macro-environmental fluctuations of certain weather and soil factors that occurred during crop growth and development (Li et al., 2018; Vidotti et al., 2019; Millet et al., 2019; Guo et al., 2020; Jarquín et al., 2020). Because of that, a multiplicative genotype by environment interaction (G×E) source of variation derived from G×E = P - (G+E) emerges generating quantitative changes in trait expression across the environments.

The G×E is mostly a consequence of the macro-environment fluctuations in the lifetime of the crops (Allard and Bradshaw, 1964; Bradshaw, 1965; Arnold et al., 2019), in which the balance between different environmental inputs are responsible sources for modulating the rate of gene expression (e.g., Jończyk et al., 2017; Liu et al., 2020) and fine-tuning epigenetic variations related to transcriptional responses (Vendramin et al., 2020; Cimen et al., 2021). Nevertheless, since 2012, several GP studies tried to deal with G×E in different manners, most of them ignoring the potential uses of enviromics as a source of ecophysiology knowledge. First studies involved the use of marker by environment interaction (M×E) models, such as the environment-specific marker effects (Burgueño et al., 2012) and with specific marker effects across the environments (Schulz-Streeck et al., 2013). Then, the integration of some environmental covariable (Heslot et al., 2014), which can be: (i) "explicit covariable", derived from pedoclimatic conditions or stressful factors; or an (ii) "implicit covariate", derived from matrix algebra decompositions (e.g, factor analytic models). This last has the advantage of proving a better structuration of the variance-covariance matrix for G×E effects, reducing the complexity and noise for providing predictions. However, the explicit covariates have the

advantage of dealing directly with the macro-environment sources of variations, which can be also used to derive additional information about crop adaptation for certain key factors.

To evolve the current GP platforms for the next level, capable to attempt predictions for complex future scenarios (e.g., climate change), the use of explicit covariates in a wise manner must also evolve. Because of that, the classical conept of reaction-norm cames back again to the GP context (Jarquím et al., 2014; Millet et al., 2019; Jarquín et al., 2020; de los Campos et al., 2020; Rogers et al., 2021). Among its implementations, the most intuitive and robust might be the linear kernel method extending the linear GBLUP (Jarquín et al. 2014), while the most biological accurate are the whole-genome regressions of the rection-norms for key factors (e.g., Li et al., 2018; Millet et al., 2019; Guo et al., 2020) and mechanistic crop growth models (Cooper et al., 2016; Messina et al., 2018; Toda et al. 2020; Robert et al., 2020). However, for running all these approaches, we envisage that at least four aspects are needed to be considered: (1) background on crop ecophysiology; (2) the quality and availability of environmental information; (3) crop and trait and; (4) about crop models, the difficult in phenotyping additional ecophysiology traits for training the crop modeling approaches.

The potential importance of enviromics in modern plant breeding has been recently emphasized (Voss-Fels et al., 2019; Bernardo, 2020; Resende et al., 2020; Crossa et al., 2021). Conversely, there is a lack of studies providing a theoretical background on this field, or even an intuitive pipeline, this present thesis interplays enviromic theory and quantitative genetics focused on predictive breeding. Here different investigations were conducted in order to check the merit of multi-environment genomic prediction platforms, with and without some enviromic enrichment, and how this last can be a cost-effective and climate-smart tool for future plant breeding. Thus, the philosophy that guided the conception of this thesis followed five key- hypothesis:

I. does enviromics really improves the prediction ability of multi-environment GP?

II. does the use of enviromics might fill the lack of phenotypic records across drastically sparse multi-environment phenotyping networks?

III. which is the best kernel method to model the environmental relatedness realized from enviromics and its subsequent reaction-norm effects for different genomic effects (e.g., additive and dominance deviations)?

IV. if enviromics is a good alternative to increase the prediction ability of GP, then we might be able to develop an intuitive and easy managed open-source software, with some basic pipeline useful to democratize the use of enviromics by plant breeders and quantitative geneticists?

V. if this software is provided, and it has pieces of evidence that enviromics might supply the lack of phenotypic information for training multi-enviroment GP models, then it is possible to develop a second pipeline that integrates genomic and enviromic tools in order to optimize field testing efforts and perform an early screening of genotypes across "virtual environments"?

All hypotheses were tested using tropical maize hybrids (single-crosses) as proof-of-concept crop. Currently, there is a wide range os modeling approaches and applications of GP in maize breeding, such as to support the selection of parental inbreeds over diverse populations and whole-genome methods for single or multiple environments (e.g., Lorenzana and Bernardo, 2009; Windhausen et al., 2012; Zhang et al., 2017; Cui et al., 2020), the prediction of double-haploid lines using crop growth models with GP across contrasting growing conditions (e.g., Cooper et al., 2016; Messina et al., 2018) and the prediction of the performance of single-crosses using additive and non-additive effects with several types of kernel methods and matrix structurations for G×E effects (e.g., Dias et al., 2018; Costa-Neto et al., 2021; Rogers et al., 2021). Guided by the five hypothesis, in the following chapters we added to this matter some novel uses of genomics and enviromics for optimizing the multi-environment genomic prediction

in maize. The first chapter is dedicated to the development of the first envirotyping pipeline, and a subsequent open-source software, capable to support the interplay between enviromics and quantitative genomics. This was focused on enviromics in plant breeding, were is discussed some key aspects relevant to better use of environmental information in an ecophysiology-smart manner. The second chapter deals with the study of which kernel method is more suitable for modeling nonadditive effects (e.g., dominance and dominance-based reaction-norms), focused on tropical maize hybrids. Finally, the third chapter deals with a conceptual study involving the development and application of a novel approach for genomic prediction – the so-called "enviromic-aided genomic prediction" (E-GP). This approach involves the use of "markers of environmental typologies" (shortly referred to as envirotype markers) over two prediction scenarios and integrating selective phenotyping approaches. At the end of the third chapter, we show how to connect a genotyping pipeline of parental inbreeds with enviromic sources in order to design super-optimized field trials for training GP, with a focus on predicting novel G×E scenarios.

REFERENCES

Allard, R. W., and Bradshaw, A. D. (1964). Implications of genotype-environmental interactions in applied plant breeding. *Crop Sci.* 4, 503–508

Antolin, L. A. S., and Heinemann, A. B. (2021). Impact assessment of common bean availability in Brazil under climate change scenarios. 191. doi:10.1016/j.agsy.2021.103174.

Arnold, P. A., Kruuk, L. E. B., and Nicotra, A. B. (2019). How to analyse plant phenotypic plasticity in response to a changing climate. *New Phytol.* 222, 1235–1241. doi:10.1111/nph.15656.

Bernardo, R. (2020). Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. *Heredity (Edinb).* doi:10.1038/s41437-020-0312-1.

Bradshaw, A. D. (1965). Evolutionary significance of phenotypic plasticity in plants. *Adv. Genet.* 13, 115–155.

Burgueño, J., de los Campos, G., Weigel, K., and Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Sci.* 52, 707–719. doi:10.2135/cropsci2011.06.0299.

Cimen, E., Jensen, S. E., and Buckler, E. S. (2021). Building a tRNA thermometer to estimate microbial adaptation to temperature. *Nucleic Acids Res.* 48, 12004–12015. doi:10.1093/nar/gkaa1030.

Cinta Romay, M., Malvar, R. A., Campo, L., Alvarez, A., Moreno-González, J., Ordás, A., et al. (2010). Climatic and genotypic effects for grain yield in maize under stress conditions. *Crop Sci.* 50, 51–58. doi:10.2135/cropsci2008.12.0695.

Cooper, M., Technow, F., Messina, C., Gho, C., and Radu Totir, L. (2016). Use of crop growth models with whole-genome prediction: Application to a maize multienvironment trial. *Crop Sci.* 56, 2141–2156. doi:10.2135/cropsci2015.08.0512.

Cooper, M., Messina, C. D., Podlich, D., Totir, L. R., Baumgarten, A., and Hausmann, N. J. (2014). Predicting the future of plant breeding: Complementing empirical evaluation with genetic prediction. *Crop Pasture Sci* 65.

Costa-Neto, G. M. F., Morais Júnior, O. P., Heinemann, A. B., de Castro, A. P., and Duarte, J. B. (2020). A novel GIS-based tool to reveal spatial trends in reaction norm: upland rice case study. *Euphytica* 216, 1–16. doi:10.1007/s10681-020-2573-4.

Costa-Neto, G., Fritsche-Neto, R., and Crossa, J. (2021). Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. *Heredity (Edinb).* 126, 92–106. doi:10.1038/s41437-020-00353-1.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends Plant Sci.* 22, 961–975. doi:10.1016/j.tplants.2017.08.011.

Crossa, J., Fritsche-Neto, R., Montesinos-lopez, O. A., Costa-Neto, G., Dreisigacker, S., Montesinos-lopez, A., et al. (2021). The Modern Plant Breeding Triangle : Optimizing the Use of Genomics , Phenomics , and Enviromics Data. *Front. Plant Sci.* 12, 1–6. doi:10.3389/fpls.2021.651480.

Cui, Z., Dong, H., Zhang, A., Ruan, Y., He, Y., and Zhang, Z. (2020). Assessment of the potential for genomic selection to improve husk traits in maize. *G3 Genes, Genomes, Genet.* 10, 3741–3749. doi:10.1534/g3.120.401600.

de los Campos, G., Pérez-Rodríguez, P., Bogard, M., Gouache, D., and Crossa, J. (2020). A data-driven simulation platform to predict cultivars' performances under uncertain weather conditions. *Nat. Commun.* 11. doi:10.1038/s41467-020-18480-y.

Dias, K. O. D. G., Gezan, S. A., Guimarães, C. T., Nazarian, A., Da Costa E Silva, L., Parentoni, S. N., et al. (2018). Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. *Heredity (Edinb).* 121, 24–37. doi:10.1038/s41437-018-0053-6.

Epinat-Le Signor, C., Dousse, S., Lorgeou, J., Denis, J.-B., Bonhomme, R., Carolo, P., et al. (2001). Interpretation of genotype x environment interactions for early maize hybrids over 12 years. *Crop Sci.* 41, 663–669. doi:10.2135/cropsci2001.413663x.

Freeman, G. H., and Perkins, J. M. (1971). Environmental and genotype–environmental components of variability: VIII - Relations between genotypes grown in different environments and measures of these environments. *Heredity (Edinb).* 27, 15–23.

Guo, T., Mu, Q., Wang, J., Vanous, A. E., Onogi, A., Iwata, H., et al. (2020). Dynamic effects of interacting genes underlying rice flowering-time phenotypic plasticity and global adaptation. *Genome Res.* 30, 673–683. doi:10.1101/gr.255703.119.

Heinemann, A. B., Ramirez-Villegas, J., Rebolledo, M. C., Costa Neto, G. M. F., and Castro, A. P. (2019). Upland rice breeding led to increased drought sensitivity in Brazil. *F. Crop. Res.* 231, 57–67. doi:10.1016/j.fcr.2018.11.009.

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, 463–480.

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., et al. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet.* 127, 595–607. doi:10.1007/s00122-013-2243-1.

Jarquin, D., Kajiya-Kanegae, H., Taishen, C., Yabe, S., Persa, R., Yu, J., et al. (2020). Coupling day length data and genomic prediction tools for predicting time-related traits under complex scenarios. *Sci. Rep.* 10, 1–12. doi:10.1038/s41598-020-70267-9.

Jończyk, M., Sobkowiak, A., Trzcinska-Danielewicz, J., Skoneczny, M., Solecka, D., Fronk, J., et al. (2017). Global analysis of gene expression in maize leaves treated with low temperature. II. Combined effect of severe cold (8 °C) and circadian rhythm. *Plant Mol. Biol.* 95, 279–302. doi:10.1007/s11103-017-0651-3.

Lorenzana, R. E., and Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theor. Appl. Genet.* 120, 151–161. doi:10.1007/s00122-009-1166-3.

Li, X., Guo, T., Mu, Q., Li, X., and Yu, J. (2018). Genomic and environmental determinants and their interplay underlying phenotypic plasticity. *Proc. Natl. Acad. Sci. U. S. A.* 115, 6679–6684. doi:10.1073/pnas.1718326115.

Liu, S., Li, C., Wang, H., Wang, S., Yang, S., Liu, X., et al. (2020). Mapping regulatory variants controlling gene expression in drought response and tolerance in maize. *Genome Biol.* 21, 1–22. doi:10.1186/s13059-020-02069-1.

Messina, C. D., Technow, F., Tang, T., Totir, R., Gho, C., and Cooper, M. (2018). Leveraging biological insight and environmental variation to improve phenotypic prediction: Integrating crop growth models (CGM) with whole genome prediction (WGP). *Eur. J. Agron.* 100, 151–162.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157.

Millet, E. J., Kruijer, W., Coupel-Ledru, A., Alvarez Prado, S., Cabrera-Bosquet, L., Lacube, S., et al. (2019). Genomic prediction of maize yield across European environmental conditions. *Nat. Genet.* 51, pages952–956. doi:10.1038/s41588-019-0414-y.

Resende, R. T., Piepho, H. P., Rosa, G. J. M., Silva-Junior, O. B., e Silva, F. F., de Resende, M. D. V., et al. (2020). Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. *Theor. Appl. Genet.* doi:10.1007/s00122-020-03684-z.

Robert, P., Le Gouis, J., and Rincent, R. (2020). Combining Crop Growth Modeling With Trait-Assisted Prediction Improved the Prediction of Genotype by Environment Interactions. *Front. Plant Sci.* 11, 1–11. doi:10.3389/fpls.2020.00827.

Rogers, A. R., Dunne, J. C., Romay, C., Bohn, M., Buckler, E. S., Ciampitti, I. A., et al. (2021). The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. *G3 Genes|Genomes|Genetics* 11. doi:10.1093/g3journal/jkaa050.

Schulz-Streeck, T., Ogutu, J. O., Gordillo, A., Karaman, Z., Knaak, C., and Piepho, H. P. (2013). Genomic selection allowing for marker-by-environment interaction. *Plant Breed.* 132, 532–538. doi:10.1111/pbr.12105.

Toda, Y., Wakatsuki, H., Aoike, T., Kajiya-Kanegae, H., Yamasaki, M., Yoshioka, T., et al. (2020). Predicting biomass of rice with intermediate traits: Modeling method combining crop growth models and genomic prediction models. *PLoS One* 15, 1–21. doi:10.1371/journal.pone.0233951.

Vendramin, S., Huang, J., Crisp, P. A., Madzima, T. F., and McGinnis, K. M. (2020). Epigenetic regulation of ABA-induced transcriptional responses in maize. *G3 Genes, Genomes, Genet.* 10, 1727–1743. doi:10.1534/g3.119.400993.

Vidotti, M. S., Matias, F. I., Alves, F. C., Rodríguez, P. P., Beltran, G. A., Burguenõ, J., et al. (2019). Maize responsiveness to Azospirillum brasilense: Insights into genetic control, heterosis and genomic prediction. *PLoS One* 14, 1–22. doi:10.1371/journal.pone.0217571.

Voss-Fels, K. P., Cooper, M., and Hayes, B. J. (2019). Accelerating crop genetic gains with genomic selection. *Theor. Appl. Genet.* 132, 669–686. doi:10.1007/s00122-018-3270-8.

Windhausen, V. S., Atlin, G. N., Hickey, J. M., Crossa, J., Jannink, J.-. L., and Sorrells, M. E. (2012). Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *Genes|Genomes|Genet* 2.

Wood, J. T. (1976). The use of environmental variables in the interpretation of genotype-environment interaction. *Heredity (Edinb).* 37, 1–7. doi:10.1038/hdy.1976.61.

Xu, Y. (2016). Envirotyping for deciphering environmental impacts on crop plants. *Theor. Appl. Genet.* 129, 653–673. doi:10.1007/s00122-016-2691-5.

Zhao, Y., Gowda, M., Liu, W., Würschum, T., Maurer, H. P., Longin, F. H., et al. (2012). Accuracy of genomic selection in European maize elite breeding populations. *Theor. Appl. Genet.* 124, 769–776. doi:10.1007/s00122-011-1745-y.

Zhang, X., Pérez-Rodríguez, P., Semagn, K., Beyene, Y., Babu, R., López-Cruz, M. A., et al. (2015). Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. *Heredity (Edinb).* 114, 291–299. doi:10.1038/hdy.2014.99.

Zhang, X., Pérez-Rodríguez, P., Burgueño, J., Olsen, M., Buckler, E., Atlin, G., et al. (2017). Rapid Cycling Genomic Selection in a Multiparental Tropical Maize Population. *G3* 7, 2315–2326. doi:10.1534/g3.117.043141.

2. EnvRtype: A SOFTWARE TO INTERPLAY QUANTITATIVE GENOMICS AND ENVIROMICS IN AGRICULTURE

ABSTRACT

Envirotyping is a technique used to unfold the non-genetic drivers associated with the phenotypic adaptation of living organisms. Here we introduce the EnvRtype R package, a novel toolkit developed to interplay large-scale envirotyping data (enviromics) into quantitative genomics. To start a user-friendly envirotyping pipeline, this package offers: (1) remote sensing tools for collecting (get_weather and extract_GIS functions) and processing ecophysiological variables (processWTH function) from raw environmental data at single locations or worldwide; (2) environmental characterization by typing environments and profiling descriptors of environmental quality (env_typing function), in addition to gathering environmental covariables as quantitative descriptors for predictive purposes (W_matrix function); and (3) identification of environmental similarity that can be used as an enviromic-based kernel (env_typing function) in whole-genome prediction (GP), aimed at increasing ecophysiological knowledge in genomic best-unbiased predictions (GBLUP) and emulating reaction norm effects (get_kernel and kernel_model functions). We highlight literature mining concepts in fine-tuning envirotyping parameters for each plant species and target growing environments. We show that envirotyping for predictive breeding collects raw data and processes it in an eco-physiologically-smart way. Examples of its use for creating global-scale envirotyping networks and integrating reaction-norm modeling in GP are also outlined. We conclude that EnvRtype provides a cost-effective envirotyping pipeline capable of providing high quality enviromic data for a diverse set of genomic-based studies, especially for increasing accuracy in GP across untested growing environments.

Keywords: G×E: genotype × environment interaction; Envirotyping; Environmental characterization

Published at G3: Genes|Genetics|Genomes 2021, 11(4), jkab040, as DOI: 10.1093/g3journal/jkab040

2.1. INTRODUCTION

Quantitative genetics divides phenotypic variation (P) into a genetic (G) and non-genetic source of variation (E). The latter may involve micro-environmental effects that can be controlled by adequate experimental designs, spatial analysis and phenotype correction strategies (e.g., Resende and Duarte, 2007; Galli et al., 2018). Conversely, most non-genetic sources are due to macro-environmental fluctuations resulting from resource availability during crop lifetime (Shelford, 1931). Despite this unfolded division, the effect of the environment on shaping gene expression (e.g., Plessis et al., 2015; Jończyk et al., 2017; Liu et al., 2020) and fine-tuning epigenetic factors (Varotto et al., 2020; Vendramin et al., 2020) creates an indissoluble envirotype-phenotype covariance in the phenotypic records (Lynch and Walsh, 1998). Thus, for any genotype-phenotype association study across multiple environments (e.g., mapping quantitative trait loci, QLT; genomic association studies, GWAS), there is a strong non-genetic influence that can be better understood using envirotyping-based data, i.e., a foundation of multiple techniques to collect, process, and integrate environmental information in genetic and genomic studies, Costa-Neto et al. (2020a).

Over the last ten years, envirotyping (Xu, 2016) has been incorporated into whole-genome prediction (GP, Meuwissen et al., 2001) aiming to better model genotype × environment interaction (G×E) as a function of reaction-

norm from environmental covariables (ECs), i.e., linearized responsiveness of a certain genotype for a target environmental gradient. Those genomic-related reaction norms can be modeled as genotype-specific coefficients for each EC due to whole-genome factorial regressions (Heslot et al., 2014; Ly et al., 2018; Millet et al., 2019), allowing a deeper understanding of which ECs may better explain the phenotypic plasticity of organisms. Furthermore, ECs can also be used to create envirotyping-based kinships (Jarquín et al., 2014; Morais-Junior et al., 2018; Costa-Neto et al., 2020a), enabling the establishment of putative environmental similarities that may drive a large amount of phenotypic variation. The integration of ecophysiologically enriched envirotyping data has led to outstanding results in modeling crops such as maize, due to the use of Crop Growth Models (Cooper et al., 2016; Messina et al., 2018) and Deep Kernel approaches (Costa-Neto et al., 2021a). Combined with phenotyping and genotyping data, the use of envirotyping data may leverage molecular breeding strategies to understand historical trends and cope with future scenarios of environmental change (Gillberg et al., 2019; de los Campos et al., 2020). Its use can also support other prediction-based pipelines in plant breeding, such as high-throughput phenotyping surveys (Krause et al., 2019; Bustos-Korts et al., 2019; Galli et al., 2020).

Despite advancements in the development of hypotheses supporting the inclusion of envirotyping data in GP, it is difficult for most breeders to deal with the interplay between envirotyping, ecophysiology, and genetics. For example, much research has been conducted to explore and associate data into the concepts and theories underlying quantitative genetics (e.g., Fisher's Infinitesimal Model) with the goal of building genomic relationship matrices (GRM). Genotyping pipelines based on bioinformatics were successfully developed to translate biochemical outputs collected from plant tissues into biologically significant markers of DNA polymorphisms, e.g., genotyping-by-sequence (GBS, Elshire et al., 2011). To the best of our knowledge, there is no publicly available user-friendly software to implement envirotyping pipelines to translate raw environmental data into a useful, highly-tailored matrix of envirotypic descriptors. Consequently, a workflow to interplay enviromics (pool of environmental types, abbreviated as envirotypes) and genomic analysis is lacking, especially for GP conditions in multi-environment testing (MET) where G×E might be a noise source of the model's accuracy.

In this study, we introduce EnvRtype, a novel R package used to integrate macro-environmental factors in various fields of plant and animal research or evolutionary ecology. We approach basic ecophysiological concepts underlying the collection and processing of raw-environmental data, both biologically and statistically. Then, we present the functions for implementing remote data collection and primary processing and its applications for deriving quantitative and qualitative descriptors of relatedness. Finally, we present a comprehensive view of how enviromе-based data can be incorporated into GP for selecting genotypes across diverse environments. We highlight the use of different envirotyping levels to discover descriptors of environmental similarity, using crop species to exemplify the concepts.

2.2. Envirotyping Pipeline

EnvRtype is an R package created for handling envirotyping by ecophysiological concepts in quantitative genetics and genomics for multiple environments. It means that envirotyping is not only a collection of raw environmental data that is used for exploratory or predictive processes but rather a pipeline based on the collection of raw data and their subsequent processing in a manner that makes sense for describing the development of an organism in the target environment using a priori ecophysiological knowledge. Here we consider enviromics as the large-scale envirotyping (Xu, 2018; Resende et al., 2020) of a theoretical population of environments for a target species or

germplasm (the so-called envirome). It may also denote the core of possible growing conditions and technological inputs that create different productivity levels. The envirotyping pipeline implemented by EnvRtype software is divided into three modules briefly described above and detailed in the following sections (Fig. 1).

Module 1 (yellow toolboxes in Fig. 1) starts by collecting raw environmental data from public platforms, such as a satellite-based weather system named 'NASA's Prediction of Worldwide Energy Resources' (NASA POWER, https://power.larc.nasa.gov/), which can access information daily anywhere on earth. This database is well consolidated and validated for use in several research fields, including crop modeling in agricultural research (White et al., 2011; Monteiro et al., 2018; Aboelkhair et al., 2019). Details about resolution and validation of this data source are given in https://power.larc.nasa.gov/docs/methodology/validation/.

Data collection may span existing experimental trials (single sampling trials) or historical trends for a given location × planting date arrangement. This module gathers the functions for remote data collection of daily weather and elevation data, as well as the computation of ecophysiological variables, such as the effect of air temperature on radiation use efficiency. The module includes a toolbox with 'Remote Data Collection' and 'Data Processing' steps, both designed to help researchers find a viable alternative for expensive in-field environmental sensing equipment. More details about the theoretical basis of environmental sensing and the module are given in the section "Module 1: Remote Environmental Sensing".

The processed environmental information can then be used for many purposes. In Module 2, we designed tools for the characterization of the macro-environmental variations, which can also be done across different time intervals of crop growth and development (when associated with a crop) or fixed time intervals (to characterize locations). The environmental characterization toolbox (green toolbox in Fig. 1) involves two types of profiling:

1) Discovering environmental types (envirotypes, hereafter abbreviated as ETs) and how frequently they occur at each growing environment (location, planting date, year). Based on the ET-discovering step, it is possible to create environmental profiles and group environments with the same ET pattern. This step is also useful for running exploratory analysis, e.g., to discover the main ET of planting dates at a target location.

2) Gathering environmental covariables (hereafter abbreviated as ECs) from point-estimates (e.g., mean air temperature, cumulative rainfall). These ECs can be used for many purposes, from a basic interpretation of G×E to estimating gene-environment interactions. At the end of this process, a matrix of ECs (**W**) is created and integrated with tools from Module 3. Further details about this module are given in the section "Module 2: Macro-Environmental Characterization".

Finally, the information from Module 2 can be used to create environmental similarity and integrate robust GP platforms for multiple environments, hereafter referred to as envirotype-informed GPs. Module 3 (the dark purple boxes Figure 1) aims to provide tools to compute environmental similarity using correlations or Euclidean distances across different trials conducted on ECs. Thus, we developed a function to integrate this enviromic source in GP as an additional source of variation to bridge the gap between genomic and phenotypic variation. For that, we provide at least four different structures into a flexible platform to integrate multiple genomic and enviromic kinships.

Figure 1 shows some possible outputs of the EnvRtype package (in red toolbox colors), in which **W** can be used to interpret G×E (e.g., factorial regression) or exploited in terms of increasing the accuracy of phenotype prediction for multiple environments. More details are given in the section "MODULE 3: Enviromic Similarity and Phenotype Prediction". Below we give some theoretical details about each module and a description of the functions used to implement it.

2.3. Software

The R package EnvRtype is available at https://github.com/allogamous/EnvRtype [verified 25th May 2021]. More details about graphical plots and additional codes can also be found on this Git Hub webpage and in the published form of the package (Costa-Neto et al., 2021b). All code boxes (BOX from 1 to 19) are detailed in the Supplementary Files. For installing the package, see BOX 1.

2.4. Data sets

EnvRtype has a toy data set for running examples, mostly involving genomic prediction (see Module 3). This data set was included in Bandeira e Souza et al. (2017) and Cuevas et al. (2019) and came from the Helix Seed Company (HEL). However, to facilitate the demonstration of functions, we provided a subset of 150 hybrids per environment (see BOX 2). Grain yield data are mean-centered and scaled (MaizeYield object). The genotyping relationship for additive effects is based on 52,811 SNPs available to make the predictions (maizeG object). The phenotypic and genomic data are credited to Helix Seed Ltda. Company. Finally, weather data are presented for each of the five environments (maizeWTH object). Additional tutorials can be found at Git Hub (https://github.com/allogamous/EnvRtype).

2.5. MODULE 1: Remote Environmental Sensing

2.5.1. Remote data collection

EnvRtype implements the remote collection of daily weather and elevation data by the *get_weather* function. This function has the following arguments: the environment name (env.id); geographic coordinates (latitude, lat; longitude, lon) in WGS84; time interval (start.day and end.day, given in 'year-month-day'); and country identification (country), which sets the raster file of elevation for the region of a specific country. Countries are specified by their 3-letter ISO codes (check in the package Git Hub or use the getData('ISO3') function from the raster package to see the codes).

Table 1 shows the names of the outputs of get_weather() and processWTH() (see Tools for basic processing). All weather information is given on a daily basis. Altitude (ALT) information is provided by SRTM 90 m resolution and can be collected from any place between -60 and 60 latitudes. This information is presented as a data.frame class output in R. It is possible to download data for several environments by country.

A practical example of get_weather is given below (BOX 3). In this example, we run a collection of environmental data for Nairobi, Kenya (latitude 1.367 N, longitude 36.834 E) from 01 march 2015 to 01 April 2015. A second function is extract_GIS, which can collect point values from large raster files from GIS databases. This function has six arguments. The object env.data indicates the name of the environmental dataset (arranged as a data.frame). It can be an output data.frame of the get_weather() function or any spreadsheet of environmental data, as long as it is organized with a column denoting the environment's name, which is defined by the env.id argument (default is env.id = 'env'). Latitude and Longitude is provided in the same manner described in get_weather().

Finally, the name.out is the argument to define the name of the collected covariable (e.g., ALT for altitude). The function extract_GIS can be useful for collecting covariables from raster files within databases such as WorldClim (Fick et al., 2017; https://www.worldclim.org/), SoilGrids (https://soilgrids.org/), EarthMaps (https://earthmap.org/) and Nasa Power (https://power.larc.nasa.gov).

A practical use of extract_GIS() is given below (BOX 4). In this example, a collection of clay content (g/kg) from 5cm to 15cm of depth for Nairobi using a raster file was downloaded from SoilGrids and the function extract_GIS(). The raster file 'clay_5_15' can be accessed in R by typing data("clay_5_15").

2.6. Summarizing raw-data

A basic data summary of the outputs from the get_weather function is done by the summaryWTH() function. This function has 10 arguments (env.data, id.names, env.id, days.id, var.id, statistic, probs, by.interval, time.window, and names.window). The common arguments with extract_GIS have the same described utility. Other identification columns (year, location, management, responsible researcher, etc.) may be indicated in the id.names argument, e.g., id.names = c('year','location','treatment').

Considering a specific environmental variable, the argument var.id can be used as, for example, var.id = 'T2M'. By default, this function considers all the names of the variables presented in Table 1. For other data sources, such as micro-station outputs, this argument is necessary for identifying which variables will be summarized. The argument days.id indicates the variable pointing to the time (days), where the default is the daysFromStart column from the get_weather function. A basic example of this use is given below (BOX 4):

Dividing the development cycle into time intervals (e.g., phenology), whether phenological or fixed time intervals (e.g., 10-day intervals), helps to understand the temporal variation of environmental factors during the crop growth cycle. Thus, specific time intervals can be created by the time.window argument. For example, time.window = c(0,14,35,60,90,120) denotes intervals of 0-14 days from the first day on record (0). If the first record denotes the crop's emergence date in the field, this can also be associated with some phenological interval. Those intervals can be named using the argument names.window, names.window = c('P-E','E-V1','V1-V4','V4-VT','VT-GF','GF-PM').

The argument statistic denotes which statistic should be used to summarize the data. The statistic can be: mean, sum or quantile. By default, all statistics are used. If statistic = 'quantile', the argument prob is useful to indicate which percentiles (from 0 to 1) will be collected from the data distribution, i.e., default is prob = c(0.25, 0.50, 0.75), denoting the first (25%) second (50%, median) and third (75%) quantiles.

2.7. Tools for basic data processing

The processWTH() function performs basic data processing. As described for summaryWTH(), this function can also process environmental data for get_weather outputs and other sources (micro-stations, in-field sensors) using the same identification arguments (env.data, id.names, env.id, days.id, var.id). This function also gathers three other sub-functions created to compute general variables related to ecophysiological processes, such as the macro effects of soil-plant-atmosphere dynamics and atmospheric temperature on crop development. The basic usage of this package is given by processWTH(env.data = env.data); in addition, crop-specific parameters such as cardinal values of

temperature and evapotranspiration, as well as site-specific characteristics, can be given in additional arguments. Below, we describe these arguments over three functions that compose processWTH(). Because of its importance for democratizing the EnvRtype pipeline, we first provide a brief description of them, in addition to the ecophysiological concepts underlying their application.

2.7.1. Radiation-related covariables

The radiation balance in crop systems is regulated by the difference between the amount of incident radiation, absorbed energy by the plants and soil surface, and the converted thermal energy. From Nasa Power, the radiation outputs are given in terms of Top-of-atmosphere Insolation (ALLSKY_TOA_SW_DWN), Insolation Incident on a Horizontal Surface (Shortwave, ALLSKY_SFC_SW_DWN), and Downward Thermal Infrared Radiative Flux (Longwave, ALLSKY_SFC_LW_DW). Thus, the net solar radiation available for the physiological process of growth (biomass production) is given by the difference between longwave and shortwave, i.e., SRAD=ALLSKY_SFC_LW_DW-ALLSKY_SFC_SW_DWN, in MJ m-2 d-1. It is possible to download more solar-related parameters directly from Nasa Power website, (https://power.larc.nasa.gov/data-access-viewer/).

In most growth modeling approaches, the effect of radiation use efficiency (RUE) is the main target to describe the relationship between the available energy in the environment and how the plants translate it into biomass (see subsection about thermal parameters). In this context, this environmental variation source is important to understand the differences in potential yield observed in genotypes evaluated across diverse environments. Radiation is also vital as a source for regulating the available energy for other biophysical processes, such as evaporation, transpiration, and temperature (see subsection Temperature-related covariables).

Thus, EnvRtype made a function called param_radiation() available to compute additional radiation-based variables that can be useful for plant breeders and researchers in several fields of agricultural research (e.g., agrometeorology). These parameters include the actual duration of sunshine hours (n, in hours) and total daylength (N, in hours), both estimated according to the altitude and latitude of the site, time of year (Julian day, from 1 to 365), and cloudiness (for n). In addition, the global solar radiation incidence (SRAD, in MJ m^2 d^{-1}) is computed as described at the beginning of this section. The latter is important in most computations of crop evapotranspiration (Allen et al., 1998) and biomass production (Muchow et al., 1990; Muchow and Sinclair, 1991). More details about those equations are given in ecophysiology and evapotranspiration literature (Allen et al., 1998; Soltani and Sinclair, 2012).

The arguments of param_radiation are: env.data and merge, in which merge denotes if the computed radiation parameters must be merged with the env.data set (merge = TRUE, by default).

2.7.2. Temperature-related covariables

Thermal variables are essential for regulating the rates of critical biochemical processes within an organism. At the cell level, the effect of temperature may regulate the rate of enzymatic reactions, in which critical values may lead to denaturation of those enzymes and the death of the cell. At the plant level, temperature-related variables regulate the balance between photosynthesis (gross and net) and respiration in the canopy, impacting radiation use efficiency (RUE). It is also related to the transpiration rates and, consequently, to the absorption of nutrients from water flux in the roots. At the reproductive stages, temperature affects the efficiency of pollination, which is directly related to the

crop's final yield, especially for species in which grain yield is the main target trait. Phenology development rates are also strongly influenced by temperature (e.g., growing degree-days, GDD), in which the balance between biomass accumulation and acceleration of the crop cycle may compromise the source: sink relations and then the final yield. Finally, the dew point (T2MDEW) is another agrometeorological factor that is greatly important for crop health. This factor determines the establishment of diseases (especially fungus) under the leaf to being related to the evaporation process in the stomata.

EnvRtype provides the param_temperature function, which computes additional thermal-related parameters, such as GDD and FRUE, and T2M_RANGE. As previously described, the first is useful to predict phenological development, while the second is an ecophysiology parameter used to quantify the impact of temperature on crop growth and biomass accumulation in crop models (Soltani and Sinclar, 2012). Thus, both can be useful to relate how temperature variations shape some species' adaptation in the target environment. GDD is also important for modeling plant-pathogen interactions because some pests and diseases have temperature-regulated growth. Finally, the daily temperature range (T2M_RANGE) impacts processes such as floral abortion in crops where the main traits are related to grain production. For more details about the impact of temperature on diverse crops, please check Luo (2011).

Thus, the param_temperature() function has eight arguments (env.data, Tmax, Tmin, Tbase1, Tbase2, Topt1, Topt2 and merge). To run this function with data sources other than get_weather(), it is necessary to indicate which columns denote maximum air temperature (Tmax, default is Tmax = 'T2M_MAX') and minimum air temperature (Tmin, default is Tmin = 'T2M_MIN') (BOX 6). The cardinal temperatures must follow the processes provided in the previously described ecophysiology literature (Soltani and Sinclar, 2012). Consider the estimations for dry beans at the same location in Nairobi, Kenya as used in the other BOX examples.

2.7.3. Atmospheric demands

Atmospheric demands are shaped by the dynamics of precipitation (rainfall) and water demand (evaporation+plant transpiration). Thus, both are regulated as a consequence of the balance of radiation and thermal-related processes in the atmosphere (Soltani and Sinclair, 2012; Allen et al., 1998). The soil-plant-atmosphere continuum involves water dynamics from the soil, passing through plant tissues, and going back to the atmosphere through the stomata. This process's rate is deeply related to the biomass production of plants and the absorption of nutrients by the mass flux in roots. Because of that, water demands are essential for measuring the quality of some growing environments.

Here we used the Priestley-Taylor equation to compute the reference crop evapotranspiration. With this equation, the empirical constant (alpha = α) may range from 1 (at humidity conditions) to 2 (at arid conditions). First, we compute the vapor pressure, determined by $e_a = RH \times e_s$ (Dingman, 2002), where e_s is the saturation vapor pressure defined as (Buck, 1981):

$$e_s = [1.007 + (3.46 \times 10^{-5} \times P)] \times 6.1121 \times exp\left(\frac{17.502 \times Tavg}{240.97 + Tavg}\right)$$

where Tavg is the average air temperature, and P is the air pressure (kPa) computed from elevation as $P = 101.3 \times (293 - 0.0065 \times ALT/293)^{5.26}$. Thus, from the daily vapor pressure (e_a), we compute the slope of the saturation vapor pressure curve (Δ), by (Dingman, 2002):

$$\Delta = \frac{4098 \times e_s}{(Tavg + 237.2)^2}$$

Finally, the reference evapotranspiration (ET_0) is computed as:

$$ET_0 = \alpha \frac{\Delta \times (R_n - G)}{\lambda_v \times (\Delta + \Upsilon)}$$

where λ_v is the volumetric latent heat of vaporization (2453 MJ m^{-3}) and Υ is the psychometric constant (kPa C^{-1}), that can be computed from air pressure as $\Upsilon = 0.665 \times 10^{-3} P$ (Allen et al., 1998). For crops, we encourage the use of crop coefficients (K_c, dimensionless) to translate ET_0 in crop-specific evapotranspiration. This K_c is computed from empirical phenotypic records (crop height, the albedo of the soil-crop surface, canopy resistance) combined with in-field sensors (evaporation from the soil) or using K_c estimates for each crop species. Allen et al. (1998) provide a wide number of general K_c values to be used in this sense. For a complete understanding of soil-water dynamics, we suggest using pedotransfer functions to derive some hydraulic properties of the soil, such as infiltration rate and water retention parameters. It can be done by soil samples or from remotely collected data from SoilGrids using extract_GIS() function.

We implemented the param_atmospheric() function to run basic computations of atmospheric demands. This function has 11 arguments: env.data; PREC (rainfall precipitation in mm, default is PREC='PRECTOT'); Tdew (dew point temperature in °C, default is Tdew='T2M_DEW'); Tmax (maximum air temperature in °C, default is Tmax='T2M_MAX'); Tmin (minimum air temperature in °C, default is Tmin='T2M_MIN'); RH (relative air humidity %, default is RH='RH2M'); Rad (net radiation, in MJ m^{-2} day^{-1}, default is Rad ='Srad'); alpha (empirical constant accounting for vapor deficit and canopy resistance values, default is alpha=1.26); Alt (altitude, in meters above sea level, default is Alt = ALT); G (soil heat flux in W m^{-2}, default is G=0); and merge (default is merge='TRUE). In BOX 7 we present an example of usage in Nairobi, Kenya. Consider the same env.data collected in the previous box and an elevation value of Alt = 1,628.

2.8. MODULE 2 - Macro-Environmental Characterization

2.8.1. Discovering Envirotypes with env_typing

An environment can be viewed as the status of multiple resource inputs (e.g., water, radiation, nutrients) across a certain time interval (e.g., from sowing to harvesting) within a specific space or location. The quality of those environments is an end-result of the daily balance of resource availability, which can be described as a function of how many resources are available and how frequently those resources occur (e.g., transitory or constant effects). Also, the relationship between resource absorption and allocation depends on plant characteristics (e.g., phenology, current health status). Then, this particular environmental-plant influence is named after the envirotype to differentiate it from the concept of raw environmental data (data collected directly from sensors). It can be referred to as environmental type (ET). Finally, the typing of environments can be done by discovering ETs; the similarity among environments is a consequence of the number of ETs shared between environments.

Before the computation of ETs, a first step was to develop a design based on ecophysiological concepts (e.g., plants' needs for some resource) or summarize the raw data from the core environments being analyzed. Then, for each ET we computed the frequency of occurrence, which represents the frequency of specific quantities of resources available for plant development. Typing by frequency of occurrence provides a deeper understanding of the distribution of events, such as rainfall distribution across different growing cycles and the occurrence of heat stress conditions in a target location (Heinemann et al. 2015). Thus, groups of environments can be better identified by analyzing the events occurring in a target location, year, or planting date. This step can be done not only by using grade point averages (e.g., accumulated sums or means for specific periods), but also by their historical similarity. In this way, we are able to not only group environments in the same year, but also through a historical series of years. Finally, this analysis deepens in resolution when the same environment is divided by time intervals, which can be fixed (e.g., 10-day intervals), or categorized by specific phenological stages of a specific crop.

To implement envirotype profiling, we created the env_typing() function. This function computes the frequency of occurrence of each envirotype across diverse environments. This function has 12 arguments, nine of which (env.data, id.names, env.id, days.id var.id, statistic, by.interval, time.window, and names.window) work in the same way as already described in the previous functions. The argument cardinals are responsible for defining the biological thresholds between envirotypes and adaptation zones. These cardinals must respect the ecophysiological limits of each crop, germplasm, or region. For that, we suggest literature on ecophysiology and crop growth modeling, such as Soltani and Sinclar (2012). The argument cardinals can be filled out as vectors (for single-environmental factors) or as a list of vectors for each environmental factor considered in the analysis. For example, considering the cardinals for air temperature in rainfed rice presented in Table 2, the cardinals are typed for Los Baños, Philippines, from 2000 to 2020, as (see BOX 8):

If cardinals = NULL, the quantiles 10%, 25%, 50%, 75% and 90% are used by default. Which quantiles will be used is determined in the same manner as prob (in summaryWTH), but now using the quantile argument, e.g., quantile = c(0.25,0.50,0.75).

For multiple environmental factors, a list of cardinals must be provided—for example, considering rainfall precipitation (PRECTOT, $mm.day^{-1}$) and dew point temperature (T2DEW, $°C.day^{-1}$). Suppose precipitation values less than 10 $mm.day^{-1}$ are insufficient to meet the studied crops' demand. Values between 11 $mm.day^{-1}$ and 40 $mm.day^{-1}$ would be considered excellent water conditions, and values greater than 40 $mm.day^{-1}$ would be considered excessive rainfall. In this scenario, such rainfall values could be negatively associated with flooding of the soil and drainage of fertilizers, among other factors related to crop lodging or disease occurrence. Thus, for PRECTOT, the cardinals will be cardinals = c(0,5,10,25,40,100). For dew point, let's assume data-driven typing (cardinals = NULL) using the previously described quantiles. Taking the same example for Los Baños, Philippines (BOX 9).

2.8.2. Environmental Covariables with W_matrix

The quality of an environment is measured by the amount of resources available to fulfill the plants' demands. In an experimental network composed of multi-environment trials (MET), the environment's quality is relative to the global environmental gradient. Finlay and Wilkinson (1963) proposed using phenotypic data as a quality index over an implicit environmental gradient. However, this implicit environmental quality index was proposed as an

alternative to explicit environmental factors, given the difficulties in obtaining high-quality envirotyping data. Here we provide the use of detailed environmental data arranged in a quantitative descriptor such as a covariate matrix (W), following the terminology used by Costa-Neto et al. (2021a) and de los Campos et al. (2020). Based on these W matrices, several analyses can be performed, such as (1) dissecting the G×E interaction; (2) modeling genotype-specific sensibility to critical environmental factors; (3) dissecting the environmental factors of QTL×E interaction; (4) integrating environmental data to model the gene × environment reaction-norm; (5) providing a basic summary of the environmental gradient in an experimental network; (6) producing environmental relationship matrices for genomic prediction.

To implement these applications, the processed environmental data must be translated into quantitative descriptors by summarizing cumulative means, sums, or quantiles, such as in summaryWTH(). However, these data must be mean-centered and scaled to assume a normal distribution and avoid variations due to differences in scale dimensions. To create environmental similarity kernels, Costa-Neto et al. (2021a) suggested using quantile statistics to better describe each variable's distribution across the experimental network. Thus, this allows a statistical approximation of the environmental variables' ecophysiological importance during crop growth and development. In this context, we developed the W_matrix() function to create a double-entry table (environments/sites/years environmental factors). Contrary to env_typing(), the W_matrix() function was designed to sample each environmental factor's quantitative values across different environments.

The same arguments for the functions summaryWTH() and env_typing() are applicable (env.data, id.names, env.id, days.id var.id, statistic, by.interval, time.window, and names.window). However, in W_matrix(), arguments center = TRUE (by default) and scale = TRUE (by default) denote mean-centered ($w - \overline{w}$) and scaled ($w - \overline{w}/\sigma$), in which w is the original variable, $\overline{w}$ and σ are the mean and standard deviation of this covariable across the environments. Quality control (QC = TRUE argument) is done by removing covariables with more than $\sigma_{TOL} \pm \sigma$, where σ_{TOL} is the tolerance limit for standard deviation, settled by default argument as sd.tol = 3.

To exemplify a basic use of W_matrix(), let us consider the maizeWTH object, involving only weather variables temperature, rainfall and precipitation, while assuming a quality control of sd.tol = 4. The time intervals were settled for every ten days (default), and statistic as 'mean' for each variable at each time interval (BOX 10).

2.9. MODULE 3 - Enviromic Similarity and Phenotype Prediction

The prediction of phenotypes across multiple environments can be conducted using different approaches, such as mechanistic crop models and empirical regressions, in which environmental and/or genomic information is necessary for training accurate models. The latter, named after whole-genome prediction (GP, Meuwissen et al., 2001), has revolutionized both plant and animal breeding pipelines around the world. Most approaches rely on increasing the accuracy of modeling genotype-phenotype patterns and exploring them as a predictive breeding tool. Among the several enrichments of computational efficacy and breeding applications, the integration of genomic by environment interaction (G×E) has boosted the ability of genomic-assisted selection to evaluate a wide number of genotypes under several growing conditions over multiple environmental trials (MET).

Heslot et al. (2014) and Jarquín et al. (2014) introduced environmental covariables to model an environmental source of the phenotypic correlation across MET. These approaches aim to model the reaction-norm of genotypes across MET, i.e., how different genotypes react to different environmental gradient variations. In most

cases, reaction-norm modeling serves as an additional source of variation for complementing the genomic relatedness among individuals tested and untested under known environmental conditions. Thus, in addition to the genomic kernels, envirotype-informed kernels can be used to capture macro-environmental relatedness that shapes the phenotypic variation of relatives, the so-called enviromic kernel (Costa-Neto et al., 2021a).

In the third module of the EnvRtype package, we present the tools used to implement this modeling approach. Three main functions were designed for this purpose. First, the function env_kernel can be used for the construction of environmental relationship kernels using environmental information. Second, the get_kernel aims to integrate these kernels into statistical models accounting for different structures capable of explaining the phenotypic variation across MET. Finally, the function kernel_model can be used to fit regression models accounting for environmental and or genomic data using a computationally efficient Bayesian approach. In the following subsections, we describe the kernel methods for modeling envirotype relatedness. Then we present the statistical models that can be built with these kernels.

2.9.1. Enviromic kernels with env_kernel

In this package, we use two types of kernel methods to compute enviromic-based similarity. The first consists of the traditional method based on the linear variance-covariance matrix (Jarquín et al., 2014). This kernel is equivalent to a genomic relationship matrix and can be described mathematically as:

$$\boldsymbol{K}_E = \frac{\boldsymbol{W}\boldsymbol{W}'}{\text{trace}(\boldsymbol{W}\boldsymbol{W}')/nrow(\boldsymbol{W})} \quad \text{(Eq. 1)}$$

where $\boldsymbol{K}_E$ is the enviromic-based kernel for similarity among environments and $\boldsymbol{W}$ matrix of ECs. Note that we use $\boldsymbol{W}$ matrix, but any other source of data from environments can be used here as EC (e.g., typologies, disease evaluations, management).

The second method is a nonlinear kernel modeled by Gaussian processes, commonly called the Gaussian Kernel or GK, and widely used in genomic-enabled prediction (Gianola and van Kaam (2008); de los Campos et al., 2010; Cuevas et al., 2017). The use of GK for modeling $\boldsymbol{K}_E$ was proposed by Costa-Neto et al (2021a) and is described in a similar way to the approach already used for modeling genomic effects:

$$\boldsymbol{K}_E = \exp\left(h\boldsymbol{D}^2_{ii'}/Q\right) \quad \text{(Eq. 2)}$$

where h is the bandwidth factor (assume as h = 1 by default) multiplied by the Euclidean Distance $\boldsymbol{D}^2_{ii'} = \sum_k (w_{ik} - w_{i'k})^2$ for each pairwise elements in the $\boldsymbol{W} = \{w_i, w_{i'}\}$. This means that the environmental similarity is a function of the distance between environments realized by ECs. The scalar variable Q denotes the quantile used to ponder the environmental distance (assumed as Q = 0.5, equal to the median value of $\boldsymbol{D}^2_{ii'}$. The h can be computed using a marginal function described by Pérez-Elizalde et al. (2015).

The env_kernel function implements both methods. It has the following main arguments: env.data, env.id, gaussian, and h.gaussian. The first two arguments work in the same manner previously described for other functions. The gaussian argument (default is gaussian = FALSE) denotes if models (1) or (2) are used to compute $\boldsymbol{K}_E$. If gaussian = TRUE, then the Gaussian kernel (equation 2) is used, and h.gaussian must be inserted to compute it. In the Y argument (default is Y = NULL), it is possible to insert a phenotypic record to be used in the marginal function to compute a data-driven h (Pérez-Elizalde et al., 2015).

The env_kernel function has two outputs called varCov (relatedness among covariables) and envCov (relatedness among environments). The first is useful to deepen the understanding of the relatedness and redundancy of the ECs. The second output is $\boldsymbol{K_E}$. This matrix is the enviromic similarity kernel integrated into the GP models.

A basic use of env_kernel is presented below. Consider the W matrix created in BOX 10 for the maizeWTH object (5 environments in Brazil). The $\boldsymbol{K_E}$. value using linear covariance and the Gaussian kernel is given as (BOX 11):

2.9.2. Phenotype prediction across multiple environments

After constructing the relationship kernels for environmental relatedness, it is possible to fit a vast number of statistical models using several packages available in R CRAN. However, it is important to consider that statistical models containing more complex structures (e.g., more than one genetic effect plus G×E and environmental information) are models that require more expensive computational effort and time. Under Bayesian inference, which demands multiple iterative sampling processes (e.g., via Gibbs sampler) to estimate the variance components, the computational effort may be more expensive. Among the R packages created to run Bayesian linear models for genomic prediction, three main packages may be highlighted: BGLR-Bayesian Generalized Linear Regression (Pérez and de los Campos, 2014), BMTME-Bayesian Multi-Trait Multi-Environment (Montesinos-López et al., 2016) and BGGE-Bayesian Genotype plus Genotype by Environment (Granato et al., 2018). However, BGGE employs an optimization process that can be up to five times faster than BGLR and permits the incorporation more kernel structures than BMTME. For this reason, we implement the kernel_model function that runs the same optimization algorithm for Hierarchical Bayesian Modeling used in BGGE (see Appendix 1).

Below we describe a generic model structure that covers the diversity of possible combinations for modeling the phenotypic variation across MET. This model considers k genomic and l enviromic effects, plus fixed-effects and a random residual variation:

$$\boldsymbol{y} = \boldsymbol{1\mu} + \boldsymbol{X_f\beta} + \sum_{\boldsymbol{s=1}}^{\boldsymbol{k}} \boldsymbol{g_s} + \sum_{\boldsymbol{r=1}}^{\boldsymbol{l}} \boldsymbol{w_r} + \boldsymbol{\varepsilon} \qquad \text{(Eq. 3)}$$

where $\boldsymbol{y}$ is the vector combining the means of each genotype across each one of the q environments in the experimental network, in which $\boldsymbol{y} = \left[\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots \boldsymbol{y_q}\right]^{\boldsymbol{T}}$. The scalar $\boldsymbol{1\mu}$ is the common intercept or the overall mean. The matrix $\boldsymbol{X_f}$ represents the design matrix associated with the vector of fixed effects $\boldsymbol{\beta}$. In some cases, this vector is associated with environmental effects (target as fixed-effect). Random vectors for genomic effects ($\boldsymbol{g_s}$) and enviromic-based effects ($\boldsymbol{w_r}$) are assumed to be independent of other random effects, such as residual variation ($\boldsymbol{\varepsilon}$). It is a generalization for a reaction-norm model because, in some scenarios, the genomic effects may be divided as additive, dominance, and other sources (epistasis) and the genomic by environment (G×E) multiplicative effect. In addition, the envirotyping-informed data can be divided into several environmental kernels and a subsequent genomic by envirotyping (G×W) reaction-norm kernels. Based on Equation 6, the theory underpinning the get_kernel function is summarized in three

Benchmark Genotypic Effects. These are baseline models accounting only for genotype-based effects, mostly associated with pedigree-based or genomic realized kinships. $\sum_{\boldsymbol{s=1}}^{\boldsymbol{p}} \boldsymbol{g_s} \neq 0$ and $\sum_{\boldsymbol{r=1}}^{\boldsymbol{q}} \boldsymbol{w_r} = 0$, in which g_s may be related to main genotype-effect (G) in the case of the main genotype-effect model (MM); and G plus a genotype by

environment deviation (G+G×E), in the case of the so-called MDs model. Note that multiple genotype-relatedness kernels may be incorporated, e.g., for additive (A) and dominance (D) deviations and other sources of information from 'omics.' All genomic kernels must have the p × p dimension, in which p is the number of genotypes. However, this model does not consider any environmental effect.

Enviromic-enriched Main Effects. We added the acronym 'E' to the MM and MDs models to denote 'enviromic-enriched' for EMM and EMDs models. These models consider $\sum_{s=1}^{p} g_s \neq 0$ and $\sum_{r=1}^{q} w_r \neq 0$, in which g_s are related to G (EMM) or G+G×E (EMDs), and w_r are only the main envirotype effects (W). In this type of model, the environmental effects can be modeled as a fixed deviation (using $X_f\beta$) plus a random envirotyping-based variation ($\sum_{r=1}^{q} w_r$).

Enviromic-based Reaction-Norm. We added the acronym 'RN' for 'reaction-norm' to the MM and MDs models, resulting in RNMM and RNMDs models, respectively. As described in (ii), the environmental effects can now be modeled as fixed deviations (using $X_f\beta$) plus a random envirotyping-based variation ($\sum_{r=1}^{q} w_r$). However, those RN models consider $\sum_{s=1}^{p} g_s \neq 0$ and $\sum_{r=1}^{q} w_r \neq 0$, in which g_s are related to G (RNMM) or G+G×E (RNMDs), and w_r are related to main envirotype effects (W) plus an envirotype × genomic interaction (G×W). In this context, RNMM accounts for the variation due to G+W+GW, whereas RNMDS considers G+GE+W+GW.

2.9.3. Getting covariance structures with get_kernel

The get_kernel function has four main arguments: a list of genomic relationship kernels (K_G); a list of environmental relationship kernels (K_E); and a phenotypic MET data set (data), composed of a vector of environment identification (env), a vector of genotype identification (gid), and a vector of trait values (y); at last, the model argument sets the statistical model used ('MM,' 'MDs', 'EMM', 'EMDs', 'RNMM' and 'RNMDs'). Each genomic kernel in K_G must have the dimension of p × p genotypes. This argument assumes K_G = NULL by default. If no structure for genetic effects is provided, the get_kernel function automatically assumes an identity matrix for genotype effects, in which it considers no relatedness among individuals. Finally, the argument stage (stage = NULL by default) states which development stages can be used to create stage-specific enviromic kernels. More detail about the latter is given in the Example 3 of the Results section.

In the same manner, K_E could have the dimension of q × q environments, but the environmental kernels can be built at the phenotypic observation level in some cases. It means that for each genotype in each environment, there is a different EC, according to particular phenology stages or envirotyping at the plant level. Thus, using the additional argument dimension_KE = c('q', 'n') (default is 'q', for environment), the K_E may accomplish a kernel with n × n, in which n = pq. The basic usage of get_kernel() is given in BOX 12 and its detailed applications are provided in the Results section.

2.9.4. Modeling the phenotypic variation with kernel_models

Finally, the kernel_model function has four main arguments: a phenotypic MET data set (data), composed of a vector of environment identification (env), a vector of genotype identification (gid), and a vector of trait values

(y), a list object for random effects (random, from get_kernel) and a matrix for fixed effects (fixed). For the Hierarchical Bayesian Modeling (see Appendix 1), the arguments for number of iterations (iterations, default is 1000) and the number of samples used for burn-in (burnin, default is 200) and thining (thining, default is 10) must be provided. The function has two main outputs: the predicted phenotypes (yHat), variance components for each random effect (VarComp). Below we show a brief example of the use of kernel_model for a MDs model (BOX 13) using the same inputs used in BOX 12.

2.10. Practical Examples

Three practical examples were implemented to present a comprehensive overview of the most important functions of EnvRtype. First, we illustrate the use of EnvRtype for starting an envirotyping pipeline across different locations in the world (Example 1). Second, we used the toy data set (maizeG, maizeWTH and maizeYield) to demonstrate different environmental similarities based on different environmental factors (Example 2). We used two envirotyping levels (per environment and per development stage at each environment) and two ECs (FRUE, PETP and FRUE+PETP) to demonstrate different ways to build environmental relatedness for GP. This type of application can be useful for researchers interested in predicting the individual genotypic responses shaped by genomic and enviromic-specific factors across existing experimental trials or for the assembly of virtual scenarios.

Finally, in Example 3 we ran a genomic prediction study case in maize (maizeG, maizeWTH and maizeYield) involving three models (M1, Baseline Genomic MDs model; M2, Reaction Norm RNMM model and M3, Reaction Norm RNMM considering a different enviromic kinship for each development stage) and two cross-validation schemes (CV1: prediction of novel genotypes, using 20% of the data as a training set; CV00: prediction of novel genotypes at novel environments, using 3 of the 5 environments plus 20% of the genotypes as a training set).

2.11. RESULTS

2.11.1. Example1: Global-scale Envirotyping

To illustrate the use of EnvRtype for a global-scale envirotyping study, we consider different periods (and years) within the summer season at nine locations around the world (BOX 14): Goiânia (Brazil, 16.67 S, 49.25 W, from March 15th, 2020 to April 04th, 2020); Texcoco (Mexico, 19.25 N, 99.50 W, from May 15th, 2019 to June 15th, 2019); Brisbane (Australia, 27.47 S, 153.02 E, from September 15th, 2018 to October 04th, 2018); Montpellier (France, 43.61 N, 3.81 E, from June 18th, 2017 to July 18th, 2017); Los Baños (the Philippines, 14.170 N, 121.431 E, from May 18th, 2017 to June 18th, 2017); Porto-Novo (Benin, 6.294 N, 2.361 E, from July 18th, 2016 to August 18th, 2016), Cali (Colombia, 3.261 N, 76.312 W, from November 18th, 2017 to December 18th, 2017); Palmas (Brazil, 10.168 S, 48.331 W, from December 18th, 2017 to January 18th, 2018); and Davis (the United States, 38.321 N, 121.442 W, from July 18th, 2018 to August 18th, 2018). In this example, we use 'GOI', 'TEX', 'BRI', 'MON', 'LOS', 'PON', 'CAL', 'PAL' and 'DAV' to identify each location (Figure 2A).

From the collected variables, it is possible to type any environmental factor or a core of environmental factors (Figure 2B). As a toy exemplification (BOX 14-15), we used the variable 'T2M' (the daily average temperature at 2 meters) to discover environmental types (ETs) and compute environmental similarity (Figure 2C). In this case, we

used the Gaussian kernel as a sign of environmental distance, but it can also be used as kinship for predictive breeding (Costa-Neto et al., 2021a).

It is possible to see in this toy example, that perhaps locations in different continents might have similar ET trends for air temperature. This process can be done for several variables (single or joint) to better describe those similarities.

2.11.2. Example 2: Modeling Genomic-enabled Reaction-Norm

To illustrate the use of different ECs in modeling genomic-enabled reaction-norms, we ran a toy example involving a tropical maize data set available in EnvRtype (see BOX 2). From equation 3, we assumed the following baseline model: $\boldsymbol{y = 1\mu + X_f\beta + g + gE + \varepsilon}$, where $\boldsymbol{X_f\beta}$ is the fixed environmental effects, $\boldsymbol{g}$ is the random genomic-additive effects and $\boldsymbol{gE}$ is the genomic × environment interaction, modeled by a block diagonal matrix of genomic effects across environments (MDs model). Thus, we added enviromic effects following two envirotyping levels: (1) envirotyping mean values per environment (entire croplife), and (2) envirotyping for each time interval (development stage) across crop life, assuming fixed stages in terms of days after emergence. For each envirotyping level, we considered two types of ECs: the factor of temperature effect over radiation use efficiency (FRUE) and the difference between rainfall precipitation and crop evapotranspiration (PETP) (Figure 3). From equation (3), these matrices of ECs (EC1, EC2, EC3, EC4, EC5 and EC6) were arranged in three kernel structures using the RNMDs model, with the baseline genomic model updated to $\boldsymbol{y = 1\mu + X_f\beta + g + gE + EC + gEC + \varepsilon}$, which resulted in 6 models (M1, M2, M3, M4, M5 and M6) according to each EC matrix used, plus a baseline genomic model (M0).

BOX 16 and BOX 17 presents the codes to implement these envirotyping levels and model structures can be implemented in EnvRtype. Table 3 presents a brief summary of the variance components for each random effect. The inclusion of enviromic sources led to a drastic reduction of variance components for residual effects (from 0.837 in M0 to 0.262 in M6), and in some cases, the increase in the variance component for genomic effects (from 0.435 in M0 to 0.548 in M5). It was possible to observe that environmental variance is a key component that explains phenotypic variation. For the same raw environmental data, each envirotyping level and modeling structure impacts on modeling the phenotypic variation, such as comparing M1 to M4 and M2 to M5. Most of the variation due to G×E effects are better captured when some enviromic information is used in the model (from 0.329 in M0 to 0.786 in M3), leading us to infer that pure genomic G×E effects are inefficient in capturing the real pattern of genotype-environment differences observed in the phenotypic records. In the present example, the joint use of different EC led to the greatest reduction in error variance (M3 and M6), but the use of single ECs can be helpful in explaining genotype-specific difference as the G×E component is better estimated by considering the reaction-norm for particular environmental factors (e.g., FRUE in models M1 and M4). Finally, the effect of envirotyping level also impacted in the model's ability in explaining phenotypic records, increasing the genomic variance (from 0.425 in M1 to 0.522 in M4, and from 0.387 in M2 to 0.548 in M5).

2.11.3. Example 3: Genomic Prediction using kernels for different development stages

Finally, we illustrate a case study of genomic prediction (GP) for two prediction scenarios (CV1 and CV00). In order to demonstrate the get_kernel function, we assumed a nonlinear kernel for enviromic effects (gaussian = TRUE). Different from the last example, in this study, we created different enviromic kernels for each development stage (M3). This model was compared to the benchmark reaction-norm model (M2, Jarquín et al., 2014) and the baseline multi-environment genomic model (M1, the MDs model López-Cruz et al., 2015). Thus, we ran the RNMM, which means that for M2 and M3 the genomic×environment effects are computed as Kronecker product between enviromic and genomic kinships, and for M1, as a block diagonal genomic matrix (MDs). In this example, we hope to demonstrate that for the same environment, it is possible to create different enviromic relatedness kernels according to the similarities found in different environments at the same development stage (Figure 4).

We assume four development stages in maize: first vegetative stage (S_1, from V1 to V6); second vegetative stage, in which the rate of biomass growth is increased in tropical maize (S_2, from V6 to VT); anthesis-silking stage (S_3, from VT to R1); and the grain filling stage (S_4, from R1 to R3). To create this relatedness, 13 ECs were computed from daily weather data. The full matrix of ECs for all development stages considered a combination of 13 ECs by 4 development stages, thus resulting in 91 environmental descriptors (BOX 18).

Figure 3 shows that for the same five environments, the crops' growing conditions, and consequently the patterns of similarity, differ according to the crop development stage. So, it's feasible to hypothesize a relatedness build up for some development stages may be more informative of the environmental-phenotype covariance among field trials than others, and perhaps more so than all environmental variables at all development stages. The biological explanation behind this hypothesis relies on ecophysiology, in which the plants are more or less sensitive to environmental variations at specific development stages. If the growing conditions drastically differ at some key development stage, and do not differ in others that do not have a strong impact on the final phenotypic expression, it is feasible to assume that the environmental variance for those "non polymorphic stages" may lead to an increased noise in the enviromic relatedness, reducing then the accuracy of GP using reaction-norms. In order to test this hypothesis, we ran the cross-validation (BOX 19) for two prediction scenarios (CV1 and CV00). At Supplementary files, its present an example of code for running GP.

Table 4 presents a summary of the variance components for the random effects of each model. An increased trend in the genomic variance was observed as the inclusion of some enviromic data (from 0.426 in M1 to 0.509 in M2 and 0.555 in M3) and reduction of residual error variance (from 0.848 in M1 to 0.269 in M2 and 0.262 in M3). Despite that, for this germplasm evaluated at this experimental network, the effect of G×E decreased when estimated using a enviromic kinship (from 0.353 in M1 to 0.269 in M2), but were better understood when dissected for each development stage (M3). The variance components for environmental relatedness plus genomic×stage interaction was higher for reproductive-related stages (S_3 and S_4), suggesting that these stages may be more important in explaining environmental-phenotype covariances across field trials than using all environmental information to build up enviromic kinships.

Finally, in Table 5, we present the accuracy of the statistical models for each prediction scenario. These results were obtained by the average Pearson's Moment Correlation (r) for each one of the 30 random samples of training sets. The use of enviromic information was beneficial for both prediction scenarios. The ability to predict novel genotypes at known growing conditions (CV1), using only the phenotypic records of 20% of the germplasm led

to an increase from r = 0.130 (baseline genomic model, M1) to r = 0.762 (benchmark reaction-norm model, M2), in which the latter was not different for the reaction-norm accounting for stage-specific enviromic kernels (r = 0.760 for M3). However, great differences between M2 and M3 were observed for predicting novel genotypes at novel growing conditions (CV00). In this scenario, based on the phenotypic records from 20% of the germplasm evaluated in 3 of the 5 environments (so the remaining 2 were used as testing-environments), the M3 model outperformed all models (r = 0.504) in comparison to M2 (r = 0.485) and M1 (r = 0.102). This last model has the worst performance due to the lack of phenotypic records.

2.12. DISCUSSION

The collection, processing, and use of envirotyping data in genomic-based studies do not depend only on the quality of the data sources. Here we demonstrate that the increased ecophysiological knowledge in envirotyping increases statistical models' accuracy in genomic prediction and provides a better explanation of the sources of variation, while increasing those efficiency models. The correct use of envirotyping data depends on the quality of data processing and is specific for each crop species (or living organism). The same 'environment' (considering a time interval for a target location) may result in different environmental types (ETs) for each organism, depending on their sensitivity to constant and transitory environmental variations. Thus, in this study, we presented some of those concepts and created functions (and gathered others from different R packages) to facilitate the use of envirotyping data in quantitative genomics.

We presented a user-friendly software, but also a cost-effective pipeline aimed at democratizing the use of envirotyping in several fields of plant research. EnvRtype is an open-source package and all advances made up until now are freely available, a situation that can ultimately boost the predictive breeding for low-budget research programs to invest in environmental sensors or perform experiments across a geographically heterogeneity region. Thus, as the remote sensing tools and databases evolve, the power of EnvRtype to perfom a quick and accurate envirotyping pipeline also evolves. In addition, other types of data sources can easily be integrated in the modeling approaches. For example, the use of high-throughput phenotypes can easy be integrated in predictive models by using get_kernel to build phenomics-realized relationship kernels (Rincent et al., 2018; Crain et al., 2018; Cuevas et al., 2019). Other kernel methods, such as Deep Kernel (Cuevas et al., 2019; Costa-Neto et al., 2021a), can be used to create kernels to be incorporated in Bayesian kernel models using kernel_model function. Thus, despite the provided end-to-end pipeline to interplay envirotyping in quantitative genomics, the users can adapt their codes and integrate different sources of information using the EnvRtype functions, such other omics data (Westhues et al., 2017).

We also showed that global envirotyping networks could be built using remote sensing tools and functions provided in EnvRtype. The combination of remote sensing + typing strategies is a powerful tool for turbocharging global partnerships of field testing and germplasm exchange. It also contributes to increasing the prediction of genotypes across a wide range of growing conditions, i.e., the so-called adaptation landscapes (Messina et al., 2018; Bustos-Korts et al., 2019). This can involve past trends and virtual scenarios (Gillberg et al., 2019; de los Campos et al., 2020). Associated with predictive GIS tools, the recommendation of cultivars could also be leveraged for specific regions (Costa-Neto et al., 2020). It could also increase a better definition of field trial positioning (Rincent et al., 2017; Resende et al., 2020) and how breeding strategies have impacted crop adaptation in the past (Heinemann et al., 2015.

Evidence of this suggestion is the increased ability of predicting novel genotypes at novel growing conditions achieved by obtaining a deeper understanding of how environments are more or less related at each development stage of crop life.

Despite the benefits and potential uses of EnvRtype, we can envisage the following limitations: (1) the resolution of satellite-based weather system (Nasa Power data base), corresponding to 0.5°×0.5° (~55,5 km × 55,5 km) of longitude by latitude, may compromise the discrimination of environments in close geographical proximity; (2) the quality of point-estimates of environmental data using extract_GIS function from public GIS databases depends on the file resolution available; (3) the need for a good registration of geographic coordinates of the target environment, but also on knowing the 'window' between harvest and sowing (for agricultural crops); and (4) management factors must be included manually in W_matrix, which we strongly suggest in order to avoid mistakes.

REFERENCES

Allen, R. G., L. S. Pereira, D. Raes, and M. Smith, 1998 Crop evapotranspiration – Guidelines for computing crop water requirements. – FAO Irrigation and drainage paper 56 / Food and Agriculture Organization of the United Nations.

Aboelkhair, H., M. Morsy, and G. El Afandi, 2019 Assessment of agroclimatology NASA POWER reanalysis datasets for temperature types and relative humidity at 2 m against ground observations over Egypt. Adv. Sp. Res. 64: 129–142.

Bartz, A. C., M. Muttoni, C. M. Alberto, N. A. Streck, G. A. Machado et al., 2017 Thermal time in sprinkler-irrigated lowland rice. Pesqui. Agropecu. Bras. 52: 475–484.

Buck AL. 1981. New equations for computing vapor pressure and enhancement factor. Journal of Applied Meteorology 20 (12): 1527-1532 DOI: 10.1175/1520-0450(1981)020<1527:NEFCVP>2.0.CO;2.

Bustos-Korts, D., M. Malosetti, K. Chenu, S. Chapman, M. P. Boer et al., 2019 From QTLs to Adaptation Landscapes: Using Genotype-To-Phenotype Models to Characterize G×E Over Time. Front. Plant Sci. 10: 1–23.

Cooper, M., F. Technow, C. Messina, C. Gho, and L. Radu Totir, 2016 Use of crop growth models with whole-genome prediction: Application to a maize multienvironment trial. Crop Sci. 56: 2141–2156.

Costa-Neto, G., R. Fritsche-Neto, and J. Crossa, 2020a Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. Heredity (Edinb).

Costa-Neto, G. M. F., O. P. Morais Júnior, A. B. Heinemann, A. P. de Castro, and J. B. Duarte, 2020b A novel GIS-based tool to reveal spatial trends in reaction norm: upland rice case study. Euphytica 216: 1–16.

Crain, J., S. Mondal, J. Rutkoski, R. P. Singh, and J. Poland, 2018 Combining High-Throughput Phenotyping and Genomic Information to Increase Prediction and Selection Accuracy in Wheat Breeding. Plant Genome 11: 0.

Cuevas, J., Crossa, J., Montesinos-López, O. A., Burgueño, J., Pérez-Rodríguez, P., and de los Campos, G. (2017). Bayesian Genomic prediction with genotype × environment kernel models. G3 (Bethesda) 7 (1), 41–53. doi: 10.1534/ g3.116.035584

Cuevas, J., O. Montesinos-López, P. Juliana, C. Guzmán, P. Pérez-Rodríguez et al., 2019 Deep Kernel for genomic and near infrared predictions in multi-environment breeding trials. G3 Genes, Genomes, Genet. 9: 2913–2924.

de los Campos, G., Gianola, D., Rosa, G. J. M., Weigel, K., and Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. 92, 295–308. doi: 10.1017/S0016672310000285

de los Campos, G., P. Pérez-Rodríguez, M. Bogard, D. Gouache, and J. Crossa, 2020 A data-driven simulation platform to predict cultivars' performances under uncertain weather conditions. Nat. Commun. 11: 1–10.

Dingman, SL, 2002. Physical hydrology. Waveland Press.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto et al., 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: 1–10.

Fick, S. E., and R. J. Hijmans, 2017 WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37: 4302–4315.

Finlay, K. W., and G. N. Wilkinson, 1963 The analysis of adaptation in a plant breeding programme. J. Agric. Res. 14: 742–754.

Galli, G., D. H. Lyra, F. C. Alves, Í. S. C. Granato, M. B. e Sousa et al., 2018 Impact of phenotypic correction method and missing phenotypic data on genomic prediction of maize hybrids. Crop Sci. 58: 1481–1491.

Galli, G., D. W. Horne, S. D. Collins, J. Jung, A. Chang et al., 2020 Optimization of UAS-based high-throughput phenotyping to estimate plant health and grain yield in sorghum. Plant Phenome J. 3: 1–14.

Gianola, D., and van Kaam, JBCHM (2008). Reproducing Kernel Hilbert spaces regression methods for genomic-assisted prediction of quantitative traits. Genet. 178, 2289–2303. doi: 10.1534/genetics.107.084285

Granato, I., J. Cuevas, F. Luna-Vázquez, J. Crossa, O. Montesinos-López et al., 2018 BGGE: A new package for genomic-enabled prediction incorporating genotype × environment interaction models. G3 Genes, Genomes, Genet. 8: 3039–3047.

Gillberg, J., P. Marttinen, H. Mamitsuka, and S. Kaski, 2019 Modelling G×E with historical weather information improves genomic prediction in new environments. Bioinformatics 35: 4045–4052.

Heinemann, A. B., C. Barrios-Perez, J. Ramirez-Villegas, D. Arango-Londoño, O. Bonilla-Findji et al., 2015 Variation and impact of drought-stress patterns across upland rice target population of environments in Brazil. J. Exp. Bot. 126: 1–14.

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

Jarquín, D., J. Crossa, X. Lacaze, P. Du Cheyron, J. Daucourt et al., 2014 A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor. Appl. Genet. 127: 595–607.

Jończyk, M., A. Sobkowiak, J. Trzcinska-Danielewicz, M. Skoneczny, D. Solecka et al., 2017 Global analysis of gene expression in maize leaves treated with low temperature. II. Combined effect of severe cold (8 C) and circadian rhythm. Plant Mol. Biol. 95: 279–302.

Krause, M. R., L. González-Pérez, J. Crossa, P. Pérez-Rodríguez, O. Montesinos-López et al., 2019 Hyperspectral reflectance-derived relationship matrices for genomic prediction of grain yield in wheat. G3 Genes, Genomes, Genet. 9: 1231–1247.

Lago, I.; Streck, N.A.; Carvalho, M.P.; Fagundes, L.K.; Paula, G.M. De; Lopes, S. ., 2009 Estimativa da temperatura base do subperíodo emergência – diferenciação da panícula em arroz cultivado e arroz vermelho. Ceres 56: 288–295.

Liu, S., C. Li, H. Wang, S. Wang, S. Yang et al., 2020 Mapping regulatory variants controlling gene expression in drought response and tolerance in maize. Genome Biol. 21: 1–22.

Luo, Q., 2011 Temperature thresholds and crop production: A review. Clim. Change 109: 583–598.

Ly, D., S. Huet, A. Gauffreteau, R. Rincent, G. Touzy et al., 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.

Lynch, M., and B. Walsh, 1998 Genetics and Analysis of Quantitative Traits. Sinauer Associates, Sunderland, Massachussets.

Messina, C. D., F. Technow, T. Tang, R. Totir, C. Gho et al., 2018 Leveraging biological insight and environmental variation to improve phenotypic prediction: Integrating crop growth models (CGM) with whole genome prediction (WGP). Eur. J. Agron. 100: 151–162.

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

Millet, E. J., W. Kruijer, A. Coupel-Ledru, S. Alvarez Prado, L. Cabrera-Bosquet et al., 2019 Genomic prediction of maize yield across European environmental conditions. Nat. Genet. 51:952–956.

Monteiro, L. A., P. C. Sentelhas, and G. U. Pedra, 2018 Assessment of NASA/POWER satellite-based weather system for Brazilian conditions and its impact on sugarcane yield simulation. Int. J. Climatol. 38: 1571–1581.

Montesinos-López, O. A., A. Montesinos-López, J. Crossa, F. H. Toledo, O. Pérez-Hernández et al., 2016 A Genomic Bayesian Multi-trait and Multi-environment Model. G3 6: 2725–2744.

Morais Junior, O. P., J. B. Duarte, F. Breseghello, A. S. G. Coelho, O. P. Morais et al., 2018 Single-Step Reaction Norm Models for Genomic Prediction in Multienvironment Recurrent Selection Trials. Crop Sci. 607: 592–607.

Muchow, R. C., J. M. Sinclair, and J. M. Bennett, 1990 Temperature and solar radiation effects on potential maize yield across locations. Agron. J. 82: 338–343.

Muchow, R. C., and T. R. Sinclair, 1991 Water Deficit Effects on Maize Yields Modeled under Current and "Greenhouse" Climates. Agron. J. 83: 1052–1059.

Pérez-Elizalde, S., J. Cuevas, P. Pérez-Rodríguez, and J. Crossa, 2015 Selection of the Bandwidth Parameter in a Bayesian Kernel Regression Model for Genomic-Enabled Prediction. J. Agric. Biol. Environ. Stat. 20: 512–532.

Pérez, P., and G. De Los Campos, 2014 Genome-wide regression and prediction with the BGLR statistical package. Genetics 198: 483–495.

Plessis, A., C. Hafemeister, O. Wilkins, Z. J. Gonzaga, R. S. Meyer et al., 2015 Multiple abiotic stimuli are integrated in the regulation of rice gene expression under field conditions. Elife 4: 1–27.

Resende, M. D. V., and J. B. Duarte, 2007 Precisão e controle de qualidade em experimentos de avaliação de cultivares. Pesqui. Agropecuária Trop. 37: 182–194.

Resende, R. T., H. P. Piepho, G. J. M. Rosa, O. B. Silva-Junior, F. F. e Silva et al., 2020 Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. Theor. Appl. Genet.

Rincent, R., E. Kuhn, H. Monod, F. X. Oury, M. Rousset et al., 2017 Optimization of multi-environment trials for genomic selection based on crop models. Theor. Appl. Genet. 130: 1735–1752.

Rincent, R., J. P. Charpentier, P. Faivre-Rampant, E. Paux, J. Le Gouis et al., 2018 Phenomic selection is a low-cost and high-throughput method based on indirect predictions: Proof of concept on wheat and poplar. G3 Genes, Genomes, Genet. 8: 3961–3972.

Shelford, V. E. 1931 Some Concepts of Bioecology. Ecology 12: 455–467.

Soltani, A., and T. R. Sinclair, 2012 Modeling physiology of crop development, growth and yield (CAB International, Ed.). International, Wallingford, Cambridge.

Souza, M. B., J. Cuevas, E. G. de O. Couto, P. Pérez-Rodríguez, D. Jarquín et al., 2017 Genomic-Enabled Prediction in Maize Using Kernel Models with Genotype × Environment Interaction. G3 7: g3.117.042341.

Sparks, A., 2018 nasapower: A NASA POWER Global Meteorology, Surface Solar Energy and Climatology Data Client for R. J. Open Source Softw. 3: 1035.

Varotto, S., E. Tani, E. Abraham, T. Krugman, A. Kapazoglou et al., 2020 Epigenetics: possible applications in climate-smart crop breeding. J. Exp. Bot. 71: 5223–5236.

Vendramin, S., J. Huang, P. A. Crisp, T. F. Madzima, and K. M. McGinnis, 2020 Epigenetic regulation of ABA-induced transcriptional responses in maize. G3 Genes, Genomes, Genet. 10: 1727–1743.

Westhues, M., T. A. Schrag, C. Heuer, G. Thaller, H. F. Utz et al., 2017 Omics-based hybrid prediction in maize. Theor. Appl. Genet.

White, J. W., G. Hoogenboom, P. W. Wilkens, P. W. Stackhouse, and J. M. Hoel, 2011 Evaluation of satellite-based, modeled-derived daily solar radiation data for the continental United States. Agron. J. 103: 1242–1251.

Xu, Y., 2016 Envirotyping for deciphering environmental impacts on crop plants. Theor. Appl. Genet. 129: 653–673

TABLES

Tabela 1. The core of environmental factors available using the 'Environmental Sensing Module' of the EnvRtype package.

Source	Environmental Factor	Unit
Nasa POWER[1]	Top-of-atmosphere insolation	MJ m^{-2} d^{-1}
	Average insolation incident on a horizontal surface	MJ m^{-2} d^{-1}
	Average downward longwave radiative flux	MJ m^{-2} d^{-1}
	Wind speed at 10 m above the surface of the earth	m s^{-1}
	Minimum air temperature at 2 m above the surface of the earth	°C d^{-1}
	Maximum air temperature at 2 m above the surface of the earth	°C d^{-1}
	The dew-point temperature at 2 m above the surface of the earth	°C d^{-1}
	Relative air humidity at 2 m above the surface of the earth	%
	Rainfall precipitation (P)	mm d^{-1}
SRTM[2]	Elevation (above sea level)	m
Computed[3]	Effect of Temperature on Radiation use Efficiency (FRUE)	-
	Evapotranspiration (ETP)	mm d^{-1}
	Atmospheric water deficit P-ETP	mm d^{-1}
	The deficit of vapor Pressure	kPa d^{-1}
	The slope of saturation vapor pressure curve	kPa C° d^{-1}
	Temperature Range	°C d^{-1}
	Global Solar Radiation based on Latitude and Julian Day	MJ m^{-2} d^{-1}

[1] collected from NASA orbital sensors (Sparks, 2018); [2] Shuttle Radar Topography Mission integrated with the raster R package; [3] processed using concepts from Allen et al. (1998) and Soltani and Sinclair (2012).

Tabela 2. Synthesis of some cardinal limits for temperature on the phenology development in the main crops. These estimates were adapted from Soltani and Sinclar (2012), Lago et al. (2009), and Bartz et al. (2017).

Specie	Suggested Cardinal Limit			
	Tbase1	**Topt1**	**Topt2**	**Tbase2**
Maize	8.0	30.0	37.0	45.0
Wheat	0.0	25.0	28.0	40.0
Rainfed Rice	8.0	30.0	37.0	45.0
Irrigated Rice (only vegetative stage)	8.0	28.0	40.0	45.0
Irrigated Rice (only reproductive stage)	15.0	25.0	35.0	45.0
Sorghum	8.0	30.0	37.0	45.0
Soybean	8.0	30.0	35.0	45.0
Peanut	8.0	30.0	35.0	45.0
Canola	0.0	25.0	28.0	40.0
Sunflower	8.0	30.0	34.0	45.0
Dry Bean	8.0	30.0	35.0	45.0
Chickpea	0.0	25.0	30.0	40.0
Barley	0.0	25.0	28.0	40.0
Sugarcane	5.0	22.5	35.0	40.0

Tabela 3. Summary of variance components [and confidence intervals] for 7 genomic-based reaction-norm models with GxE (RNMDs), considering three envirotyping levels (no envirotyping, envirotyping by environment and envirotyping by development stage at environment) and three combinations of two environmental covariates (FRUE, PETP and FRUE+PETP). Models were fitted using all phenotypic records available (n = 150 genotypes at 5 environments = 750 records). Genomic kinships were based on additive effects. Enviromic kinships were built using a linear-covariance matrix (gaussian = FALSE). FRUE and PETP denote the covariates "effect of temperature on radiation use efficiency" (from 0 to 1) and the "difference between daily precipitation and daily evapotranspiration" (mm day^{-1}), respectively.

Envirotyping Level	**Model**	**Random Effect**			
		Environment (E)	**Genotype (G)**	**G×E**	**Residual**
No Envirotyping	M0	-	0.435 [0.397;0.480]	0.329 [0.299;0.362]	0.837 [0.763;0.924]
Envirotyping by environment	M1 (EC1 = FRUE)	4.117 [3.789;4.493]	0.425 [0.387;0.468]	0.764 [0.696;0.843]	0.849 [0.773;0.936]
	M2 (EC 2 = PETP)	3.440 [3.165;3.754]	0.384 [0.350;0.423]	0.786 [0.716;0.867]	0.726 [0.662;0.801]
	M3 (EC3=FRUE+PETP)	4.279 [3.938;4.670]	0.497 [0.453;0.548]	0.664 [0.605;0.733]	0.456 [0.416;0.503]
Envirotyping by development stage at each environment	M4 (EC4 = FRUE)	8.802 [8.099;9.605]	0.522 [0.476;0.576]	0.484 [0.441;0.534]	0.266 [0.243;0.294]
	M5 (EC5 = PETP)	3.514 [3.233;3.835]	0.548 [0.500;0.604]	0.425 [0.388;0.469]	0.267 [0.243;0.295]
	M6 (EC6 = FRUE+PETP)	1.595 [1.468;1.740]	0.514 [0.468;0.566]	0.464 [0.423;0.512]	0.262 [0.238;0.289]

Tabela 4. Summary of the variance components for three modeling structures (M1, Baseline Genomic Model; M2, Benchmark Reaction-Norm model; M3, Reaction-Norm for each development stage) considering different sources of phenotypic variation due to genomic and enviromic effects. Confidence intervals (α = 5%) for each variance component is given between square brackets. Horizontal dashed lines separate the genomics, environmental and genomic × environment effects. Genomic kinships were based on additive effects. Enviromic kinships were built using a nonlinear method (gaussian = TRUE).

Random Effect	Model		
	M1	**M2**	**M3**
Genomic (G)	0.426 [0.389;0.470]	0.509 [0.464;0.562]	0.555 [0.506;0.612]
Environment (E)	-	2.686 [2.470;2.505]	-
Stage 1 (S_1: V1 to V6)	-	-	3.507 [3.227;3.827]
Stage 2 (S_2: V6 to VT)	-	-	2.711 [2.494;2.958]
Stage 3 (S_3: VT to R1)	-	-	3.940 [3.626;4.300]
Stage 4 (S_4: R1 to R3)	-	-	4.018 [3.697;4.385]
GxE*	0.353 [0.322;0.390]	0.269 [0.246;0.297]	-
GxS_1	-	-	0.308 [0.269;0.326]
GxS_2	-	-	0.295 [0.278;0.337]
GxS_3	-	-	0.306 [0.279;0.336]
GxS_4	-	-	0.304 [0.280;0.339]
Residual	0.848 [0.773;0.936]	0.269 [0.245;0.296]	0.262 [0.238;0.289]

* For M1 model, GxE is based on the Kronecker product between an identity environment matrix and the genomic kinship matrix. For M2, this Kronecker product is based on the enviromic kinship matrix instead of an identity matrix, in which this enviromic kinship were built up using envirotyping data to mimic an environmental relatedness among field trials.

Tabela 5. Accuracy (± standard deviation) of statistical models for phenotype prediction using genomic (M1) and genomic plus enviromic sources of variation (M2 and M3) in predicting novel genotypes at known environments (CV1), using 20% of the genotypes as a training set, and novel genotypes and novel environments (CV00), used as a training set 20% of the genotypes phenotyped at 3 from the 5 environments.

Model	Prediction Scenario	
	CV1	**CV00**
M1 (Baseline Genomic × Environment)	0.130 ± 0.047	0.102 ± 0.045
M2 (Benchmark Reaction-Norm)	0.762 ± 0.024	0.485 ± 0.211
M3 (Reaction-Norm for Each Development Stage)	0.760 ± 0.028	0.504 ± 0.194

FIGURES

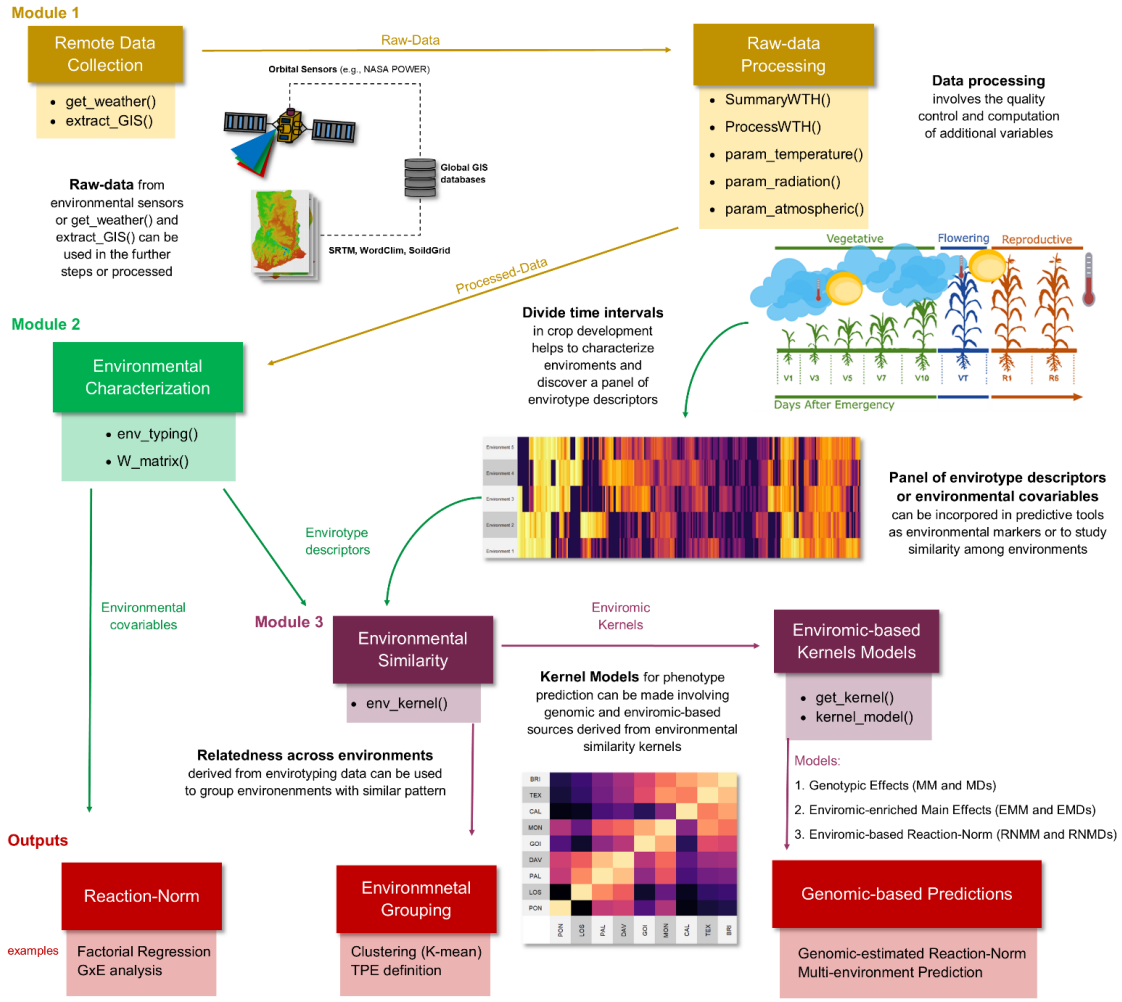

Figura 1. The workflow of the envirotyping pipeline implemented using EnvRtype in R. Yellow, green, dark purple, and red-colored boxes denote the steps related to Modules 1, 2, 3, and the examples of outputs from EnvRtype. Straight arrows indicate the flux of the envirotyping pipeline passing by each module. Curved arrows represent a process between Modules 1 and 2 in which field-growing conditions can be described as a panel of envirotype descriptors from each environmental factor processed and organized in Module 2.

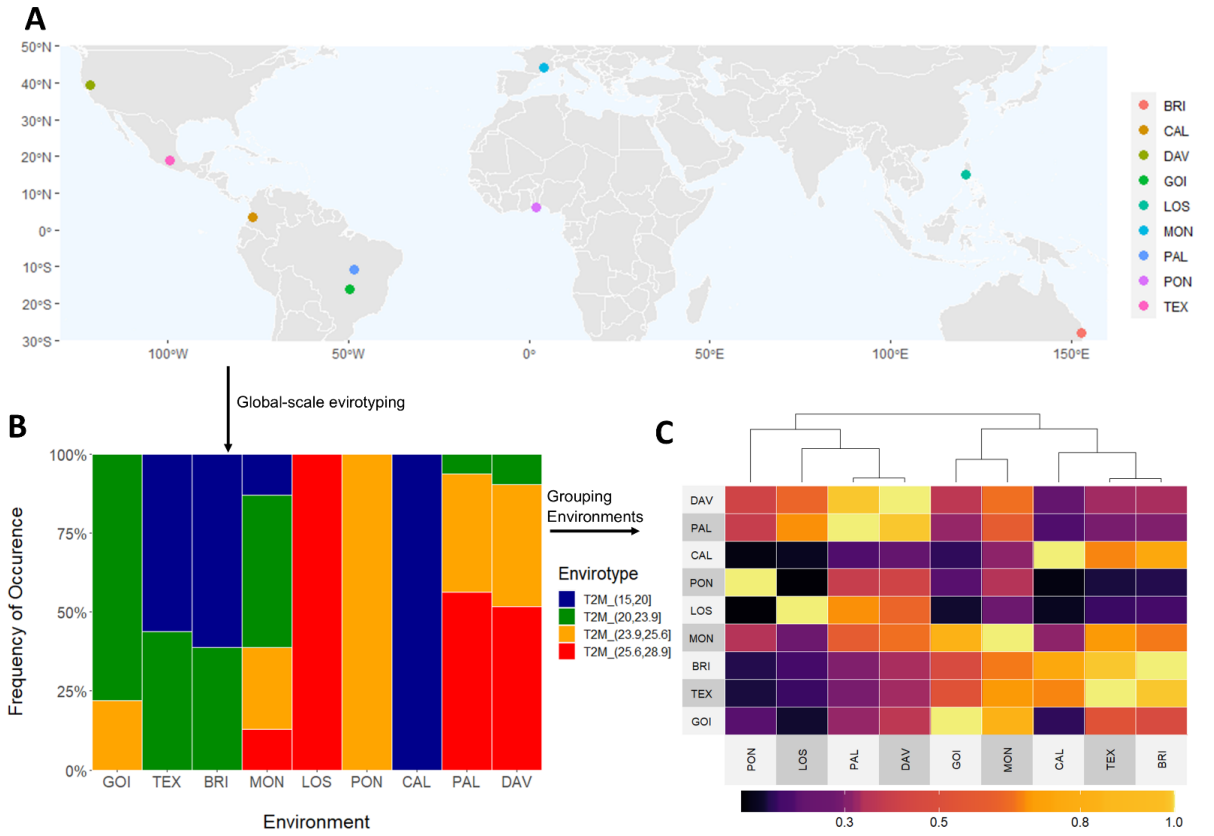

Figura 2. Workflow for a global-scale envirotyping analysis for air temperature effects in maize growing environments over diverse locations. A. Worldwide geographic positions of 9 locations used as toy examples. B. Panel of environmental types (ETs) for average air temperature during a specific month of a particular year in the summer season of each location. C. Environmental similarity matrix using Gaussian kernel method among of the observed locations using the information of the observable ETs. From A, B and C, it is possible to highlight that environmental similarity among locations in different continents can be visualized to support the creation of a global-scale experimental network and support germplasm exchange between countries.

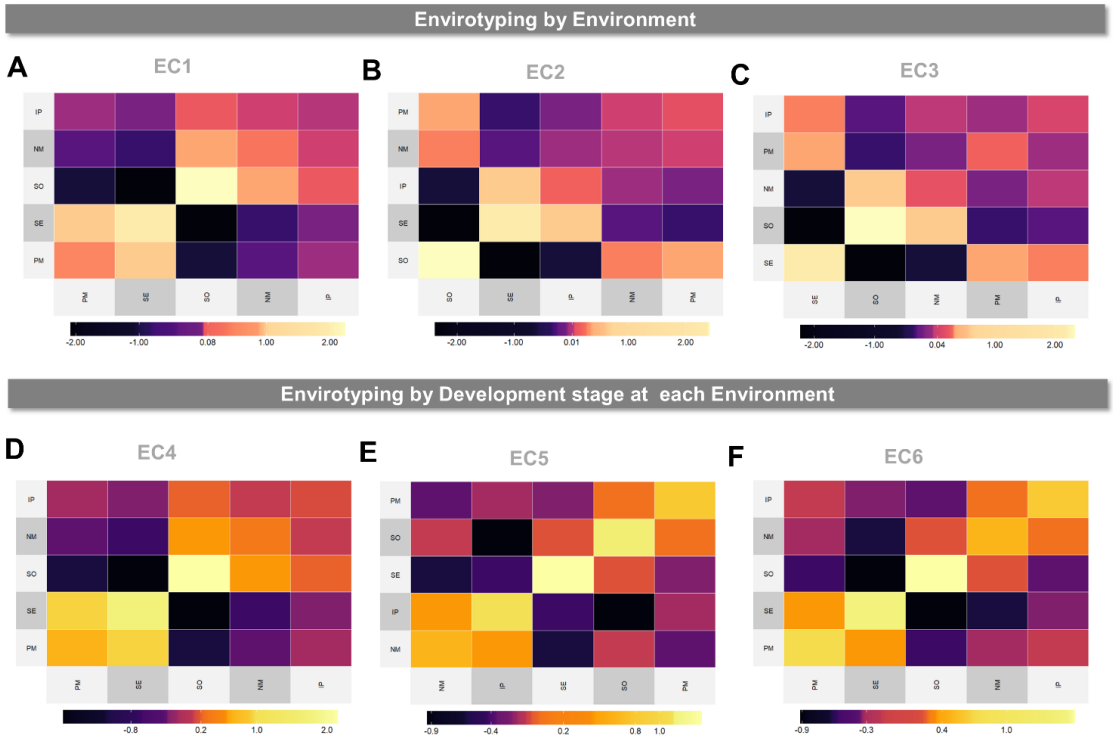

Figura 3. Linear enviromic kernels based on the combination of two environmental covariates (ECs) and two envirotyping levels (by environment and by development stage at each environment) for 5 locations (SE, PM, NV, IP and SO) across an experimental network of tropical maize. A. enviromic kernel considering only the FRUE variable (impact of temperature on radiation use efficiency) at environmental level (EC1 matrix). B. enviromic kernel considering only the PETP variable (deficit of evapotranspiration, mm.day^{-1}) for the entire crop life (EC2 matrix). C. enviromic kernel considering both FRUE and PETP for the entire live crop (envirotyping per environment, EC3 matrix). D. enviromic kernel considering only FRUE for each development stage at each environment (EC4 matrix). E. enviromic kernel considering only PETP for each development stage at each environment (EC5 matrix) and the combination of FRUE and PETP for each development stage (D, EC6 matrix).

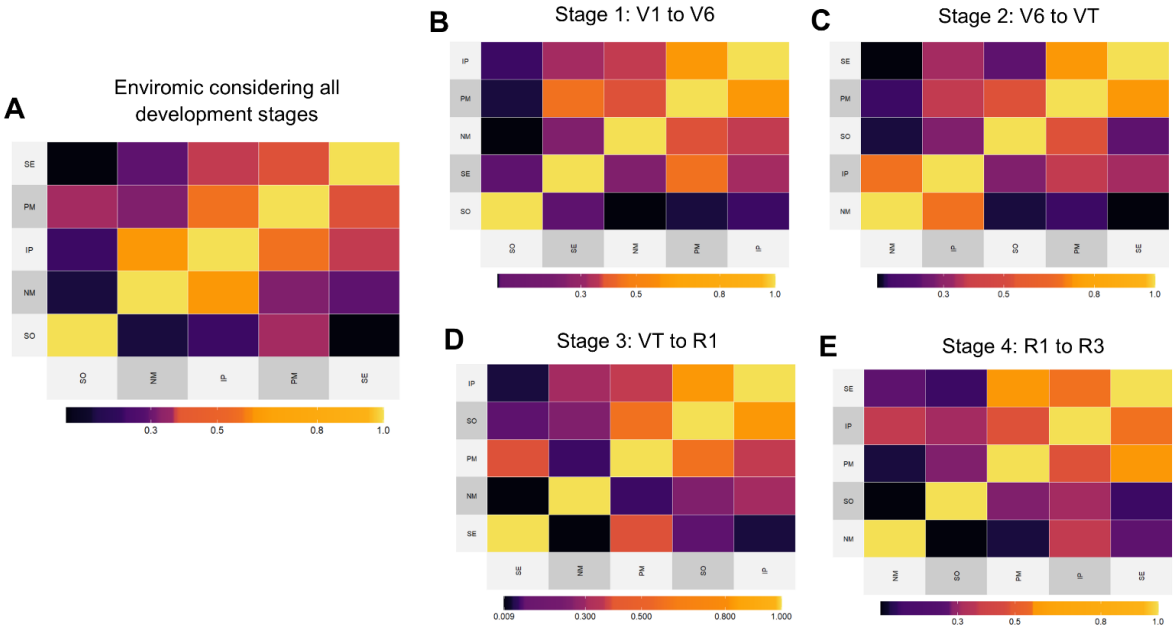

Figura 4. Figure 4. Nonlinear enviromic kernels (Gaussian) based on 13 environmental covariates over five tropical maize environments (locations (SE, PM, NV, IP and SO). A. enviromic kernel using a combination of 13 covariates at 4 development stages in maize (total of 91 ECs). B. enviromic kernel using variables 13 covariates at the initial vegetative stage (from V1 to V6). C. enviromic kernel using variables 13 covariates at the leaf growing stage (from V6 to VT). D. enviromic kernel using variables 13 covariates at the anthesis-silking interval (from VT to R1). E. enviromic kernel using variables 13 covariates at the grain filling interval (from R1 to R3).

APPENDIX

Appendix1 - Hierarchical Bayesian Modeling used in kernel_model : In this appendix, we present the hierarchical Bayesian modeling used in kernel_model function of EnvRtype. From the package for Bayesian Genotype plus Genotype × Environment (BGGE), which contains a function called BGGE(), we collected the main code and adapted it for our purposes. This function aims to solve mixed linear models through Hierarchical Bayesian Modeling--more detail about that can be found at Granato et al. (2018). Thus, we integrated the packages EnvRtype with BGGE into a single platform. If the users want to run genome-enabled models without enviromic data, we strongly suggest the use of BGGE() instead of kernel_models() because the BGGE package permits the construction of other modeling structures beyond the MM and MDs presented in this study. Below, we briefly describe the main distributions and priors used by this package.

The algorithm starts with a reparameterization of each variance-covariance matrix (K) provided by using the get_kernel() function. Each K is reparametrized using an eigen-decomposition procedure as suggested by De Los Campos et al. (2010), in which $\boldsymbol{K = USU'}$, where $\boldsymbol{S}$ is a diagonal matrix with n non-zero eigenvalues and $\boldsymbol{U}$ is an orthogonal matrix with eigenvectors. Then, an orthogonal transformation is applied to increase the computational efficiency of the further steps of the Bayesian approach. This transformations consists of a phenotypic parametrization, represented as $\boldsymbol{d = U'y}$, and any kernel-based random effect ($\boldsymbol{b = U'u}$) and error variation ($\mathbf{e} = \boldsymbol{U'\varepsilon}$) is now represented into a reparametrized normal distribution as $\boldsymbol{b} \sim N(0, \boldsymbol{U'KU\sigma_u^2}) = N(0, \boldsymbol{S\sigma_u^2})$ and $\mathbf{e} \sim N(0, \mathbf{U'U\sigma}_\varepsilon^2) = N(0, \mathbf{I}\sigma_\varepsilon^2)$ (Cuevas et al., 2017, 2019). Finally, the distribution of the transformed data is given by:

$$f(\boldsymbol{d} \mid \boldsymbol{b}, \boldsymbol{\sigma}_\varepsilon^2) = \prod_{i=1}^{n} N\left(d_i \mid b_i, \sigma_\varepsilon^2\right)$$

where the acronym i now denotes each random effect (variance-covariance) considered (from get_kernel). As this Bayesian linear model assumes $p(\boldsymbol{u} \mid \sigma_u^2) = N(\boldsymbol{u} \mid 0, \boldsymbol{K\sigma}_u^2)$, the conditional of any b_i is given as $p(b_i \mid \sigma_u^2) = N(b_i \mid 0, \sigma_u^2 s_i)$, where s_i are the eigenvalues.

Thus, we assume a conjugate prior distribution of σ_u^2 and σ_ε^2, given by inverse chi-squared with, $\mathrm{p}(\sigma_u^2) \sim \chi^{-2}(\nu_u, Sc_u)$ and $\mathrm{p}(\sigma_\varepsilon^2) \sim \chi^{-2}(\nu_\varepsilon, Sc_\varepsilon)$ respectively, in which ν_u and ν_ε denote the degree of freedom, and Sc_u and Sc_ε the scale factors for u and e. Finally, The Markov Chain Monte Carlo (MCMC) procedure is then used to generate the conditional distributions through a Gibbs sampler using the joint posterior distribution ($\boldsymbol{J} = \boldsymbol{b}, \sigma_u^2, \sigma_\varepsilon^2$), given the parameters ($\boldsymbol{P} = \boldsymbol{d}, \nu_u, \nu_\varepsilon, Sc_u, Sc_\varepsilon$ and S) as:

$$p(\boldsymbol{J} \mid \boldsymbol{P}) \propto \left\{ \prod_{i=1}^{n} N\left(d_i \mid b_i, \sigma_\varepsilon^2\right) N(b_i \mid 0, \sigma_u^2 s_i) \right\} \times \chi^{-2}(\sigma_u^2 \mid \nu_u, \nu_u Sc_u) \times \chi^{-2}(\sigma_\varepsilon^2 \mid \nu_\varepsilon, \nu_\varepsilon Sc_\varepsilon)$$

SUPPLEMENTARY FILES

Supplementary Box Codes: R codes for running the envirotyping pipelined described in this chapter.

```
################ BOX 1: Install EnvRtype #################

install.packages("devtools")
devtools::install_github("allogamous/EnvRtype")
require("EnvRtype")

################ BOX 2: Data sets ################
data("maizeYield") # toy set of phenotype data (grain yield per environment)
data("maizeG") # toy set of genomic relationship for additive effects
data("maizeWTH") # toy set of environmental data

################ BOX 3: Practical use of get_weather #################

env.data = get_weather(env.id = 'NAIROBI',country = 'KEN',
                       lat = -1.367,lon = 36.834,
                       start.day = '2015-03-01',end.day = '2015-04-01')

head(env.data)

################  BOX 4: Practical use of extract_GIS ################
data("clay_5_15")
env.data = extract_GIS(covraster = clay_5_15,name.out = 'clay_5_15',env.data = env.data)
head(env.data)

################ BOX 5: Practical use of SummaryWTH ################
summaryWTH(env.data = env.data, env.id = 'env', days.id = 'daysFromStart',statistic = 'mean')
summaryWTH(env.data = env.data) # by default

################ BOX 6: Practical use of param_temperature for Dry Beans in Nairobi, Kenya
################
TempData = param_temperature(env.data = env.data,Tbase1 = 8,Tbase2 = 45,Topt1 = 30,Topt2 = 35)
head(TempData)
env.data = param_temperature(env.data = env.data,Tbase1 = 8,Tbase2 = 45,Topt1 = 30,Topt2 =
35,merge = TRUE)
head(env.data) # merging TempData automatically

################ BOX 7: Practical use of param_atmospheric for Dry Bean Crop in Nairobi, Kenya
################
RadData = param_radiation(env.data = env.data) # first need to compute radiation parameters
head(RadData)
env.data = param_radiation(env.data = env.data,merge = TRUE)
AtmData  = param_atmospheric(env.data = env.data, Alt = 1628)
head(AtmData)
env.data = param_atmospheric(env.data = env.data, Alt = 1628,merge = TRUE)
head(env.data)

################ BOX 8: Basic use of env_typing for typing temperature in Los Baños,
Philipines, from 2000 to 2020 ################
env.data = get_weather(env.id = 'LOSBANOS',country = 'PHL',
                       lat = 14.170,lon = 121.241,variables.names = 'T2M',
                       start.day = '2000-03-01',end.day = '2020-03-01')

card = list(T2M=c(0,8,15,28,40,45,Inf)) # a list of vectors containing empirical and cardinal
thresholds
env_typing(env.data = env.data,env.id = 'env', var.id = 'T2M', cardinals = card)

################ BOX 9: Basic use of env_typing for more than one variable ################
var = c("PRECTOT", "T2MDEW") # variables
env.data = get_weather(env.id = 'LOSBANOS',country = 'PHL',
                       lat = 14.170,lon = 121.241,variables.names = var,
                       start.day = '2000-03-01',end.day = '2020-03-01')
card = list(PRECTOT = c(0,5,10,25,40,100), T2MDEW = NULL) # cardinals and data-driven limits
env_typing(env.data = env.data,env.id = 'env', var.id = var, cardinals = card)
################ BOX 10: Basic use of env_typing for more than one variable ################
data("maizeWTH") # toy set of environmental data
var = c("PRECTOT", "T2MDEW", "T2M_MAX", "T2M_MIN") # variables
W = W_matrix(env.data = maizeWTH[maizeWTH$daysFromStart < 100,],
             var.id=var, statistic="mean", by.interval=TRUE)
```

```
dim(W)

################ BOX 11: Basic use of env_kernel ################
env_kernel(env.data = W, gaussian = FALSE)
env_kernel(env.data = W, gaussian = TRUE)

################ BOX 12: Basic usage of get_kernel function ################
data("maizeYield") # toy set of phenotype data (grain yield per environment)
data("maizeG"    ) # toy set of genomic relationship for additive effects
data("maizeWTH")   # toy set of environmental data
y   = "value"      # name of the vector of phenotypes
gid = "gid"        # name of the vector of genotypes
env = "env"        # name of the vector of environments

ECs  = W_matrix(env.data = maizeWTH, var.id = c("FRUE",'PETP',"SRAD","T2M_MAX"),statistic =
'mean')
## KG and KE might be a list of kernels
KE = list(W = env_kernel(env.data = ECs)[[2]])
KG = list(G=maizeG);
## Creating kernel models with get_kernel
MM    = get_kernel(K_G = KG, y = y,gid = gid,env = env, data = maizeYield,model = "MM")
MDs   = get_kernel(K_G = KG, y = y,gid = gid,env = env,  data = maizeYield, model = "MDs")
EMM   = get_kernel(K_G = KG, K_E = KE, y = y,gid = gid,env = env,  data = maizeYield, model =
"EMM")
EMDs  = get_kernel(K_G = KG, K_E = KE, y = y,gid = gid,env = env,  data = maizeYield, model =
"EMDs")
RMMM  = get_kernel(K_G = KG, K_E = KE, y = y,gid = gid,env = env,  data = maizeYield, model =
"RNMM")
RNMDs = get_kernel(K_G = KG, K_E = KE, y = y,gid = gid,env = env,  data = maizeYield, model =
"RNMDs")

################ BOX 13: Basic usage of kernel_model ################
fixed = model.matrix(~0+env, maizeYield)
MDs   = get_kernel(K_G = KG, y = y,gid = gid,env = env,  data = maizeYield, model = "MDs")
fit   = kernel_model(y = y,env = env,gid = gid, data = maizeYield,random = MDs,fixed = fixed)

fit$yHat    # predicted phenotype values
fit$VarComp # variance components and confidence intervals
fit$BGGE    # full output of Hierarchical Bayesian Modeling

################ BOX 14: Remote Sensing for Several Places ################
env = c('GOI','TEX','BRI','MON','LOS','PON','CAL','PAL','DAV')
lat = c(-16.67,19.25,-27.47,43.61,14.170,6.294,3.261,-10.168,38.321)
lon = c(-49.25,-99.50,153.02,3.87,121.241,2.361,-76.312,-48.331,-121.442)
start = c('2020-03-15','2019-05-15','2018-09-15',
            '2017-06-18','2017-05-18','2016-07-18',
            '2017-11-18','2017-12-18','2018-07-18')
end = c('2020-04-15','2019-06-15','2018-10-15',
          '2017-07-18','2017-06-18','2016-08-18',
          '2017-12-18','2018-01-18','2018-08-18')
env.data = get_weather(env.id = env, lat = lat, lon = lon, start.day = start, end.day = end)


################ BOX 15: Discovering ETs and similarity among locations

ET = env_typing(env.data = env.data,env.id = 'env',var.id = 'T2M',format = 'wide')
EC = W_matrix(env.data = env.data,var.id = 'T2M')
distances = env_kernel(env.data = ET,gaussian = T)[[2]]
kinship   = env_kernel(env.data = EC,gaussian = F, sd.tol = 3)[[2]]

ET = env_typing(env.data = env.data,env.id = 'env',var.id = 'T2M',format = 'wide')

################  BOX 16: Envirotyping levels and model structures for GP with ECs
################
data("maizeYield") # toy set of phenotype data (grain yield per environment)
data("maizeG"    ) # toy set of genomic relationship for additive effects
data("maizeWTH")   # toy set of environmental data
y   = "value"      # name of the vector of phenotypes
gid = "gid"        # name of the vector of genotypes
env = "env"        # name of the vector of environments
### 1- Environmental Covariables (ECs)
stages    = c('VE','V1_V6','V6_VT','VT_R1','R1_R3','R3_R6',"H")
interval  = c(0,7,30,65,70,84,105)
EC1  = W_matrix(env.data = maizeWTH, var.id = 'FRUE')
EC2  = W_matrix(env.data = maizeWTH, var.id = 'PETP')
EC3  = W_matrix(env.data = maizeWTH, var.id = c('FRUE','PETP'))
EC4  = W_matrix(env.data = maizeWTH, var.id = 'FRUE',
```

```
                by.interval = T,time.window = interval,names.window = stages)
EC5  = W_matrix(env.data = maizeWTH, var.id = 'PETP',
                by.interval = T,time.window = interval,names.window = stages)
EC6  = W_matrix(env.data = maizeWTH, var.id = c('FRUE','PETP'),
                by.interval = T,time.window = interval,names.window = stages)

### 2- Kernels
K1 = list(FRUE = env_kernel(env.data = EC1)[[2]])
K2 = list(PETP = env_kernel(env.data = EC2)[[2]])
K3 = list(FRUE_PETP = env_kernel(env.data = EC3)[[2]])
K4 = list(FRUE = env_kernel(env.data = EC4)[[2]])
K5 = list(PETP = env_kernel(env.data = EC5)[[2]])
K6 = list(FRUE_PETP = env_kernel(env.data = EC6)[[2]])
### 3- Obtain Kernel Models
M0  = get_kernel(K_G = KG,  y = y,gid = gid,env = env,  data = maizeYield, model = "MDs")
M1  = get_kernel(K_G = KG, K_E = K1, y = y,gid = gid,env = env,  data = maizeYield, model =
"RNMDs")
M2  = get_kernel(K_G = KG, K_E = K2, y = y,gid = gid,env = env,  data = maizeYield, model =
"RNMDs")
M3  = get_kernel(K_G = KG, K_E = K3, y = y,gid = gid,env = env,  data = maizeYield, model =
"RNMDs")
M4  = get_kernel(K_G = KG, K_E = K4, y = y,gid = gid,env = env,  data = maizeYield, model =
"RNMDs")
M5  = get_kernel(K_G = KG, K_E = K5, y = y,gid = gid,env = env,  data = maizeYield, model =
"RNMDs")
M6  = get_kernel(K_G = KG, K_E = K6, y = y,gid = gid,env = env,  data = maizeYield, model =
"RNMDs")

################  BOX 17: Fitting Genomic-enabled models with enviromic data
fixed = model.matrix(~0+env,maizeWTH)

iter = 1000
burn = 500
seed = 78172
thin = 10
model = paste0('M',0:6)
Models = list(M0,M1,M2,M3,M4,M5,M6)
Vcomp <- c()
for(MODEL in 1:length(Models)){
  set.seed(seed)
  fit <- kernel_model(data = maizeYield,y = y,env = env,gid = gid,
                      random = Models[[MODEL]],fixed = Z_E,
                      iterations = iter,burnin = burn,thining = thin)

  Vcomp <- rbind(Vcomp,data.frame(fit$VarComp,Model=model[MODEL]))
}

Vcomp

reshape2::dcast(Vcomp,Model~Type,sum,value.var = 'Var')
reshape2::dcast(Vcomp,Model~Type,sum,value.var = 'CI_lower')
reshape2::dcast(Vcomp,Model~Type,sum,value.var = 'CI_upper')

################  BOX 18: Bulding enviromic kernels for each development stage
################
data("maizeYield") # toy set of phenotype data (grain yield per environment)
data("maizeG"    ) # toy set of genomic relationship for additive effects
data("maizeWTH")   # toy set of environmental data
y   = "value"      # name of the vector of phenotypes
gid = "gid"        # name of the vector of genotypes
env = "env"        # name of the vector of environments

## Organizing Environmental Covariables (ECs) in W matrix
stages    = c('VE','V1_V6','V6_VT','VT_R1','R1_R3','R3_R6',"H")
interval = c(0,7,30,65,70,84,105)
id.vars  = names(maizeWTH)[c(10:15,23,25:30)]

W.matrix = W_matrix(env.data = maizeWTH,env.id = 'env',
                    var.id = id.vars,by.interval = T,time.window = interval,
                    names.window = stages,center = F,scale = F )


## Kernel for the involving all development stages
K_F <- env_kernel(env.data = W.matrix,gaussian = T)[[2]]
```

```
## Kernels for each development stage
K_S <- env_kernel(env.data = W.matrix,gaussian = T,stages = stages[2:5])[[2]]
# K_G (genotype) and K_E (environment) must be a list of kernels
# So:
K_G = list(G = maizeG)
# And:
K_F <- list(E = K_F)
# for K_S, we dont need to create a list because K_S is already a list of kernels for each
development stage
## Assembly Genomic and Enviromic Kernel Models
M1 = get_kernel(K_G = K_G,data = maizeYield,env = env,gid = gid,y = y, model = "MDs") #
baseline model
M2 = get_kernel(K_G = K_G, K_E = K_F, data = maizeYield,env = env,gid = gid,
                y = y, model = "RNMM",dimension_KE = 'q') # reaction-norm 1
M3 = get_kernel(K_G = K_G, K_E = K_S, data = maizeYield,env = env,gid = gid,
                y = y,model = "RNMM",dimension_KE = 'q') # reaction-norm 2

model = c('Baseline Genomic','Reaction-Norm','Reaction-Norm for each Dev.Stage')
Models = list(M1,M2,M3) # for running all models in loop

iter = 10E3 # number of iterations
burn = 5E3  # number of burn in
thin = 10   # number for thining

Z_E = model.matrix(~0+env,data=maizeYield) # fixed environmental effects

Vcomp <- c()
for(MODEL in 1:length(Models)){
  set.seed(seed)
  fit <- kernel_model(data = maizeYield,y = y,env = env,gid = gid,
                      random = Models[[MODEL]],fixed = Z_E,
                      iterations = iter,burnin = burn,thining = thin)

  Vcomp <- rbind(Vcomp,data.frame(fit$VarComp,Model=model[MODEL]))
}

Vcomp

#################  BOX 19: Genomic Prediction using kernel_model ################
## Cross-validation to assess predictive ability of GP models (kernel_model function)
source('https://raw.githubusercontent.com/gcostaneto/SelectivePhenotyping/master/cvrandom.R')

rep  = 3
seed = 1010
f    = 0.20
iter = 5E3
burn = 1E3
thin = 10

## CV1
TS = Sampling.CV1(gids = maizeYield$gid,f = f,seed = seed,rep = rep,gidlevel = F)

require(foreach)
require(EnvRtype)

Y = maizeYield
results <-foreach(REP = 1:rep, .combine = "rbind")%:%
  foreach(MODEL = 1:length(model), .combine = "rbind")%dopar% {


    yNA      <- Y
    tr       <- TS[[REP]]
    yNA$value[-tr] <- NA

    Z_E = model.matrix(~0+env,data=yNA) # fixed environmental effects

    fit <- kernel_model(data = yNA,y = y,env = env,gid = gid,
                        random = Models[[MODEL]],fixed = Z_E,
                        iterations = iter,burnin = burn,thining = thin)


    df<-data.frame(Model = model[MODEL],rep=REP,
                   rTr=cor(Y$value[tr ], fit$yHat[tr ],use = 'complete.obs'),
                   rTs=cor(Y$value[-tr], fit$yHat[-tr],use = 'complete.obs'))

    write.table(x = df,file = 'point-estimate_r.txt',sep=',',append = T,row.names=T)
```

```
    output <- data.frame(obs=Y$value,pred=fit$yHat,
                         gid=Y$gid, env=Y$env,
                         Model = model[MODEL],rep=REP,pop=NA)


    output$pop[tr ] <- 'training'
    output$pop[-tr] <- 'testing'

    return(output)
  }
#stopCluster(cl)

results
require(plyr)
pa = ddply(results,.(rep,pop,Model),summarise,r = cor(obs,pred))
ddply(pa,.(pop,Model),summarise, pa = round(mean(r),3),sd = round(sd(r),3))


## CV00 (sampling genotypes and environments)

seed = 8172
# out environments (as testing set) = 2
TS=Sampling.CV0(gids = Y$gid,envs = Y$env,out.env = 2,f = f,seed = seed,rep = rep)

cl <- makeCluster(3)
registerDoParallel(cl)

results <-foreach(REP = 1:rep, .combine = "rbind")%:%
  foreach(MODEL = 1:length(model), .combine = "rbind")%dopar% {


    yNA      <- Y
    tr       <- TS[[REP]]$training
    yNA$value[-tr] <- NA

    Z_E = model.matrix(~0+env,data=Y)
    fit <- kernel_model(data = yNA,y = y,env = env,gid = gid,
                        random = Models[[MODEL]],fixed = Z_E,
                        iterations = iter,burnin = burn,thining = thin)


    output <- data.frame(obs=Y$value,pred=fit$yHat,
                         gid=Y$gid, env=Y$env,
                         Model = model[MODEL],rep=REP,pop=NA)


    output$pop[tr ] <- 'training'
    output$pop[-tr] <- 'testing'

    return(output)
  }

stopCluster(cl)

pa = ddply(results,.(rep,pop,Model),summarise,r = cor(obs,pred))
ddply(pa,.(pop,Model),summarise, pa = round(mean(r),3),sd = round(sd(r),3))
```

3. NONLINEAR KERNELS, DOMINANCE AND ENVIROTYPING DATA INCREASES ACCURACY OF GENOMIC PREDICTION IN MULTI-ENVIRONMENTS

ABSTRACT

Modern whole-genome prediction (WGP) frameworks that focus on multi-environment trials (MET) integrate large-scale genomics, phenomics, and envirotyping data. However, the more complex the statistical model, the longer the computational processing times, which do not always result in accuracy gains. We investigated the use of new kernel methods and modeling structures involving genomics, and non-genomic sources of variation in two MET maize datasets. Five WGP models were considered, advancing in complexity from a main effect additive model (A) to more complex structures including dominance deviations (D), genotype × environment interaction (AE and DE), and the reaction norm model using environmental covariables (W) and their interaction with A and D (AW+DW). A combination of those models built with three different kernel methods, Gaussian kernel (GK), Deep kernel (DK) and the benchmark genomic best linear unbiased predictor (GBLUP/GB), was tested under three prediction scenarios: newly developed hybrids (CV1); sparse MET conditions (CV2), and new environments (CV0). GK and DK outperformed GB in prediction accuracy and reduction of computation time (~up to 20%) under all model-kernel scenarios. GK was more efficient in capturing the variation due to A+AE and D+DE effects and translated it into accuracy gains (~up to 85% compared to GB). DK provides more consistent predictions, even for more complex structures such as W+AW+DW. Our results suggest that DK and GK are more efficient in translating model complexity into accuracy, and more suitable for including dominance and reaction norm effects in a biologically accurate and faster way.

Keywords: Genotype × environment interaction; Non-additive effects; Reaction-norm; Enviromics

Published at Heredity (2021) 126:92–106 , as DOI: https://doi.org/10.1038/s41437-020-00353-1

3.1. INTRODUCTION

Historically, utilizing the best linear unbiased prediction (BLUP) has been useful for predicting the performance of unobserved maize hybrids utilizing pedigree or molecular marker relationships of all crosses (Bernardo 1994, 1996). The assessment and prediction of hybrid performance have two main sources of variation: the estimated additive (A) effects among lines based on the variance of the general combining ability of the two parents, and the dominance (D) (and/or epistatic) effects among lines based on the variance of the specific combining ability of the cross between parents (Alves et al. 2019). These two sources are fundamental for prediction based on either pedigree or genome-wide marker information (or both) of the lines forming the single cross. Multi-environment testing (MET) of single crosses facilitates sampling of genotype × environment interactions (GE), as well as additive × environment (AE) and dominance × environment (DE) interactions, and it allows hybrids unobserved in field evaluation to be predicted based on existing data from other observed hybrids derived from related lines.

Prediction-based strategies employing genomic-assisted data (Meuwissen et al. 2001) are responsible for the greatest leaps in genetic gain and reduction of time between selection cycles in both animal and plant breeding programs (Crossa et al. 2017; Voss-Fels et al. 2019). Whole genomic prediction (WGP) focuses on modeling genomic effects due to dense molecular markers related to quantitative-genetics concepts, such as additive and non-additive

variation. WGP studies conducted over the last decade include BLUPs based on different prediction methods, i.e., Ridge Regression and the Genomic Best Linear Unbiased Predictor (GBLUP, VanRaden 2008). These methods have been extensively and intensively employed in maize and wheat hybrid prediction (Windhausen et al. 2012; Lehermeier et al. 2014; Technow et al. 2014; Acosta-Pech et al. 2017; Zhang et al. 2017; Basnet et al. 2019).

However, most genomic hybrid prediction studies ignore GE interaction and do not incorporate environmental covariables to model similarities between environments. In maize, Acosta-Pech et al. (2017) incorporated GE and marker information to predict hybrid performance. A recent study on hybrid wheat investigated the genomic-enabled prediction of single-cross wheat hybrids using models with various combinations of pedigree, markers, and/or their interaction with environments (Basnet et al. 2019). This study on hybrid wheat showed that hybrid prediction accuracy increases when environmental covariables are incorporated and when additive × environmental covariables and dominance × environmental covariables are included in the GBLUP reaction norm model (Jarquin et al., 2014). Thus, selection guided by genomic-enabled prediction in multiple environment trials (WGP-MET) can result in optimization of the breeding pipeline by increasing the number of possible hybrids and evaluated environments, especially when aiming to choose the best hybrids for certain environmental conditions, i.e., capable of capturing the effects of the GE. Usually, the WGP-MET models in maize have integrated mainly A and their interaction with environment (AE). However, more recently, some authors have suggested that the inclusion of dominance effects and their interaction with environments (D and DE) may lead to more accurate WGP-based selection in MET (Wang et al., 2017; Dias et al., 2018; Ferrão et al., 2020).

On the other hand, the use of data derived from environmental typing analysis (e.g., environmental covariables, W) can be an important source to bridge the gap between phenotypic correlations and genomic correlations across MET (Cooper et al., 2014). WGP models including the so-called envirotyping (Xu, 2016) analysis can be used to mimic the linear response of the phenotypic performance of genotypes over a certain type of environmental gradient (envirotype), i.e., reaction norm (Jarquín et al., 2014; Crossa et al., 2017), in which the GE effects are studied as an extension of the GBLUP. The theoretical basis of this modeling approach relies on assuming that the differential envirotype-to-phenotype dynamics for different genotypes drivers the GE variation over MET (Millet et al., 2019; Costa-Neto et al., 2020; Porker et al. 2020). In this context, there is a genomic background impacting the phenotypic responses across environments. As the genotypes differ in terms of their allelic constitution, the number of copies of an allele (additivity) and intra-allelic interactions (dominance) are expected to have different degrees of influence on how genotypes respond to environmental variations and how meaningful AW and DW interactions are. For this reason, efforts have focused on a more in-depth search for the genomic causes that are linked to the ecophysiological responses of cultivation, either through genomic association studies (Li et al., 2018) or by genomic prediction considering reaction norm kernels (Jarquín et al., 2014; Morais Júnior et al., 2018) or whole-genome × envirotyping-based factorial regression models (Ly et al., 2018; Millet et al., 2019).

As already mentioned, the GBLUP (GB) (VanRaden, 2008) uses a linear kernel. Other methods consider the complete genetic values of individuals, including both additive and non-additive (dominance and epistasis) effects, thereby estimating the genetic performance of the lines or hybrids (Crossa et al., 2017). The complexity of applying genomic-based prediction breeding is influenced by various factors acting at different levels. Some of the statistical complexities can be addressed by using semi-parametric genomic regression methods to account for non-additive variation (Gianola et al., 2006, 2011; Gianola and Van Kaam, 2008; Morota and Gianola, 2014). These methods have been used to predict complex traits with promising practical results (González-Camacho et al. 2012; Pérez-Rodríguez et al. 2012). Semi-parametric models often used non-linear kernel methods for addressing complex gene actions, thus

capturing non-linear relations between phenotype and genotype. A commonly used kernel is the Gaussian kernel (GK) based on molecular markers (Gianola et al. 2014). Cuevas et al. (2016, 2018) and Souza et al. (2017) showed that using the GK within the multi-environment genomic GE model of Jarquín et al. (2014) led to higher prediction accuracy than the same method with the linear kernel GB. Parametric alternatives for modeling epistasis have also been broadly discussed in the literature (Jiang and Reif 2015; Martini et al. 2016).

Recently, Cuevas et al. (2019) introduced the arc-cosine kernel (AK) function for genome-enabled prediction. The non-linear AK is defined by a covariance matrix that emulates a deep learning model with one hidden layer and a large number of neurons. A recursive formula allows altering the covariance matrix stepwise, thus adding more hidden layers to the emulated deep neural network. The AK kernel method has been used in both single-environment and multi-environment models, including genomic × environment interaction (GE) (Crossa et al. 2019; Cuevas et al. 2019). The results of these authors show that AK genomic-enabled prediction accuracy is similar to that of the GK, but AK has the advantage over GK that it is computationally more straightforward, since no bandwidth parameter is required, while performing similarly or slightly better than GK. The tuning parameter "number of layers" required for AK can be determined by a maximum marginal likelihood procedure (Cuevas et al. 2019). Because the AK emulates the action of the deep learning method, we also name the AK kernel method as Deep kernel (DK) (Crossa et al. 2019). In this article, we will use AK and DK interchangeably.

Based on the previous studies and on the advantage of using several linear and non-linear kernel relationships between the covariables (markers and environmental covariables), in this study, we tested the practical aspects of 5 WGP models. There are only three main-effects models including environments (E), additive (A), dominance (D) and envirotype (W) (EA, EAD, EADW), and two are main effects plus GE and GW interactions (EAD+GE, EADW+GW) accounting for different genomic and GE and GW covariance structures and using three kernel methods (GB, GK, and DK). Note that the GE interaction includes EA+ED, whereas GW includes AW+DW. First, we compare the differences between WGP and kernel methods to explain the sources of variation and reduction error variance in MET. Next, we check the computational efficiency of running these models under a Bayesian framework. Finally, we compute the accuracy of each model-kernel method combination using three prediction problems faced by most hybrid maize breeding programs:

- Predicting hybrids untested in any environment (CV1);
- Predicting hybrids over incomplete trials (the so-called sparse testing, CV2);
- Predicting hybrids in entirely novel environments (CV0).

The three kernel methods were used on the two types of covariables employed, (1) dense molecular markers, and (2) dense environmental covariables collected in all the environments considered in the two data sets.

3.2. MATERIAL AND METHODS

The material and methods are organized as follows. First, in sections “Environmental Typing” and "Maize Data," we describe the maize data sets used, including genomic and phenotypic data (grain yield, tons per ha), and how environmental data were collected and processed. Next, in the sections “Kernel Methods” and "Implemented Models," we describe the combinations of the 5 MET-WGP models, including different structures to accommodate genomic and envirotypic data, and the three kernel methods used to model them (GB, GK, and DK). Finally, in “Assessing

prediction accuracy by cross-validation", we present the statistical efforts used in testing each combination of the model-kernel method under different experimental network scenarios (CV1, CV2, and CV0).

3.2.1. Environmental typing

Environmental typing (envirotyping) is a core of procedures used to collect, process, and integrate environmental factors as non-genomic covariates into genetic-informed studies (Cooper et al. 2014; Xu 2016). In this study, a total of 16 environmental factors was used to create what we call envirotype covariable matrix **W** (Table 1).

First, daily environmental data were obtained from NASA orbital sensors (Sparks 2018). Next, additional variables describing ecophysiological processes (e.g., evapotranspiration, the impact of air temperature on radiation use efficiency) were computed as extensively described by Allen et al. (1998) and Soltani and Sinclair (2012). Finally, to capture the temporal variation of the environmental information across crop development, the crop cycles were divided into five-time intervals:

- From 0 DAE (emergence day) to 14 DAE (appearance of the first leaf, V1);
- From 15 DAE (V1) to 35 DAE (appearance of the fourth leaf, V4);
- From 36 DAE (V4) to 65 DAE (tasseling stage, VT);
- From 66 DAE (VT) to 90 DAE (kernel milk stage, R3);
- From 91 DAE (R3) to 120 DAE (physiological maturation).

These time intervals were defined based on agronomic expertise of how tropical maize grows in Brazil's environments. For each variable-phenology combination, we calculated the first (25%), second (50%), and third (75%) percentiles of each combination of environmental variable × time interval across different environments. By using three percentiles, we hope to better capture the statistical distribution of each environmental variable in order to better represent the similarities between environments. In this sense, each combination of environmental variable × time interval × quantile has become an envirotype descriptor of the environmental relatedness. Finally, quality control was done by removing covariables with more than 3±SD, where SD is the standard deviation of the covariables across environments (Morais Júnior et al. 2018). This envirotyping pipeline was developed using the core of functions present in the R package *EnvRtype* (Costa-Neto et al., 2021b).

3.2.2. Maize data

The phenotypic data consisted of grain yield (ton/ha) records collected from two data sets of tropical maize hybrids in Brazil (HEL and USP). Both sets include data from Souza et al. (2017) that have been used in previous proof-of-concept studies. Details about the experimental design, cultivation practices, and fundamental statistical analysis are given in Souza et al. (2017) and Alves et al. (2019). Below, we summarize the number of hybrids, the number of environments, and the genomic and envirotyping data used.

Phenotypes, genotypes and environmental covariables for the HEL data set: The HEL data set is based on the germplasm developed by the Helix Seeds Company (HEL) in South America. It includes a set of 247 maize hybrids from a core of 452 F_1 hybrids obtained by crossing 106 inbred lines. Those hybrids were evaluated in 2015 in five sites in Brazil (S1-S3 in the southern region and S4-S5 in the mid-west region). Parent lines were genotyped with

an Affymetrix Axiom Maize Genotyping Array of 616 K SNPs (Single Nucleotide Polymorphisms) (Unterseer et al. 2014). Then, standard quality controls (QC) were applied to the data, by removing markers with a Call Rate ≥ 0.95. After this process, the remaining missing data in the lines were imputed with the *Synbreed* package (Wimmer et al. 2012) using the algorithms from the *Beagle 4.0* software (Browning and Browning 2008). Finally, markers with a Minor Allele Frequency (MAF) of ≤ 0.05 were removed, resulting in a total of 52,811 high-quality SNPs. Souza et al. (2017) described both phenotypic and genomic data of inbred lines as credited to the Helix Seeds Ltda. Company. According to the geographic coordinates, environmental data were collected for each of the five sites (Supplementary Table S1). At the end of the process described in the Environmental Typing section, a total of 243 envirotype covariables were obtained (combinations of environmental variables × time intervals × percentiles).

Phenotypes, genotypes and environmental covariables for the USP data set: The USP data set is based on the germplasm developed by the Luiz de Queiroz College of Agriculture of the University of São Paulo (USP), Brazil. From 2016 to 2017, a partial diallel experiment involving 49 inbred lines resulting in 906 F_1 hybrids was conducted, and 570 of those hybrids were evaluated across eight environments (E1-E8), involving an arrangement of 2 locations, 2 years, and 2 nitrogen levels. The two sites used in this study involved two distinct biomes with different edaphoclimatic patterns, i.e., Piracicaba (Atlantic Forest, clay soil) and Anhumas (Savannah, silt-sandy soil). At each site, two contrasting nitrogen (N) fertilization levels were used. One experiment was conducted under ideal N conditions and received 100 kg ha^{-1} of N (30 kg ha^{-1} at sowing and 70 kg ha^{-1} in a coverage application at the V8 plant stage), while the second experiment under low N conditions received only 30 kg ha^{-1} of N at sowing. As described in the HEL data set, the parent lines were genotyped with an Affymetrix Axiom Maize Genotyping Array of 616 K SNPs. Markers with a Minor Allele Frequency (MAF) of ≤ 0.05 were removed. After all QC procedures, a total of 54,113 high-quality SNPs was available for predictions. Environmental data were collected for each of the two sites and two years according to the planting date and geographic coordinates (Supplementary Table S1). A nitrogen management variable was inserted, designating the amount of nitrogen applied in the development cycle (ideal N = 100; low N = 30). At the end of the process described in the Environmental Typing section, a total of 248 envirotype covariables was obtained.

3.2.3. Kernel methods

In this study, we tested three methods to estimate the relationship kernels for additive effects ($\boldsymbol{K}_{\mathbf{A}}$), dominance deviations ($\boldsymbol{K}_{\mathbf{D}}$), and envirotype-informed environmental relatedness ($\boldsymbol{K}_{\mathbf{W}}$). The additive effects were modeled from the molecular data, assuming $\mathbf{A} = \{0 = A^2A^2; 1 = A^1A^2; 2 = A^1A^1\}$. Dominance deviations were computed by re-coding the matrix of molecular markers for each individual as $\mathbf{D} = \{-2f_l^2 = A^2A^2; 2f(1-f_l) = A^1A^2; -2f(1-f_l)^2 = A^1A^1\}$ (Vitezica et al. 2013), where f_l is the frequency of the favorable allele at locus l. Finally, the envirotyping-based matrix W (q environments × k covariables), with $w \sim N(0,1)$, was constructed by mean-centering and scaling the environmental information (Environmental Typing Section). Each of the three kernel methods is detailed below.

Benchmark Genomic Best Linear Unbiased Predictor: The first method is the traditional GBLUP (in short refered as GB), where we obtained the covariance matrix from the following expression:

$$GB: \ \boldsymbol{K} = \frac{\boldsymbol{XX'}}{\text{trace}(\boldsymbol{XX'})/nrow(\boldsymbol{X})}$$

where *K* is a generic representation of the relationship kernel ($\boldsymbol{K_A}$, $\boldsymbol{K_D}$, and $\boldsymbol{K_W}$), and $\boldsymbol{X}$ is a generic representation of the molecular or envirotyping informed matrix (A, D and W). By $nrow(\boldsymbol{X})$ we denote the number of rows in $\boldsymbol{X}$ matrix. The GB method was also used as a benchmark for comparisons with the following two methods.

Gaussian Kernel: The non-linear Gaussian Kernel (GK) method was the second type of kernel method used in this study. Unlike GB, this kernel is estimated from an exponential relation based on the Euclidean Distance $\boldsymbol{D}^2_{ii'} = \sum_k (x_{ik} - x_{i'k})^2$ matrix for each pairwise elements in the $\boldsymbol{X} = \{x_i, x_{i'}\}$ pondered by its median (a scalar variable, Q) and a bandwidth parameter (a scalar variable, h) that controls the rate of decay of the covariance between individuals, resulting in:

$$GK: \boldsymbol{K} = \exp(h\boldsymbol{D}^2/Q)$$

where the diagonal of the GK-based covariance matrix is equal to 1. The bandwidth parameter (h) was estimated for each relationship kernel ($\boldsymbol{K_A}$, $\boldsymbol{K_D}$, and $\boldsymbol{K_W}$) following the marginal function described in Pérez-Elizalde et al. (2015).

Deep Kernel: The arc-cosine kernel (referred to here as DK) is the third kernel method tested in this study. Cuevas et al. (2019) and Crossa et al. (2019) introduced the use of deep kernels in genomic prediction for multiple environments based on the additive relationship effects. Here we introduce the frequent use of DK for the joint modeling of additive, dominance, and reaction-norm kernels.

The general formulation of the DK method is based on the proposition of Neal (1996) for a Bayesian method for deep artificial neural networks (ANN). After that, Williams (1998) and Cho and Saul (2009) established the relationship between the DK method and a deep neural network with one hidden layer. In this context, the DK method aims to emulate a deep learning approach, exploring the relationship between individuals within an $\boldsymbol{X}$ matrix of inputs (e.g., molecular markers, near-infrared data) through the angle ($\theta_{i,i'}$) between two designed vectors of individuals ($\mathbf{x}_i \cdot \mathbf{x}_{i'}$):

$$\theta_{i,i'} = \cos^{-1}\left(\frac{\mathbf{x}_i \cdot \mathbf{x}_{i'}}{\|\mathbf{x}_i\|\|\mathbf{x}_{i'}\|}\right)$$

where $\cdot$ denotes the inner product, and $\|\mathbf{x}_i\|$ is the norm of hybrid *i*. Cuevas et al. (2019) described a maximum marginal likelihood method used to select the number of hidden layers (l) for the DK kernel. As described by Cuevas et al. (2019), the following kernel is positive semidefinite and related to an ANN with a single hidden layer, in which Cho and Saul (2009) describes the activation function as:

$$DK^1(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{1}{\pi}\|\mathbf{x}_i\|\|\mathbf{x}_{i'}\|J\left(\theta_{i,i'}\right)$$

where π is the pi constant and $J\left(\theta_{i,i'}\right)$ is computed by $J\left(\theta_{i,i'}\right) = \left[\sin\left(\theta_{i,i'}\right) + \left(\pi - \theta_{i,i}\right)\cos\left(\theta_{i,i'}\right)\right]$. The DK^1 is the base kernel defined by a symmetric positive semidefinite matrix, capable of preserving the norm of the entries such as $\text{DK}(x_i, x_i) = \|x_i\|^2$, and $DK(x_i, -x_i) = 0$ models the non-linear and orthogonal relationships. Cho and Saul (2009) and Cuevas et al. (2019) present a recursive relationship approach to shape a basic DK^1 into a final DK emulating ANN hidden layers (l), repeating l times the interior product:

$$DK: \boldsymbol{K} = \ DK^{(l+1)}(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{1}{\pi}\left[DK^{(l)}(\mathbf{x}_i, \mathbf{x}_i)DK(x_{i'}, x_{i'})\right]^{\frac{1}{2}}J\left(\theta^{(l)}_{i,i'}\right)$$

where $\theta_{i,i'}^{(l)} = \cos^{-1}\left\{DK^{(l)}(\mathbf{x}_i, \mathbf{x}_{i'})\left[DK^{(l)}(\mathbf{x}_i, \mathbf{x}_i)DK(\mathbf{x}_{i'}, \mathbf{x}_{i'})\right]^{-\frac{1}{2}}\right\}$. Thus, computing $DK^{(l+1)}$ at level (layer) l+1 is done from the previous layer $DK^{(l)}$. To select the number of hidden layers l to fill this process for each relationship kernel ($\boldsymbol{K}_{\mathbf{A}}$, $\boldsymbol{K}_{\mathbf{D}}$, and $\boldsymbol{K}_{\mathbf{W}}$), at each cross-validation fold, we adopted a maximum likelihood method described by Cuevas et al. (2019).

3.2.4. Statistical models

The merit of including additive effects ($\boldsymbol{K}_{\mathbf{A}}$), dominance deviation ($\boldsymbol{K}_{\mathbf{D}}$), GE interaction ($\boldsymbol{K}_{\mathbf{AE}}$ and $\boldsymbol{K}_{\mathbf{DE}}$), and envirotyping-based kinships ($\boldsymbol{K}_{\mathbf{W}}$, $\boldsymbol{K}_{\mathbf{AW}}$, and $\boldsymbol{K}_{\mathbf{DW}}$) to estimate reaction norms in MET was assessed using 5 WGP models. A description of each model structure is given below.

Model 1: Main additive effect model (EA): The main additive effect model (*EA*) is our benchmark baseline; it is also the simplest modeling structure for WGP in multi-environment trials, following:

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \boldsymbol{Z}_{\mathbf{E}}\boldsymbol{\beta} + \boldsymbol{Z}_{\mathbf{A}}\boldsymbol{u}_{\mathbf{A}} + \boldsymbol{\varepsilon} \quad \text{(Eq.1)}$$

where $\boldsymbol{y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n]'$ are the vectors of observations collected in each of the q environments with p hybrids, and $\mathbf{1}\boldsymbol{\mu} + \boldsymbol{Z}_{\mathbf{E}}\boldsymbol{\beta}$ is the general mean and the fixed effect of the environments with the incidence matrix $\boldsymbol{Z}_{\boldsymbol{E}}$. Genetic variations are modeled by the main additive effects ($\boldsymbol{u}_{\mathbf{A}}$), with $\boldsymbol{u}_{\mathbf{A}} \sim N(\mathbf{0}, \boldsymbol{J}_{\boldsymbol{q}} \otimes \boldsymbol{K}_{\mathbf{A}}\boldsymbol{\sigma}^2_{\mathbf{A}})$, where $\boldsymbol{Z}_{\mathbf{A}}$ is the incidence matrix for additive effects (absence=0, presence=1), $\boldsymbol{J}_q$ is a q×q matrix of 1s and $\boldsymbol{\sigma}^2_{\mathbf{A}}$ is the variance component for additive effects and $\otimes$ denotes the Kronecker Product. Residual deviations ($\boldsymbol{\varepsilon}$) were assumed as $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{I}_{\boldsymbol{n}}\boldsymbol{\sigma}^2)$, where n is the number of genotype-environment observations.

Model 2: Main Additive plus Dominance effects (EAD): Model EAD (Eq. 2) is a version of model 1 that includes the dominance deviation effects, as follows:

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \boldsymbol{Z}_{\mathbf{E}}\boldsymbol{\beta} + \boldsymbol{Z}_{\mathbf{A}}\boldsymbol{u}_{\mathbf{A}} + \boldsymbol{Z}_{\mathbf{D}}\boldsymbol{u}_{\mathbf{D}} + \boldsymbol{\varepsilon} \quad \text{(Eq.2)}$$

where $\boldsymbol{Z}_{\mathbf{D}}$ is the incidence matrix for dominance effects. Note that $\boldsymbol{Z}_{\mathbf{A}}$ and $\boldsymbol{Z}_{\mathbf{D}}$ are the same incidence matrix for genotypic effects. However, we put the respective acronyms A and D to facilitate the understanding that we are modeling two different genetic-based sources: additive random variation (as described in 1), plus dominance random variation ($\boldsymbol{u}_{\mathbf{D}}$), with $\boldsymbol{u}_{\mathbf{D}} \sim N(\mathbf{0}, \boldsymbol{J}_{\boldsymbol{q}} \otimes \boldsymbol{K}_{\mathbf{D}}\boldsymbol{\sigma}^2_{\mathbf{D}})$, where $\boldsymbol{\sigma}^2_{\mathbf{D}}$ is the variance component for dominance deviation effects.

Model 3: Main Effect EAD plus GE deviation (EAD+GE): The third model (EAD+GE, Eq. 3) is an update of the model (2) accounting for main effects ($\boldsymbol{u}_{\mathbf{A}}$ and $\boldsymbol{u}_{\mathbf{D}}$) plus genotype × environment interactions (GE). The inclusion of two multiplicative effects modeled these GE effects, one for additive × environment (AE = $\boldsymbol{u}_{\mathbf{AE}}$) interaction and a second for dominance × environment (DE = $\boldsymbol{u}_{\mathbf{DE}}$) interaction:

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \boldsymbol{Z}_{\mathbf{E}}\boldsymbol{\beta} + \boldsymbol{Z}_{\mathbf{A}}\boldsymbol{u}_{\mathbf{A}} + \boldsymbol{Z}_{\mathbf{D}}\boldsymbol{u}_{\mathbf{D}} + \boldsymbol{u}_{\mathbf{AE}} + \boldsymbol{u}_{\mathbf{DE}} + \boldsymbol{\varepsilon} \quad \text{(Eq.3)}$$

where $\boldsymbol{u}_{\mathbf{AE}} \sim N(\mathbf{0}, \boldsymbol{K}_{\mathbf{AE}}\boldsymbol{\sigma}^2_{\mathbf{AE}})$ and $\boldsymbol{u}_{\mathbf{DE}} \sim N(\mathbf{0}, \boldsymbol{K}_{\mathbf{DE}}\boldsymbol{\sigma}^2_{\mathbf{DE}})$, where $\boldsymbol{K}_{\mathbf{AE}} = \boldsymbol{Z}_{\mathbf{E}}\boldsymbol{I}_{\mathbf{q}}\boldsymbol{Z}'_{\mathbf{E}} \odot \boldsymbol{Z}_{\mathbf{A}}\boldsymbol{K}_{\mathbf{A}}\boldsymbol{Z}'_{\mathbf{A}}$ and $\boldsymbol{K}_{\boldsymbol{DE}} = \boldsymbol{Z}_{\mathbf{E}}\boldsymbol{I}_{\boldsymbol{q}}\boldsymbol{Z}'_{\mathbf{E}} \odot \boldsymbol{Z}_{\mathbf{D}}\boldsymbol{K}_{\mathbf{D}}\boldsymbol{Z}'_{\mathbf{D}}$, where $\boldsymbol{\sigma}^2_{\mathbf{AE}}$ and $\boldsymbol{\sigma}^2_{\mathbf{DE}}$ are the variance components for AE and DE interaction effects, respectively, as suggested by Jarquín et al. (2014), Lopez-Cruz et al. (2015) and Bandeira e Souza et al. (2017); $\boldsymbol{I}_{\boldsymbol{q}}$ is an identity matrix denoting a lack of environmental relatedness, and $\odot$ denotes the Hadamard product.

Model 4: Main Effect EAD with Main Envirotype Information (EADW): The next two models are updates of models 2 and 3, including non-genetic information (W) from envirotyping data. Jarquín et al. (2014) introduced a strategy to integrate these data in WGP by using environmental covariables to estimate an environmental relatedness kinship ($\boldsymbol{K_W}$) for $q \times q$ environments. Thus, the objective of including the W effects is bridging the gap between the pure genomic information and phenotypic variation observed across the environments. In this context, we tested the incorporation of some envirotype-phenotype relations as main effects (model 4, equation 4) and for GE effects (model 5, equation 5 in the next subsection).

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \boldsymbol{Z_E\beta} + \boldsymbol{Z_A u_A} + \boldsymbol{Z_D u_D} + \boldsymbol{u_W} + \boldsymbol{\varepsilon} \quad \text{(Eq.4)}$$

where $\boldsymbol{u_W} \sim N(\mathbf{0}, \boldsymbol{J_p} \otimes \boldsymbol{K_W}\boldsymbol{\sigma^2_W})$, $\boldsymbol{\sigma^2_W}$ is the variance component related to the variation due to envirotype data, and $\boldsymbol{J_p}$ is a matrix of 1s with dimension $p \times p$.

Model 5: Main Effect EADW plus Reaction Norm for GE (EADW+GW): The last model (EADW+GW) is an update of (Eq. 3) reaction norm variation based on the genomic × envirotype effects (GW). In model EADW+GW, we perform the traditional genomic-enabled reaction norm, but discriminating the reaction norm due to additive effects (AW = $\boldsymbol{u_{AW}}$) and dominance deviations (DW = $\boldsymbol{u_{DW}}$) as follows:

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \boldsymbol{Z_E\beta} + \boldsymbol{Z_G u_A} + \boldsymbol{Z_G u_D} + \boldsymbol{u_W} + \boldsymbol{u_{AW}} + \boldsymbol{u_{DW}} + \boldsymbol{\varepsilon} \quad \text{(Eq.5)}$$

where $\boldsymbol{u_{AW}} \sim N(\mathbf{0}, \boldsymbol{K_{AW}}\boldsymbol{\sigma^2_{AW}})$ and $\boldsymbol{u_{DW}} \sim N(\mathbf{0}, \boldsymbol{K_{DW}}\boldsymbol{\sigma^2_{DW}})$, with $\boldsymbol{K_{AW}} = \boldsymbol{Z_E K_W Z'_E} \odot \boldsymbol{Z_A K_A Z'_A}$ and $\boldsymbol{K_{DE}} = \boldsymbol{Z_E K_W Z'_E} \odot \boldsymbol{Z_D K_D Z'_D}$, where $\boldsymbol{\sigma^2_{AW}}$ and $\boldsymbol{\sigma^2_{DW}}$ are the variance components for AW and DW interaction effects. Note that in (Eq. 3) we described how to estimate the GE kernels using the Hadamard product between fixed environment and genomic sources. At that point, the GE kernels are estimated using a block diagonal matrix of genomic effects. In contrast, now in (Eq. 5) we replace the identity matrix $\boldsymbol{I_q}$ with the envirotype-informed kinship $\boldsymbol{K_W}$, in which a dense matrix models GW kernels. Then it is possible to assume that now there are different relationship levels between genotypes across environments according to the envirotyping-based kinships.

3.2.5. Assessing prediction accuracy by cross-validation

In this study, three cross-validation schemes were used to evaluate the predictive ability (PA) of each model-kernel method combination. The first scheme aimed to quantify the accuracy of WGP models when predicting new genotypes within the experimental network, i.e., maize hybrids not yet tested in any environment. This validation scheme is called CV1, which was run 50 times using random samplings of 70% of phenotypic information, while the remaining data were predicted. The second scheme aimed to quantify the predictability of WGP models under sparse experimental network conditions. In contrast to CV1, in this scheme (CV2), the sparse phenotypic information of one genotype not evaluated in one environment but evaluated across other different environments can help increase PA. For this scheme, 50 random repetitions were also used, but sampling 70% of the phenotypic information (genotype-environment combinations) as the training population, and the remaining 30% as the test population. Finally, the third scheme aimed to quantify WGP models' ability to predict new environmental conditions. For this, we adopted a leave-one-environment-out scheme (CV0).

PAs were evaluated at two levels: (1) the model level, in which we computed Pearson's correlation between observed ($\boldsymbol{y}$) and predicted values ($\hat{\boldsymbol{y}}$) and, finally, for CV0, the general average of these correlations; and (2) the genotype level, in which we computed the predictability related to the observed and predicted performance of a genotype in all environments. The standard error (SE) was computed for each average PA following $\mathrm{SE} =$

$\mathrm{SD} \times \sqrt{\frac{1}{\mathrm{n}} + \frac{\mathrm{n_2}}{\mathrm{n_1}}}$ where SD is the standard deviation of the correlations, $n = pq$ for p genotypes (hybrids) and q environments, and n_1 and n_2 denote the size of the training and testing populations for each CV scheme (Bouckaert and Frank 2004).

3.2.6. Hierarchical Bayesian modeling

Genomic predictions were performed using the Bayesian Genotype plus Genotype × Environment (*BGGE*) package (Granato et al. 2018). This package contains a function called "BGGE()" in which it solves mixed linear models through Hierarchical Bayesian Modeling. Below, we briefly describe the main distributions and priors used by this package. First, each variance-covariance matrix (*K*) is reparametrized using an eigen-decomposition procedure suggested by De Los Campos et al. (2010), $\boldsymbol{K = USU'}$ where $\boldsymbol{S}$ is a diagonal matrix with n non-zero eigenvalues and $\boldsymbol{U}$ is an orthogonal matrix with eigenvectors. Hence, an orthogonal transformation suggested by Cuevas et al (2014). In this transformation, the phenotypic parametrization is represented as $\boldsymbol{d = U'y}$, and any kernel-based random effect ($\boldsymbol{b = U'u}$) and error variation ($\mathbf{e} = \boldsymbol{U'\varepsilon}$) is now represented into a reparametrized normal distribution as $\boldsymbol{b} \sim N(0, \boldsymbol{U'KU\sigma_u^2}) = N(0, \boldsymbol{S\sigma_u^2})$ and $\mathbf{e} \sim N(0, \mathbf{U'U\sigma}_\varepsilon^2) = N(0, \mathbf{I}\sigma_\varepsilon^2)$.Both process are employed to increase the computational efficiency of the further steps. Thus, the distribution of the transformed data is now given by:

$$f(\boldsymbol{d} \mid \boldsymbol{b}, \boldsymbol{\sigma}_\varepsilon^2) = \prod_{i=1}^{n} N\left(d_i \mid b_i, \sigma_\varepsilon^2\right)$$

where the acronym *i* now denotes each random effect (variance-covariance) considered (e.g, for additive, dominance, envirotyping data). As this Bayesian linear model assumes $p(\boldsymbol{u} \mid \sigma_u^2) = N(\boldsymbol{u} \mid 0, \boldsymbol{K\sigma}_u^2)$, the conditional of any b_i is given as $p(b_i \mid \sigma_u^2) = N(b_i \mid 0, \sigma_u^2 s_i)$, where s_i are the eigenvalues. Thus, the BGGE package assumes that conjugate prior distribution of σ_u^2 and σ_ε^2 is given by inverse chi-squared with, $\mathrm{p}(\sigma_u^2) \sim \chi^{-2}(\nu_u, Sc_u)$ and $\mathrm{p}(\sigma_\varepsilon^2) \sim \chi^{-2}(\nu_\varepsilon, Sc_\varepsilon)$ respectively, in which ν_u and ν_ε denotes the degree of freedom, and Sc_u and Sc_ε the scale factors for *u* and *e*. Then, the joint posterior distribution ($\boldsymbol{J = b}, \sigma_u^2, \sigma_\varepsilon^2$), given the parameters ($\boldsymbol{P = d}, \nu_u, \nu_\varepsilon, Sc_u, Sc_\varepsilon$ and *S*) is:

$$p(\boldsymbol{J} \mid \boldsymbol{P}) \propto \left\{\prod_{i=1}^{n} N\left(d_i \mid b_i, \sigma_\varepsilon^2\right) N(b_i \mid 0, \sigma_u^2 s_i)\right\} \times \chi^{-2}(\sigma_u^2 \mid \nu_u, \nu_u Sc_u) \times \chi^{-2}(\sigma_\varepsilon^2 \mid \nu_\varepsilon, \nu_\varepsilon Sc_\varepsilon)$$

Finally, BGGE use the Markov Chain Monte Carlo (MCMC) procedure to generate the conditional distributions through a Gibbs sampler. Details of this package and functions are we given deeply in Granato et al (2018). For all combinations of model and kernel methods tested in this study, the MCMC through a Gibbs sampler were performed for 10,000 iterations with the first 1,000 cycles removed as burn-in with thinning equal to 2.

3.2.7. Software and data availability

All analyses were conducted using R statistical software (R Core Team, 2019). Data and codes are available at https://github.com/gcostaneto/KernelMethods [verified 25th May 2021].

3.3. RESULTS

3.3.1. Differences in explaining the sources of variation

When including new sources of variation, as well as when modeling these sources by different kernels, it is expected that differences in the proportion of variance explained by WGP can be detected (Figure 1 and Supplementary Tables S2 and S3).

Additive effects (A) are the main source of genomic variation in all models. In the EA model, the A effects are best explained by the DK (HEL set) and GK (USP set) kernels. The inclusion of D effects increased the genomic prediction ability to explain phenotypic variation. In the additive-dominant model (EAD), the use of GK was more efficient in capturing dominance effects (D, light blue color in Fig1) in both data sets. For the HEL set, using the DK kernel to model dominance effects resulted in an increase in the additive genomic variance and a reduction in the residual variance. In the USP set, the dominance effects were better modeled by GK, while the traditional GB kernel better captured the total genomic effects (A + D).

The biggest differences between kernel methods were observed in the most complex models involving GE interaction and envirotyping data. In GB, it is possible to verify that the interaction between home and environment (DE, light green color in Figure 1) was an important variation to describe the phenotypic variance in the tests. In general terms, in models with GE interaction (GE = AE + DE), the GK kernel was more efficient in explaining the main additive and dominant effects in both sets. However, for the HEL set, the DK kernel was more efficient in reducing the residual variance by capturing the effects of additive × environment interaction (green color in Fig1) better. Upon comparing GB with GK and DK, these last two kernels increased the variance explained by the genomic prediction model.

Reaction norm models tend to capture a large amount of variance and drastically reduce the residual error. The inclusion of the main effect of envirotyping-informed relationships (W, orange color in Fig1) produced similar results as those observed for models with EAD+GE effects. There was a drastic reduction in the residual variation of EADW benchmarked with the EAD model for all models and kernels. In models involving the reaction norm for the effects of GW=AW+DW (model 4, EADW+GW) for the HEL set, there was an increase in the capacity of the models to explain D effects using GB and GK, especially in the reaction norm for dominance (DW, purple colors in Fig1) using GB. When a reaction norm for AW+DW is integrated, most of the phenotypic variance is explained by non-genomic effects from W. For the USP data set, the DK kernel was more conservative in modeling W effects; in contrast, it was better able to model main A, D and AW interaction. Despite this, it was the model whose proportion of residual variance was highest.

3.3.2. Computational efficiency

The processing time of the models is a key issue for their widespread use in WGP-MET. All the benefits of complex models involve different genomic and environmental structures, but are computationally costly and unlikely to achieve wide approval by plant breeders. Here we calculated the processing time of a Bayesian Markov

Chain involving 10,000 iterations for each model and kernel method combination involving all the phenotypic data of *p* hybrids in *q* environments in both data sets (Table 2).

As expected, more complex models tend to take more processing time, which can range from 47 seconds (EA) to 330 s (EADW + GW) in smaller data sets like HEL ($q = 5, p = 247$), or 660 s (EA) and up to 4368 s (EADW + GW) in larger data sets, such as USP ($q = 8, p = 570$). In the simplest model (EA), GB is faster than GK and DK in both sets. However, as the complexity of the models increases, GB becomes increasingly slower than DK and GK. The DK kernel is significantly faster than GB and GK, even running the same Markov chain in 39% less time than GB. It is possible to run more complex models using GK and DK in similar time as simpler models using GB. For the USP set, it was possible to see that GK was faster than DK under most scenarios, even running a more complex model with environmental data and additive-dominant effects (EADW) at almost the same speed as a traditional GE interaction model via GB.

3.3.3. Accuracy in the HEL set

Table 3 presents the results from the three cross-validation schemes (CV1, CV2, and CV0) for each model-kernel method combination in the HEL set. For CV1 and CV2, the simplest model structures (EA and EAD) were unable to produce an accurate prediction of grain yield concerning the most complex models (EAD+GE, EADW, and EADW+GW). The inclusion of D effects (EAD) led to an increase in PA for CV1 schemes. In contrast, there was a reduction in PA for CV2 when the main D effects were included (EAD model). For the EA and EAD models, there were no great differences in PA between the three-kernel method adopted in both CV1 and CV2 schemes.

For the most complex models, however, there was a drastic difference between kernel methods. For model 3 (EAD+GE), the GB was unable to reproduce the GE effects of AE and DE interactions. On the other hand, the GK and DK kernels satisfactorily exploited the AE+DE effects, translating model complexity in PA, with increments ranging from 54% (DK at CV1) to 73% (GK at CV2) compared to the baseline EA model. EAD+GE outperformed the best GB-based models for both CV1 and CV2 schemes (EADW, with r = 0.832 for CV1 and r = 0.839 for CV2) based on GK (r=0.871 in CV1 and r = 0.892 in CV2). The reaction norm models (EADW and EADW+GW) using DK were similar to the GB models for both CV1 and CV2, but it took less computational time to run them (see Table 2).

The results for CV0 are presented in the last part of Table 3. As expected, the PA values were higher than CV1 and CV2 because this scheme uses much more phenotypic information than the other schemes. However, in CV0, it faced the problem of predicting the performance of the hybrids in an entirely new environment. All GK- and DK-based models outperformed the GB models. The use of complex structures from environmental data was useful for GB kernels, but in contrast, modeling structures based on GK and DK led to a similar result just by the inclusion of dominance effects (EAD for DK) and GE interaction (EAD for GK and DK). In summary, it was possible to achieve the same results for reaction norm GB using dominance effects or GE interaction in DK.

3.3.4. Accuracy in the USP set

Table 4 shows the results from the three cross-validation schemes (CV1, CV2, and CV0) for each model-kernel method combination in the USP set. As expected, the PA values were higher for CV0, followed by CV2 and CV1. In this last scheme, the inclusion of D effects led to an increment in PA for all kernels, except GK. As observed in the HEL set, model 3 (EAD+GE) based on GB was not satisfactory in exploring GE interaction. PA values were higher in models including non-genetic effects derived from envirotype data (EADW and EADW+GW) than in pure genomic models (EA, EAD, and EAD+GE). In CV1, the best GB model (EADW+GW) was the same as the EAD+GE model using GK and DK. This last kernel led to greater PA values when some envirotyping data were used (r = 0.822 for EADW and r = 0 .818 for EADW+GW).

The DK method was also efficient in exploring main D effects (r = 0.338 in EAD) and GE interaction (r=0.669 in EAD+GE, an increment of 54% compared to the EA model). However, in the CV2 scheme, it was possible to see how the DK method was efficient in providing a more computationally efficient approach that captures AE+DE effects better. Model EAD+GE based on DK achieved the highest PA value for all CV schemes (r= 0.891), while the best GB model (EADW+GW) had a PA value equal to r = 0.731. GK was also efficient in exploring genomic AE+DE effects (r = 0.733) and the inclusion of non-genomic reaction norm effects (r = 0.751). Finally, in CV0, it was possible to measure the models' ability to predict novel environments. The DK outperformed the GK and GB kernels and produced more precise predictions incorporating D, GE effects, and envirotyping data.

3.3.5. Resolution of genomic prediction for specific hybrids

Most studies involving WGP-MET only assess the accuracy of the models in predicting the entire data set over a specific cross-validation scenario, as presented in the previous sections. Here we introduce the concept of resolution of the WGP models by evaluating the models' ability to reproduce the phenotypic performance of specific maize hybrids within MET. The phenotypic data used as a training set in these models were obtained from (q-1) environments, where the one-environment-out is a novel growing condition in which the hybrid was not tested (CV0). Thus, the following results are a scenario in which maize breeders have already evaluated the genotypes over a MET but are interested in making predictions of the phenotypic performance of desirable target hybrids.

Figure 2 presents the PA values for specific hybrids (rows) (Fig 2A) and the typology (distribution pattern) of those predictions for each model-kernel method combination (Fig 2B) and each data set (HEL and USP). For both data sets, it is possible to observe that different model-kernel method combinations can predict different hybrids (Fig 2A). The same hybrid can be well predicted by a simpler model, but not predicted by a more complex model. In contrast, the inclusion of more complex structures such as the reaction norm may not always lead to a better description of a target hybrid. For this reason, we analyzed the typology of those predictions (Fig 2B), aiming to observe which model-kernel method combinations are more accurate in reproducing most of the hybrids.

The simplest modeling structures (EA and EAD) are incapable of reproducing the performance of almost 50% of the hybrids in both sets (green colors in Fig 2A and red colors in Fig 2B). For those models, the use of any kernel method has led to almost the same result. The greatest differences are observed when genotype × environment (GE) interaction effects are included (EAD+GE). GB was the worst kernel method for exploring the GE effects and translating them into a higher resolution of WGP. GK was the best kernel method, where the most frequent type of PA was equal to up to 0.25, but still above to 0.50 (blue color in Fig. 2A; yellow bars in Fig. 2B). DK was very efficient in the USP set, but it was not observed in the HEL set. An explanation of that may be that the DK was overfitted for the HEL set, with a smaller sample of phenotypic data.

The higher resolution of WGP was achieved by the inclusion of envirotyping-based data to model main environmental effects (EADW) or reaction norm variation (EADW+GW) into the additive-dominance models. For the HEL set, the EADW model with DK was the best modeling approach, with the highest PA values (blue and dark blue colors in Fig 2A) and with less than 4% of the hybrids not well predicted (values above 0, in red bars in Fig 2B). The most frequent PA type had values from 0.26 to 0.50 (green colors in Fig 2 B). For the USP set, all kernel methods drastically improved the resolution of WGP for both EADW and EADW+GW models (Fig 2A). The model-kernel method differences were better represented in the EADW and EADW+GW panels in Fig 2B. GK outperformed GB in increasing the frequency of higher PA values (green and blue bars in Fig 2B). In the same way, DK outperformed GK for both EADW and EADW+GW models. The typology of the EADW + GW model based on DK presents negative PA values at a frequency of less than 3%. Conversely, the predominant type is between 0.26 and 0.50 (~ 50% of the hybrids) and values between 0.51 and 0.75 (~ 20% of the hybrids).

3.3.6. Accuracy trends for novel environments

Based on the results presented in the previous section, we selected six model-kernel method combinations to be jointly evaluated in terms of their capacity to predict novel environments (Fig 3). It was difficult to determine which models were better in the less predictable environment (S4, from the HEL set). However, as the predictability of environments increases, it is possible to better understand how different kernel methods and models can reproduce the phenotypic information of a novel environmental condition. The use of the main effect additive-dominant GB (GB-EAD, red dotted line in Fig 3) was the most unstable framework in CV0. In contrast, the incorporation of envirotypic data (GB-EADW, the green dotted line in Fig 3) was responsible for increasing the PA for less predictable environments and stabilizing the response of the additive-dominant model in reproducing novel environments.

The GB-EADW model had a similar performance as models DK-EADW (solid green line in Fig 3) and GK-EAD+GE (golden dashed line in Fig 3). In contrast to the other models, the inclusion of the AW and DW effects (blue lines) combined with the GK (dashed blue line) and DK (solid blue line) kernels increased the PA for all environments, especially for E2, E3, and E6, corresponding to ideal N conditions in Piracicaba in 2016, low N conditions in Anhumas in 2016, and ideal N conditions in Piracicaba in 2017. Between these two reaction norm models, the GK outperformed the DK and achieved higher PA values for most of the environments.

3.4. DISCUSSION

In this study, we presented the first report on (1) the joint modeling of additive and dominance effects with reaction norm variation; (2) the modeling of these effects performed by Gaussian Kernel and Deep Kernel; and (3) their comparison with benchmark GBLUP-based modeling. We reported that the Gaussian Kernel and Deep Kernel outperformed GBLUP in reducing the computational time and increased predictive ability for all testing scenarios in tropical maize. Below we discuss how the use of dominance effects and envirotyping-aided reaction norm modeling is the main bottleneck for increasing predictive ability in GBLUP-based models over MET. In addition, we suggest that the Gaussian Kernel is the best alternative to model dominance variation and translate it into predictive ability gains. Finally, we discussed that Deep Kernels also have greater potential to be used on large-scale genomics and

"enviromics" (the core of envirotyping-based big data). They are faster, capture better additive and dominance effects, and have greater predictive accuracy than other kernels under several prediction conditions faced by maize breeders in the development of hybrids.

3.4.1. Importance of Dominance Effects in GBLUP

In all the predicted scenarios evaluated (CV1, CV2, and CV0), the models integrating both genomic and envirotyping data tended to have better ability to reproduce the phenotypic performance of maize hybrids. As reported in other studies in plants, the inclusion of dominance effects in traditional WGP-MET resulted in increased predictive accuracy in models based on GBLUP in front to other methods. Azevedo et al. (2015) showed that GBLUP-based models outperform methods such as Ridge Regression, (e.g., BayesA, Bayes/LASSO) in modeling A+D genetic effects in simulated populations. Dias et al (2018) demonstrated that GBLUP models containing A+D effects doubled the predictive capacity for grain yield in maize under diverse environmental conditions, such as environments with limited water availability (i.e., drought-stress screening trials). In a study based on simulations for a pine breeding population, De Almeida Filho et al. (2016) suggests that the gains predictive capacity obtained by the A+D model compared to the model based only on A are only relevant if the D effects explain at least 20% of the phenotypic variation. Here we show that not only the main D effect but their interaction with the environment (D+DE and D+DW) were responsible for 25% to 40% of the phenotypic variation in both maize sets. This can explain the excellent results found in this study, especially when the GK and DK kernels, better able to capture such effects, are used in the prediction. Despite the aforementioned factors, the inclusion of D effects is essential for the accurate modeling of phenotypic variation in species with some degree of heterosis (Technow et al. 2014; Dos Santos et al. 2016), such as in this study using F_1 single-crosses.

For the prediction of new environments (CV0) in our study, we observed a leap in accuracy from 0.402 to 0.558 (+ 39%) in HEL, and from 0.335 to 0.425 in USP (+ 27%), which can be explained by the fact that dominance effects are important to control the stability and adaptability of single maize hybrids, making them more predictable. However, without any envirotyping data, the possible accuracy achieved by those models for grain yield is limited. This trait is quantitatively inherited, controlled by many genes of small effects, and has strong epistatic relationships with several other traits highly influenced by the environment, such as the number of grains per ear and ear size. In this sense, within MET, the use of dominance effects produced by a covariance-based kinship may not be enough. Details about how dominance effects were better modeled using Gaussian kernel and Deep kernel are discussed in the next sections.

3.4.2. Envirotyping data are a limit breaker for MET GBLUP

For the prediction of novel maize hybrids, the greatest leap in accuracy in GBLUP was due to the ability to integrate the envirotyping information in the modeling of the reaction norm at the level of additive effects (AW) and dominance deviations (DW). This fact suggests that dominance effects are indispensable for a deep understanding of the genomic causes driving genomic × environment (GE) interaction for each hybrid. In the HEL data set, the models including only the main effects (EADW) had a performance similar to that of the models containing GW effects (EADW+GW). This can be explained by the fact that, in this data set, GE interaction was not as important as in USP;

therefore, the inclusion of envirotyping data was enough to adjust the genomic responses according to the degree of similarity between environments.

In contrast to the reaction norm models (EADW and EADW+GW), the GBLUP was not efficient in reproducing GE interactions in the models assuming that environments are not related (EAD and EAD+GE). Thus, the inclusion of envirotyping data (W and GW) may be the only alternative to breaking the limits of predictive ability achieved in MET-WGP employing the benchmark GBLUP kernel in maize. The prediction of novel environments is restricted to models including envirotyping data, even if the dominance effects are taken into account. However, despite the higher accuracy gains achieved by including W or GW effects, those models are computationally expensive and were outperformed by other kernel methods employing the same molecular and envirotyping data.

3.4.3. DK and GK better model interaction effects

In contrast to GBLUP, both Gaussian kernel and Deep kernel methods were successful in reproducing genomic × environment (GE) interaction, even in those models that assume that environments are not related. In the case of the Gaussian kernel, its higher efficiency in capturing interaction effects from intra-allelic (dominance) and whole GE interaction may be because such effects are better understood in terms of non-linear relationships and Euclidean distances, and not as linear covariances as given in GBLUP. The use of covariances to estimate an existing relationship between individuals has its origins in the work of VanRaden (2008), which focused on modeling pedigree and additive-genomic effects. On the other hand, the Gaussian kernel assumes a diagonal equal to 1.0 and an off-diagonal based on the Euclidean distance regulated by a bandwidth factor. Thus, the genetic sense of this matrix property for an F_1 hybrid individual is that the effects of dominance are highest within an individual. The relationship between individuals depends on the distance between the effects of intra-allelic interaction shared between related individuals. Similarly, the GE interaction corresponds to whole genomic effects being differentially activated/deactivated, for each genotype, as a function of the total existing environmental inputs (E →GE). The inclusion of envirotyping data leads to a deeper understanding of this dynamic, which is converted as a function of the known environmental inputs and of how a particular genomic response of different genotypes is distanced. On the other hand, the use of a Deep kernel seeks to model the genomic relationship matrix based on emulating hidden layers capable of capturing different levels of depth of the same genomic effect. In this work, we introduced simultaneous and independent modeling of hidden layers for additive and dominance effects, which capture different relationship patterns between individuals based on the phenotypic information provided in the training set. Unlike the Gaussian kernel, the diagonal elements of the Deep kernel are not identical (Supplementary Figures S1-S3), for they express heterogeneous variances of the genetic and environmental effects. This may be why the Gaussian Kernel overcame the Deep kernel in the EAD + GE models in CV1 and CV0. As for CV2, the Deep kernel benefited from the fact that the borrowing of phenotypic information across multiple environments helped shape the covariance structure carried out by the hidden layers.

3.4.4. Approaching envirotype-to-phenotype modeling

In this work, we also introduce the use of the non-linear methods (Gaussian kernel and Deep kernel) in the modeling of genomic and non-genomic (environmental) kinships. Since the first report of a genomic-enabled prediction considering the reaction norm, as proposed by Jarquín et al. (2014), the environmental relationship kernel ($\boldsymbol{K_W}$) was modeled by the benchmark GBLUP approach. Here we show that the similarity among environments is better modeled in terms of Gaussian processes than the covariance, as traditionally done in GBLUP for modeling the dominance effects. The use of Deep Kernels is also favored because the environmental kinship accounted for based on environmental distances due to non-genomic covariables, is regulated by the phenotypic information in the training set, thereby resulting in more accurate modeling of the envirotype-to-phenotype (E-to-P) dynamics in the prediction of new genotypes and new environments. This stems from the fact that indirectly, in the phenotype provided in the training population set, there is a genomic similarity relationship that determines the E-to-P relationship, part of which is captured by the genomic kernels and the rest by the environmental kernel. Despite these advantages, both the Gaussian Kernel and the Deep Kernel are faster, more accurate, and have a better resolution in predicting specific genotypes than the GBLUP models. In contrast with other reaction norm proposals, such as the use of factorial regression to dissect E-to-P in secondary traits, the use of crop growth models and the use of envirotyping data to group environments and target WGP models, here we can use in a faster way the large-scale envirotypic data (*enviromics*) to explore alternative kinships across the benchmark genomic data.

3.4.5. Large-scale genomics and enviromics with GK or DK

We demonstrate that the use of several sources of genomic variation (additive + dominance + GE interaction) guided by envirotyping is useful for increasing model accuracy. The use of the Gaussian kernel or Deep kernel makes it possible to capitalize on these effects, translating them into a drastic increase in predictive ability, reduction of computational processing time, a greater explanation of phenotypic variation, and reduction of residual variation. New sources of non-genomic variation can be incorporated into WGP models through GK or DK to seek greater gains in predictive ability under WGP-MET, as they are efficient in dealing with large-scale data. Here we also show that the use of environmental information through distribution quantiles is efficient for characterizing environments and, consequently, gives the kernels the ability to reproduce environmental similarities that can be explored in prediction. The field of large-scale enviromics still has a long pathway, but strategies that integrate E-to-P modeling are a bottleneck to overcome in genomic prediction, which benchmark GBLUP models are unable to achieve.

REFERENCES

Acosta-Pech R, Crossa J, de los Campos G, Teyssèdre S, Claustres B, Pérez-Elizalde S, *et al.* (2017). Genomic models with genotype × environment interaction for predicting hybrid performance: an application in maize hybrids. *Theor Appl Genet* **130**: 1431–1440.

Allen RG, Pereira LS, Raes D, Smith M (1998). *Crop Evapotranspiration (guidelines for computing crop water requirements)*. FAO Irrigation and Drainage Paper Nº 56.

Alves FC, Granato ÍSC, Galli G, Lyra DH, Fritsche-Neto R, De Los Campos G (2019). Bayesian analysis and prediction of hybrid performance. *Plant Methods* **15**: 1–18.

Azevedo CF, de Resende MDV, e Silva FF, Viana JMS, Valente MSF, Resende MFR, *et al.* (2015). Ridge, Lasso and Bayesian additive-dominance genomic models. *BMC Genet* **16**: 1–13.

Basnet BR, Crossa J, Dreisigacker S, Pérez-Rodríguez P, Manes Y, Singh RP, *et al.* (2019). Hybrid Wheat Prediction Using Genomic, Pedigree, and Environmental Covariables Interaction Models. *Plant Genome* **12**: 1–13.

Bernardo R (1994). Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Sci* **34**: 20–25.

Bernardo R (1996). Testcross additive and dominance effects in best linear unbiased prediction of maize single-cross performance. *Theor Appl Genet* **93**: 1098–1102.

Bouckaert RR, Frank E (2004). Evaluating the replicability of significance tests for comparing learning algorithms. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol 3056, pp 3–12.

Browning BL, Browning SR (2008). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**: 210–223.

Cho Y, Saul LK (2009). Kernel methods for deep learning. In: *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*,

Cooper M, Messina CD, Podlich D, Totir LR, Baumgarten A, Hausmann NJ, *et al.* (2014). Predicting the future of plant breeding: Complementing empirical evaluation with genetic prediction. *Crop Pasture Sci* **65**: 311–336.

Costa-Neto GMF, Morais Júnior OP, Heinemann AB, de Castro AP, Duarte JB (2020). A novel GIS-based tool to reveal spatial trends in reaction norm: upland rice case study. *Euphytica* **216**: 1–16.

Crossa J, Martini JWR, Gianola D, Pérez-Rodríguez P, Jarquin D, Juliana P, *et al.* (2019). Deep Kernel and Deep Learning for Genome-Based Prediction of Single Traits in Multienvironment Breeding Trials. *Front Genet* **10**: 1–13.

Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, *et al.* (2017). Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends Plant Sci* **22**: 961–975.

Cuevas J, Crossa J, Soberanis V, Perez-Elizalde S, Perez-Rodriguez P, de Los Campos G, *et al.* (2016). Genomic Prediction of Genotype x Environment Interaction Kernel Regression Models. *Plant Genome* **9**: 1–20.

Cuevas J, Granato I, Fritsche-Neto R, Montesinos-López OA, Burgueño J (2018). Genomic-Enabled Prediction Kernel Models with Random Intercepts for Multi-environment Trials. **8**: 1347–1365.

Cuevas J, Montesinos-López O, Juliana P, Guzmán C, Pérez-Rodríguez P, González-Bucio J, *et al.* (2019). Deep Kernel for genomic and near infrared predictions in multi-environment breeding trials. *G3 Genes, Genomes, Genet* **9**: 2913–2924.

Cuevas J, Pérez-Elizalde S, Soberanis V, Pérez-Rodríguez P, Gianola D, Crossa J (2014). Bayesian genomic-enabled prediction as an inverse problem. *G3 Genes, Genomes, Genet* **4**: 1991–2001.

De Almeida Filho JE, Guimarães JFR, E Silva FF, De Resende MDV, Muñoz P, Kirst M, *et al.* (2016). The contribution of dominance to phenotype prediction in a pine breeding and simulated population. *Heredity (Edinb)* **117**: 33–41.

De Los Campos G, Gianola D, Rosa GJM, Weigel KA, Crossa J (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet Res (Camb)* **92**: 295–308.

Dias KODG, Gezan SA, Guimarães CT, Nazarian A, Da Costa E Silva L, Parentoni SN, *et al.* (2018). Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. *Heredity (Edinb)* **121**: 24–37.

Ferrão LF V, Marinho CD, Munoz PR, Resende MFR (2020). Improvement of predictive ability in maize hybrids by including dominance effects and marker × environment models. *Crop Sci*.

Gianola D, Fernando RL, Stella A (2006). Genomic-Assisted Prediction of Genetic Value with Semiparametric Procedures. *Genetics* **173**: 1761–1776.

Gianola D, Morota G, Crossa J (2014). Genome-enabled prediction of complex traits with kernel methods: What have we learned? In: *Proceedings, 10th World Congress of Genetics Applied to Livestock Production*,

Gianola D, Okut H, Weigel KA, Rosa GJM (2011). Predicting complex quantitative traits with Bayesian neural networks: A case study with Jersey cows and wheat. *BMC Genet* **12**: 1–14.

Gianola D, Van Kaam JBCHM (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* **178**: 2289–2303.

González-Camacho JM, de los Campos G, Pérez P, Gianola D, Cairns JE, Mahuku G, *et al.* (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor Appl Genet* **125**: 759–771.

Granato I, Cuevas J, Luna-Vázquez F, Crossa J, Montesinos-lópez O, Burgueño J, *et al.* (2018). BGGE : A New Package for Genomic-Enabled Prediction Incorporating Genotype x Environment Interaction Models. **8**: 3039–3047.

Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, *et al.* (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor Appl Genet* **127**: 595–607.

Jiang Y, Reif JC (2015). Modeling epistasis in genomic selection. *Genetics* **201**: 759–768.

Lehermeier C, Krämer N, Bauer E, Bauland C, Camisan C, Campo L, *et al.* (2014). Usefulness of multiparental populations of maize (Zea mays L.) for genome-based prediction. *Genetics* **198**: 3–16.

Li X, Guo T, Mu Q, Li X, Yu J (2018). Genomic and environmental determinants and their interplay underlying phenotypic plasticity. *Proc Natl Acad Sci* **11**: 6679–6684.

Lopez-Cruz M, Crossa J, Bonnett D, Dreisigacker S, Poland J, Jannink J-L, *et al.* (2015). Increased Prediction Accuracy in Wheat Breeding Trials Using a Marker × Environment Interaction Genomic Selection Model. *G3* **5**: 569–582.

Ly D, Huet S, Gauffreteau A, Rincent R, Touzy G, Mini A, *et al.* (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.

Martini JWR, Wimmer V, Erbe M, Simianer H (2016). Epistasis and covariance: how gene interaction translates into genomic relationship. *Theor Appl Genet* **129**: 963–976.

Meuwissen THE, Hayes BJ, Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819–1829.

Millet EJ, Kruijer W, Coupel-Ledru A, Alvarez Prado S, Cabrera-Bosquet L, Lacube S, *et al.* (2019). Genomic prediction of maize yield across European environmental conditions. *Nat Genet* **51**: pages952–956.

Morais Júnior OP, Duarte JB, Breseghello F, Coelho ASG, Magalhães Júnior AM (2018). Single-step reaction norm models for genomic prediction in multienvironment recurrent selection trials. *Crop Sci* **58**: 592–607.

Morota G, Gianola D (2014). Kernel-based whole-genome prediction of complex traits: A review. *Front Genet* **5**: 1–13.

Neal R (1996). Bayesian Learning for Neural Networks. *Lect notes stat -new york- springer verlag-* **1**.

Pérez-Elizalde S, Cuevas J, Pérez-Rodríguez P, Crossa J (2015). Selection of the Bandwidth Parameter in a Bayesian Kernel Regression Model for Genomic-Enabled Prediction. *J Agric Biol Environ Stat* **20**: 512–532.

Pérez-Rodríguez P, Gianola D, González-Camacho JM, Crossa J, Manès Y, Dreisigacker S (2012). Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3 Genes, Genomes, Genet* **2**: 1595–1605.

Porker K, Coventry S, Fettell NA, Cozzolino D, Eglinton J (2020). Using a novel PLS approach for envirotyping of barley phenology and adaptation. *F Crop Res* **246**: 1–11.

R Core Team (2019). A language and environment for statistical computing. *R Found Stat Comput Austria* **Vienna**, **Au**.

Soltani A, Sinclair TR (2012). *Modeling physiology of crop development, growth and yield* (CAB International, Ed.). International, Wallingford: Cambridge.

Souza MB, Cuevas J, Couto EG de O, Pérez-Rodríguez P, Jarquín D, Fritsche-Neto R, *et al.* (2017). Genomic-Enabled Prediction in Maize Using Kernel Models with Genotype × Environment Interaction. *G3* **7**: g3.117.042341.

Sparks A (2018). nasapower: A NASA POWER Global Meteorology, Surface Solar Energy and Climatology Data Client for R. *J Open Source Softw* **3**: 1035.

Technow F, Schrag TA, Schipprack W, Bauer E, Simianer H, Melchinger AE (2014). Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. *Genetics* **197**: 1343–1355.

Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, *et al.* (2014). A powerful tool for genome analysis in maize: Development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics* **15**: 1–15.

VanRaden PM (2008). Efficient methods to compute genomic predictions. *J Dairy Sci* **91**: 4414–4423.

Vitezica ZG, Varona L, Legarra A (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics* **195**: 1223–1230.

Voss-Fels KP, Cooper M, Hayes BJ (2019). Accelerating crop genetic gains with genomic selection. *Theor Appl Genet* **132**: 669–686.

Wang X, Li L, Yang Z, Zheng X, Yu S, Xu C, *et al.* (2017). Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. *Heredity (Edinb)* **118**: 302–310.

Williams CKI (1998). Computing with infinite networks. *Adv Neural Inf Process Syst*: 295–301.

Wimmer V, Albrecht T, Auinger HJ, Schön CC (2012). Synbreed: A framework for the analysis of genomic prediction data using R. *Bioinformatics* **28**: 2086–2087.

Windhausen VS, Atlin GN, Hickey JM, Crossa J, Jannink J-L, Sorrells ME, *et al.* (2012). Effectiveness of Genomic Prediction of Maize Hybrid Performance in Different Breeding Populations and Environments. *G3: Genes|Genomes|Genetics* **2**: 1427–1436.

Xu Y (2016). Envirotyping for deciphering environmental impacts on crop plants. *Theor Appl Genet* **129**: 653–673.

Zhang X, Pérez-Rodríguez P, Burgueño J, Olsen M, Buckler E, Atlin G, *et al.* (2017). Rapid Cycling Genomic Selection in a Multiparental Tropical Maize Population. *G3* **7**: 2315–2326.

TABLES

Tabela 6. List of environmental factors considered in the study, estimated from NASA orbital sensors (Stackhouse Jr., 2014) and processed using concepts from Allen et al (1998) and Soltani and Sinclair (2012).

Source	Environmental Factor	Unit
NASA Power	Top-of-atmosphere insolation	MJ m^{-2} d^{-1}
	Average insolation incident on a horizontal surface	MJ m−2 d^{-1}
	Average downward longwave radiative flux	MJ m−2 d^{-1}
	Wind speed at 10 m above the surface of the earth	m s^{-1}
	Minimum air temperature at 2 m above the surface of the earth	°C d^{-1}
	Maximum air temperature at 2 m above the surface of the earth	°C d^{-1}
	Dew-point temperature at 2 m above the surface of the earth	°C d^{-1}
	Relative air humidity at 2 m above the surface of the earth	%
	Rainfall precipitation (P)	mm d^{-1}
Calculated[1]	Effect of Temperature on Radiation use Efficiency	-
	Evapotranspiration (ETP)	mm d^{-1}
	Atmospheric water deficit P-ETP	mm d^{-1}
	Deficit of vapor Pressure	kPa d^{-1}
	Slope of saturation vapor pressure curve	kPa C° d^{-1}
	Temperature Range	°C d^{-1}
	Global Solar Radiation based on Latitude and Julian Day	MJ $m^{-2}d^{-1}$

[1]Environmental data were collected, processed and organized by time intervals (phenology) using the functions get_weather(), summaryWTH() and W.matrix() from the *EnvRtype* package.

Tabela 7. Total time (in seconds) to execute a Markov Chain containing 10,000 iterations using BGGE package for each combination of genomic prediction model, kernel method and maize data set. Values between parenthesis denote the relative gain/reduction in computational time using GK and DK in comparison with the same model based on GB.

Set	Model	GB	GK	DK
HEL	EA	47	58 (+19%)	54 (+13%)
	EAD	97	101 (+4%)	88 (-10%)
	EAD+GE	134	139 (+4%)	126 (-6%)
	EADW	175	139 (-26%)	126 (-39%)
	EADW+GW	330	294 (-12%)	280 (-18%)
USP	EA	660	718 (+8%)	655 (-1%)
	EAD	1442	1341 (-8%)	1360 (-6%)
	EAD+GE	1684	1585 (-6%)	1600 (-5%)
	EADW	2440	1884 (-30%)	2202 (-11%)
	EADW+GW	4368	3800 (-15%)	4087 (-7%)

Tabela 8. Average correlations between predicted and observed values for grain yield (tons per ha) using 5 statistical models, 3 kernel methods and cross-validation schemes (CV1, CV2 and CV0) for a HEL maize set with 247 hybrids in 5 environments. Standard error values (*SE*) and the predictability gains in relation to the baseline model (EA) are given parenthesis and %, respectively. Bold numbers denote the best models for each kernel method.

		Model				
CV	**Kernel**	**EA**	**EAD**	**EAD+GE**	**EADW**	**EADW+GW**
		0.247	0.345	0.220	**0.832**	0.819
		(0.028)	(0.023)	(0.037)	(0.007)	(0.006)
	GB	-	28%	-12%	70%	70%
		0.306	0.350	**0.871**	0.831	0.824
		(0.024)	(0.021)	(0.006)	(0.016)	(0.007)
	GK	-	13%	65%	63%	63%
		0.305	0.338	0.669	**0.822**	**0.819**
		(0.016)	(0.020)	(0.019)	(0.008)	(0.007)
CV1	**DK**	-	10%	54%	63%	63%
		0.231	0.208	0.132	**0.839**	0.824
		(0.033)	(0.026)	(0.041)	(0.007)	(0.007)
	GB	-	-11%	-75%	73%	72%
		0.240	0.197	**0.892**	0.838	0.835
		(0.029)	(0.025)	(0.008)	(0.017)	(0.007)
	GK	-	-22%	73%	71%	71%
		0.209	0.172	0.734	**0.839**	**0.836**
		(0.022)	(0.031)	(0.006)	(0.009)	(0.009)
CV2	**DK**	-	-21%	72%	75%	75%
		0.402	0.558	0.551	**0.567**	0.537
		(0.059)	(0.045)	(0.046)	(0.041)	(0.046)
	GB	-	28%	27%	29%	25%
		0.505	0.560	**0.569**	**0.568**	**0.567**
		(0.055)	(0.047)	(0.034)	(0.041)	(0.042)
	GK	-	10%	11%	11%	11%
		0.533	**0.570**	**0.571**	**0.572**	**0.569**
		(0.064)	(0.049)	(0.036)	(0.042)	(0.041)
CV0	**DK**	-	7%	7%	7%	6%

Tabela 9. Average correlations between predicted and observed values for grain yield (tons per ha) using 5 statistical models, 3 kernel methods and cross-validation schemes (CV1, CV2 and CV0) for a USP maize set with 570 hybrids in 8 environments. Standard error values (*SE*) and the predictability gains in relation to the baseline model (EA) are given parenthesis and %, respectively. Bold numbers denote the best models for each kernel method.

		Model				
CV	**Kernel**	**EA**	**EAD**	**EAD+GE**	**EADW**	**EADW+GW**
		0.306	0.328	0.287	**0.658**	**0.669**
		(0.019)	(0.018)	(0.020)	(0.010)	(0.009)
	GB	-	7%	-7%	53%	54%
		0.324	0.323	**0.673**	**0.671**	**0.689**
		(0.018)	(0.017)	(0.009)	(0.010)	(0.009)
	GK	-	0%	52%	52%	53%
		0.305	0.338	0.669	**0.822**	**0.819**
		(0.016)	(0.02)	(0.009)	(0.008)	(0.007)
CV1	**DK**	-	10%	54%	63%	63%
		0.339	0.367	0.316	0.714	**0.731**
		(0.015)	(0.012)	(0.016)	(0.005)	(0.006)
	GB	-	8%	-7%	53%	54%
		0.370	0.362	0.733	0.730	**0.751**
		(0.013)	(0.012)	(0.005)	(0.005)	(0.005)
	GK	-	-2%	50%	49%	51%
		0.349	0.349	**0.891**	0.724	0.745
		(0.014)	(0.014)	(0.008)	(0.007)	(0.006)
CV2	**DK**	-	0%	61%	52%	53%
		0.335	0.425	0.427	**0.489**	**0.515**
		(0.014)	(0.015)	(0.016)	(0.088)	(0.105)
	GB	-	21%	22%	32%	35%
		0.406	0.429	**0.456**	**0.498**	**0.493**
		(0.015)	(0.015)	(0.021)	(0.098)	(0.090)
	GK	-	5%	11%	19%	18%
		0.403	0.428	**0.458**	**0.526**	**0.566**
		(0.015)	(0.016)	(0.034)	(0.098)	(0.092)
CV0	**DK**	-	6%	12%	23%	29%

FIGURES

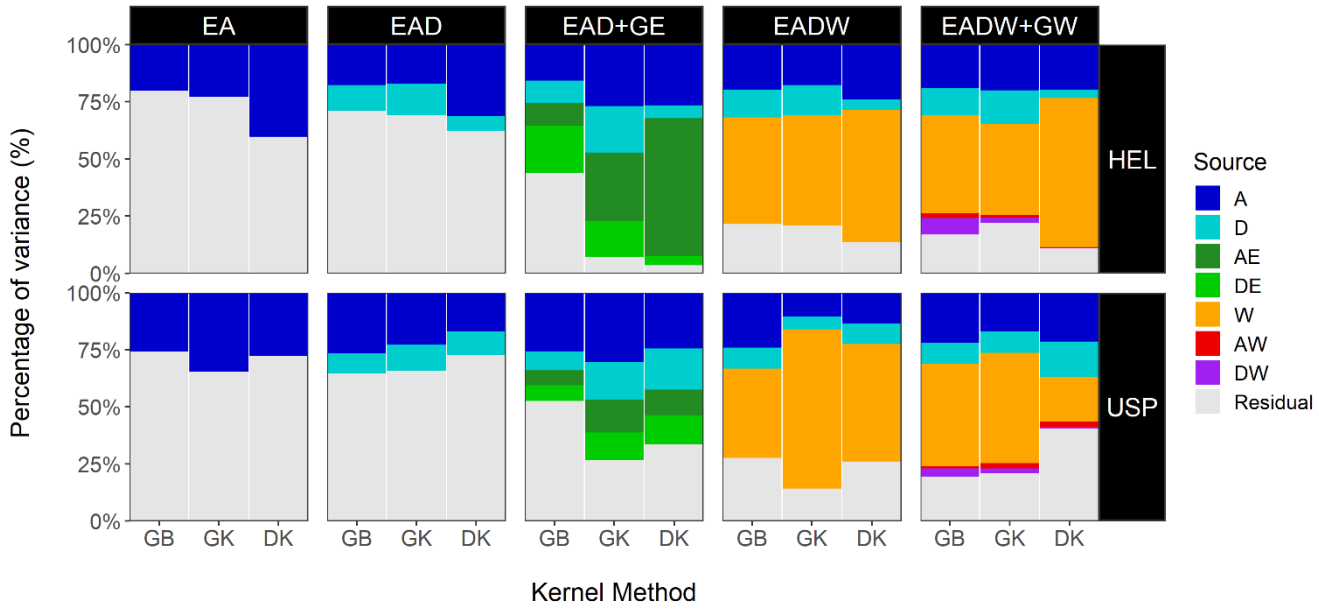

Figura 5. Partition of the variance components related to the different genetics, environmental and residual sources of variation for each of the five WGP models built with the three different kernel methods in HEL and USP data sets.

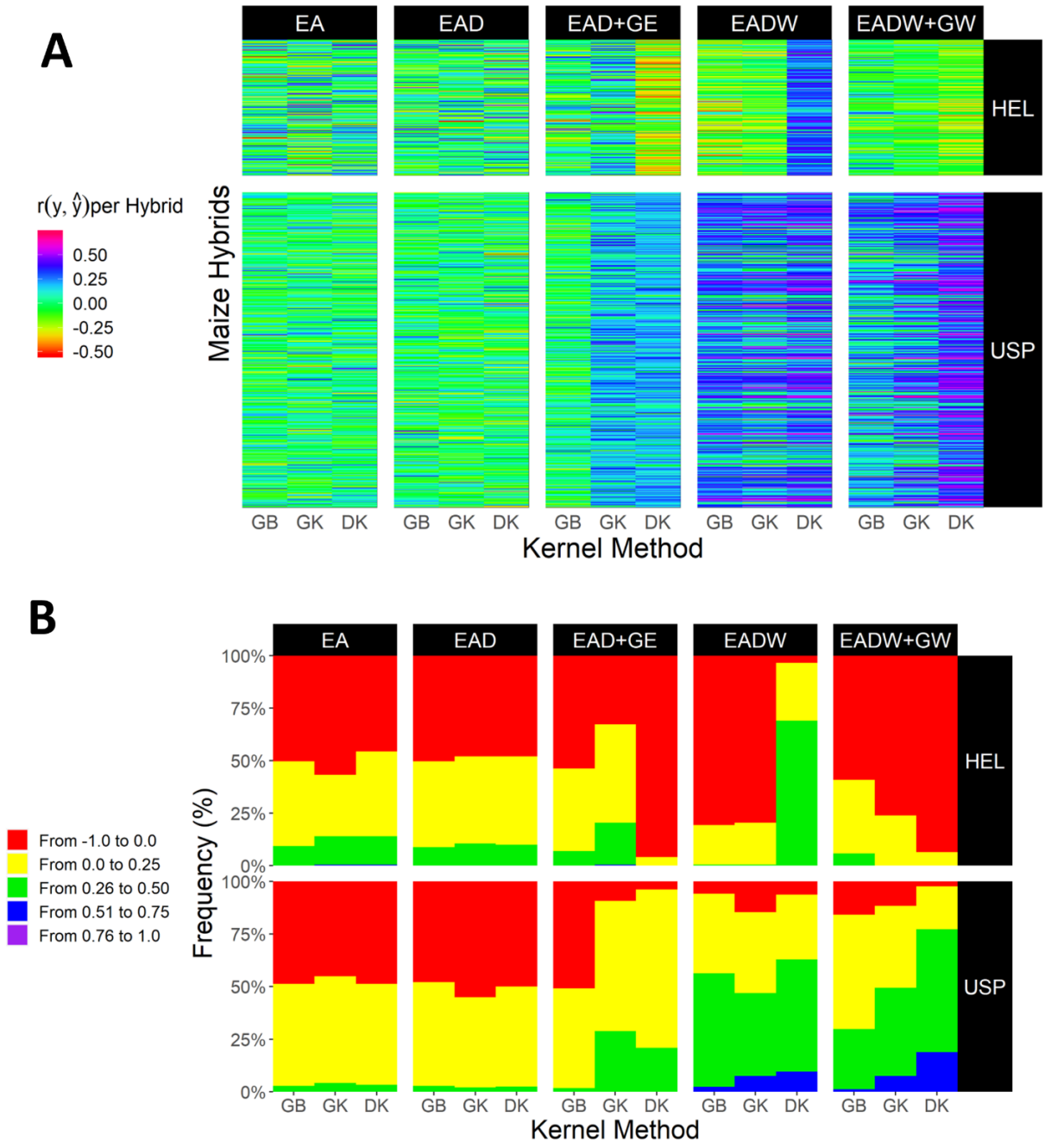

Figura 6. Resolution of the genomic-enabled models and kernel methods in predicting genotypes in novel environments. (**A**) predictive ability of specific hybrids (each row) involving (q-1) tested environments plus a one novel environment for sets HEL and USP, respectively; (**B**) typology of predictive abilities for sets HEL and USP, respectively. Predictive ability values are represented from warm colors (red, worst results) to cold colors (blue and purple, better results).

Figura 7. Joint accuracy trends of best combinations of kernel method and model in predicting novel environments (CV0) for both maize data sets (HEL and USP). On the X axis, the environments were ordered from the less predictable (S4) to the higher predictable (S3). Environments with the acronym S denote sites (from 1 to 5, in the HEL set) and with E denoting environments (from 1 to 8, in the USP set).

4. ENVIROMIC ASSEMBLY INCREASES ACCURACY AND REDUCES COSTS OF THE GENOMIC PREDICTION FOR YIELD PLASTICITY

ABSTRACT

Quantitative genetics states that phenotypic variation is a consequence of genetic and environmental factors and their subsequent interaction. Here, we present an enviromic assembly approach, which includes the use of ecophysiology knowledge in shaping environmental relatedness into whole-genome predictions (GP) for plant breeding (referred to as E-GP). We propose that the quality of an environment is defined by the core of environmental typologies (envirotype) and their frequencies, which describe different zones of plant adaptation. From that, we derive markers of environmental similarity cost-effectively. Combined with the traditional genomic sources (e.g., additive and dominance effects), this approach may better represent the putative phenotypic variation across diverse growing conditions (i.e., phenotypic plasticity). Additionally, we couple a genetic algorithm scheme to design optimized multi-environment field trials (MET), combining enviromic assembly and genomic kinships to provide in-silico realizations of the future genotype-environment combinations that must be phenotyped in the field. As a proof-of-concept, we highlight E-GP applications: (1) managing the lack of phenotypic information in training accurate GP models across diverse environments and (2) guiding an early screening for yield plasticity using optimized phenotyping efforts. Our approach was tested using two non-conventional cross-validation schemes to better visualize the benefits of enviromic assembly in sparse experimental networks. Results on tropical maize show that E-GP outperforms benchmark GP in all scenarios. We show that for training accurate GP models, the genotype-environment combinations' representativeness is more critical than the MET size. Furthermore, we discuss theoretical backgrounds underlying how the intrinsic envirotype-phenotype covariances within the phenotypic records of (MET) can impact the accuracy of GP and limits the potentialities of predictive breeding approaches. The E-GP is an efficient approach to better use environmental databases to deliver climate-smart solutions, reduce field costs, and anticipate future scenarios.

Keywords: Genomic selection, Adaptability, Genotype × environment, Climate-smart

4.1. INTRODUCTION

Environmental changing scenarios challenge agricultural research to deliver climate-smart solutions in a time-reduced and cost-effective manner (Tigchelaar et al., 2018; Ramírez-Villegas et al 2020; Cortés et al., 2020). Characterizing crop growth conditions is crucial for this purpose (Xu, 2016), allowing a deeper understanding of how the environment shapes past, present, and future phenotypic variations (e.g., Ramírez-Villegas et al. 2018; Heinemann et al., 2019; Cooper et al., 2014; de los Campos et al., 2020; Costa-Neto et al., 2021b; Antolin et al., 2021). For plant breeding research, mostly based on selecting the best-evaluated genotypes for a target population of environments (TPE), this approach is useful to discriminate genomic and non-genomic sources of crop adaptation. Thus, the concept of 'envirotyping' (environmental + typing, Cooper et al., 2014; Xu, 2016) emerges to establish the quality of a given environment in the delivery of quality phenotypic records, mostly to train accurate predictive breeding approaches capable of guiding the selection of most productive and adapted genotypes (Resende et al., 2020; Costa Neto et al., 2021a; Crossa et al., 2021).

From envirotyping, it is possible to check the quality of a certain environment, which is directly related to how the observed growing conditions in a particular field trial could be related to the most frequent environment-

types (envirotypes) that occur in the breeding program TPE or target region (e.g., Heinemann et al., 2019; Cooper et al., 2021; Antolin et al., 2021). In agricultural research, the quality of a certain environment is directly related to how it can limit the expression of the genetic potential of the certain crop for a certain trait, such as suggested by the movement called 'School of de Wit' since 1965 (see Bouman et al., 1996). Thus, for the plant breeding research, this is also direct factors such as genotype × environment interaction (e.g., Allard, 1964; Finlay and Wilkinson, 1963) and its implications of how the target germplasm under selection (or testing) can perform across the target growing conditions in which the candidate cultivars will be cropped.

Prediction-based tools have leveraged modern plant breeding research to an extent in which phenotyping is still required (Crossa et al., 2017), although prediction-based tools and simulations can support more comprehensive and faster selection decisions (Galli et al., 2020; Cooper et al., 2021; Crossa et al., 2021). One of the most widely used predictive tools is the whole-genome prediction (GP, Meuwissen et al., 2001), developed and validated for several crop species and application scenarios (Crossa et al., 2017; Voss-Fels et al., 2019), such as the selection among populations and the prediction of the performance of single-crosses across multiple environments. For the latter, the most important use of GP mostly relies on the better use of the available phenotyping records and large-scale easy-managed genomic information to expand the spectrum of evaluated single-crosses in silico (Messina et al., 2018; Rogers et al., 2021). Those phenotypic records (e.g., grain yield and plant height) are collected from existing field trials that experience a diverse set of growing conditions, carrying within them an intrinsic environment-phenotype covariance. Consequently, the GP has a limited accuracy under multiple-environment testing (MET) due to genotype × environment interaction (G×E) (Crossa et al., 2017), meaning that each genotype has a differential response for each environmental factor that assembles what we call 'environment' (time interval across crop lifetime involving a specific geographic location and agronomic practice for a particular crop). Therefore, novel ways to include environmental data (Heslot et al., 2014; Jarquín et al., 2014; Ly et al., 2018; Millet et al., 2019; Gillberg et al., 2019; Costa-Neto et al., 2021a) and process-based crop growth models (CGM) (Messina et al., 2018; Toda et al. 2020; Robert et al., 2020; Cooper et al., 2021) in GP are considered the best pathways to fix it. Most of the success achieved by such approaches lies in a better understanding of the visible ecophysiology interplay between genomics and environment variation (Gage et al., 2017; Li et al., 2018; Guo et al., 2020; Costa-Neto et al., 2021b).

The explicit integration of enviromic and genomic sources is an easy way to lead GP to a wide range of novel applications (Crossa et al., 2021), such as improving the predictive ability for untested growing conditions (Guo et al., 2020; de los Campos et al., 2020; Jarquín et al., 2020; Costa-Neto et al., 2021a), to optimize MET networks and to screen genotype-specific reaction-norms (Ly et al., 2018; Millet et al., 2019). This is excellent progress for predictive breeding (i.e., the range of prediction-based selection tools for crop improvement) and accelerating research pipelines to deliver higher yields and adapted genotypes for target scenarios. However, most of the current studies on this topic vary in accuracy and applicability, mostly due to (1) the processing protocols used to translate the raw-data into explicit environmental covariables (ECs) with biological meaning in explaining G×E over complex traits, (2) the lack of a widely-used envirotyping pipeline that, not only supports the design of field trials, but also increases the accuracy of the trained GP models and, in addition, (3) for CGM, a possible limitation is the increased demand for the phenotyping of additional intermediate phenotypes (i.e., biomass accumulation and partitioning, specific leaf area), which can involve managed iso-environments and expert knowledge in crop modeling (Cooper et al., 2016; Toda et al., 2020; Robert et al., 2020). The latter can be expensive or difficult for plant research programs in developing countries, which generally have low budgets to increase the phenotyping network and install environmental sensors. In addition, most

developing countries are located in regions where environments are subject to a broader range of stress factors (e.g., heat stress).

Therefore, here we revisit Shelford's Law (Shelford, 1931) and other ecophysiology concepts that can provide the foundations for translating raw-environmental information into an enviromic source for predictive breeding, hereafter denominated as enviromic assembly. The benefits of using the so-called 'enviromics-aided GBLUP' (E-GP) under existing experimental networks are presented, followed by the E-GP application to optimize field-based phenotyping. Finally, we benchmark E-GP with the traditional genomic-best unbiased prediction (GBLUP) to discuss the benefits of enviromic data to reproduce G×E patterns and provide a virtual screening for yield plasticity.

4.2. MATERIAL AND METHODS

The material methods are organized in the following manner: First, we briefly address the concepts underlying the novel approach of enviromic assembly inspired by Shelford's Law. The data sets are then presented, along with the statistical models and prediction scenarios used to show the benefits of large-scale environmental information in GP across multi-environment trials (MET). Finally, we present a scheme to optimize phenotyping efforts in training GP over MET and support the screening for maize single-crosses' yield plasticity.

4.2.1. Theory: adapting the Shelford Law of Minimum

Consider two experimental networks (MET) of the same target population of environments (TPE, e.g., the different locations, years, and crop management) under different environmental gradients due to year or location variations (Fig.1). For two genotypes evaluated under both conditions (G1, G2), the potential genetic-specific phenotypic plasticity (Allard and Bradshaw, 1964) (curves) is expressed as different reaction-norms (dotted lines), resulting in distinct observable G×E patterns (Fig.1a-b). In the former MET (Fig.1a), both genotypes experience a wider range of possible growing conditions (large interval between the two vertical solid lines), which result in an intricate G×E pattern (crossover). Conversely, in the latter MET (Fig.1b), the same genotypes experience a reduced range of growing conditions yet lead to a simple G×E pattern (non-crossover). It is feasible to conclude that, although the genetic variation is essential for modeling potential phenotypic plasticity of genotypes (curves, Fig.1a-c), the diversity of environmental growing conditions dictates the observable G×E patterns (Bradshaw, 1965). Thus, the GP platforms for MET may be unbiased with no diversity, and the quality of environments is not considered.

Approaches such as CGM try to reproduce the phenotypic plasticity curves, while benchmark reaction-norm models try to reproduce the observable reaction-norm. Both approaches can achieve adequate results, although we have observed that (1) CGM demands greater phenotyping efforts to train computational approaches capable of reproducing the achievable phenotypic plasticity from a reduced core of phenotypic records from field trials at near-iso environments (e.g., well-watered conditions versus water-limited conditions for the same planting date and management), (2) CGM demands additional programing efforts, which, for some regions or crops, can be expensive and limit the applicability of the method, (3) adequate reaction-norm models over well-designed phenotyping platforms are not a reality for certain regions of the world with limited resources to invest in precision phenotyping efforts.

We understand that Shelford's Law of Tolerance (Shelford, 1931) is suitable to explain how the environment drives plant plasticity and can be incorporated into the traditional GP platforms in a cost-effective way (Fig.1c). It states that a target population's adaptation is modulated as a certain range of minimum, maximum and optimum threshold limits achieved over an environment gradient (vertical solid green lines). The genotypes' potential phenotypic plasticity (curves) is not regarded as a linearized reaction-norm variation across an environmental gradient (Arnold et al., 2019). Instead, it is the distribution of possible phenotypic expressions dictated by the cardinal thresholds for each biophysical factor with ecophysiological relevance. Therefore, crops may experience stressful conditions due to the excess or lack of a target environmental factor, depending on the cardinal thresholds (vertical solid green lines in Fig.1c), which also rely on some key development stages germplasm-specific characteristics (e.g., tropical maize versus temperate maize). Consequently, the expected variation of environmental conditions across different field trials results from a series of environment-types (envirotypes) acting consistently yet varying in impact depending on the genetic-specific sensibility. The quality of a certain growing condition depends on the balance between crop necessity and resource availability, which involves constant effects, such as the type of treatments in a trial (e.g., fertilizer inputs) and transitory effects variables, such as weather events (e.g., heat-stress).

From these concepts, we observe that with the use of envirotyping (e.g., typing the profiles of a particular environment), the environment part of the G×E pattern can be visualized based on the shared frequency of envirotypes among different field trials. Thus, the enviromic of a certain experimental network or TPE (the core of possible growing conditions) can be mathematically assembled by (1) collecting large-scale environmental data, (2) processing this raw data in envirotyping entries for each real or virtual environment, and (3) processing these envirotyping-derived entries to achieve theoretical relatedness between the buildup of different environments from the shared frequency of envirotypes. Thus, the expected envirotypes can be designed relying on the adaptation zones inspired by the model proposed here, based on Shelford's Law, in which we can envisage the process of deriving environmental covariables for GP into an ecophysiological-smart way.

4.2.2. Proof of concept data sets

This study used maize as a proof-of-concept crop due to its importance for food security in developed and developing regions. Two data sets of maize hybrids (single-crosses of inbreed lines) from different germplasm sources developed under tropical conditions in Brazil (hereafter referred to as Multi-Regional and N-level) were used. Both data sets involve phenotypic records of grain yield (Mg per ha) collected across multiple environments. Details on the experimental design, cultivation practices, and fundamental statistical analysis are given in Bandeira e Souza et al. (2017) and Alves et al. (2019). Below we provide a short description of the number of genotypes and environments tested and the nature of this study's genotyping data.

Multi-Regional Set: The so-called "Multi-Regional set" is based on the germplasm developed by the Helix Seeds Company (HEL) in South America. It includes 247 maize lines evaluated in 2015 in five locations in three regions of Brazil (Supplementary Table 5). Genotypes were obtained using the Affymetrix Axiom Maize Genotyping Array containing 616 K SNPs (single-nucleotide polymorphisms) (Unterseer et al., 2014). Only SNPs with a minor allele frequency > 0.05 were considered. Finally, a total of 52,811 high-quality SNPs that achieved the quality control level were used in further analysis.

N-level set: The so-called "N-level set" is based on the germplasm developed by the Luiz de Queiroz College of Agriculture of the University of São Paulo (USP), Brazil. A total of 570 tropical maize hybrids were evaluated across

eight environments, involving an arrangement of two locations, two years, and two nitrogen levels (Supplementary Table 5). This study's sites involved two distinct edaphoclimatic patterns, i.e., Piracicaba (Atlantic forest, clay soil) and Anhumas (savannah, silt–sandy soil). In each site, two contrasting nitrogen (N) fertilization levels were managed. One experiment was conducted under ideal N conditions and received 30 kg ha−1 at sowing, along with 70 kg ha−1 in a coverage application at the V8 plant stage. That is the main recommendation for fertilization in tropical maize growing environments in Brazil. In contrast, the second experiment under low N conditions received only 30 kg ha−1 of N at sowing, resulting in an N-limited growing condition. This set's genotypes were also obtained using the Affymetrix Axiom Maize Genotyping Array containing 616 K SNPs (Unterseer et al., 2014) and minor allele frequency > 0.05. At the end of this process, a total of 54,113 SNPs were considered in the GP modeling step.

4.2.3. Envirotyping Pipeline

Below, we present the methods used for data collection, data processing, and implementing what we call 'enviromic assembly'. This envirotyping pipeline was developed using the functions of the R package EnvRtype (Chapter 1 of this thesis, published as Costa-Neto et al., 2021b) and is available at no cost.

Step 1. Data Collection: In this study, environmental information was used for the main abiotic plant-environment interactions related to daily weather, soil type, and crop management (available only for N-level set). Daily weather information was collected from NASA POWER (Sparks, 2018) and consisted of eight variables: rainfall (P, mm day-1), maximum air temperature (TMAX, °C day-1), minimum air temperature (TMIN, °C day-1), average air temperature (TAVG, °C day-1), dew point temperature (TDEW, °C day-1), global solar radiation (SRAD, MJ per m^2), wind speed at 2 meters (WS, m s-1 day-1) and relative air humidity (RH, % day-1). Elevation above sea level was obtained from NASA's Shuttle Radar Topography Mission (SRTM). Both sources were imported into R statistical-computational environments using the functions and libraries organized within the EnvRtype package (Costa-Neto et al., 2021b). A third GIS database was used to import soil types from Brazilian soil classification provided by EMBRAPA and available from Supplementary Data and Software.

Step 2. Data Processing: Quality control was adopted by removing variables outside the mean ± three standard deviation and repeated columns. After checking for outliers, the daily weather variables were used to model ecophysiological interactions related to soil-plant-atmosphere dynamics.

Step 3. Thermal and evapotranspiration covariates: The thermal-radiation interactions computed potential atmospheric evapotranspiration (ET0) following the Priestley-Taylor method. The slope of the saturation vapor pressure curve (SPV) and vapor pressure deficit (VPD) was computed as given in the FAO manual (Allen et al., 1998). An FAO-based generic function was used to estimate crop development as a function of days after emergence (DAE). We assume a 3-segment leaf growing function to estimate the crop canopy coefficient (Kc) of evapotranspiration using the following Kc values: Kc1 (0.3), Kc2 (1.2), Kc3 (0.35), equivalent to the water demand of tropical maize for initial phases, reproduction phases, and end-season stages, respectively. Using the same 3-segment function, we estimate the crop canopy using a leaf area index (LAI) of LAI = 0.7 (initial vegetative phases), LAI = 3.0 (maximum LAI for tropical maize growing conditions observed in our fields), and LAI = 2.0 (LAI tasseling stage). We computed the daily crop evapotranspiration (ETc) estimated by the product between ET0 and the Kc from those two estimations. Then, we computed the difference between daily precipitation and crop evapotranspiration as P-ETc.

Step 4. Ecophysiology covariates: The apparent photosynthetic radiation intercepted by the canopy (aPAR) was computed following aPAR=SRAD×(1-exp(-k×LAI)), where k is the coefficient of canopy, considered as 0.5. Water deficiency was computed using the atmospheric water balance between input (precipitation) and output of atmospheric demands (crop evapotranspiration). The effect of temperature on the radiation use efficiency (FRUE) was described by a three-segment function based on cardinal temperatures for maize, using the cardinal temperatures 8°C (Tb1, base lower), 30°C (To1, base optimum), 37°C (To2, upper optimum) and 45°C (Tb2, base upper). This function assumes values from 0 to 1, depending on: FRUE= 0 if TAVG ≤ Tb1; FRUE = (TAVG - Tb1)/(To1-Tb1) if Tb1 < TAVG < To1; FRUE = 1 if To1 < TAVG < To2; FRUE = (Tb2 – TAVG)/(Tb2 – To2) if To2 < TAVG < Tb2; and FRUE = 0 if TAVG > Tb2.

Step 5. Divining crop lifetime: Finally, we sampled each piece of weather and ecophysiological information across five-time intervals in the crop lifetime: from emergence to the appearance of the first leaf (V1, 14 DAE), from V1 to the fourth leaf (V4, 35 DAE), from V4 to the tasseling stage (VT, 65 DAE), from VT to the kernel milk stage (R3, 90 DAE) and from R3 to physiological maturity (120 DAE), in which DAE stands for days after emergence.

4.2.4. Enviromic assembly using typologies (T matrix)

The raw envirotyping data were used to assemble markers for environmental similarity, depending on the group of the ECs. The first group of ECs involves the transitory effect variables, which vary in the frequency of occurrence, depending on the crop development cycle. Thus, we design the expected envirotypes using the number of inputs required to lead crops in at least three levels of adaptation: (1) stress by deficit, (2) optimum growing conditions, and (3) stress by excess. These levels were defined using cardinal thresholds or frequency tables concerning the growing conditions archived in the experimental network range. Then, from having reviewed the literature, we consider the intervals for thermal-related variables: 0°C to 9°C (death), 9.1°C to 23°C (stress by deficit), 23.1°C to 32°C (optimum growing conditions), 32.1°C to 45°C (stress by excess) and 45°C to ∞°C (death). We computed the classes for accumulated prediction according to our agronomic expertise on rainfall requirements for tropical maize growing environments: 0mm to 10mm, 10.1mm to 20mm, and 20.1mm to ∞ mm. For crop evapotranspiration (ETc), we assume the envirotypes 0-6 mm.day-1, 7-10 mm.day-1 ,10-15 mm.day-1 and 16 to ∞ mm.day-1. Finally, for FRUE, we assume the envirotypes based on the following adaptation zones: impact from 0% to 25% (0-0.25), from 26% to 50% (0.26-0.50), 51% to 75% (0.51-0.75) and 76% to 100% (0.76-1.0). We preferred to adopt a simple discretization for the remaining variables using a histogram of percentiles (0-25%, 26-50%, 51-75%, 75-100%) of occurrence for a target envirotype.

The second group involves constant effect variables. In this group, we consider the elevation, crop management, and soil classification in each environment. Soil information was entered as an incidence matrix (0 or 1) based on each environment's occurrence. In addition, for the N-level set, nitrogen input levels were computed as two discrete classes: ideal N = 10 and low N = 30; we entered this same incidence matrix for soil information. Because both sets have a gradient for elevation, we used a histogram of percentiles (0-25%, 26-50%, 51-75%, 75-100%) as in the transitory group of variables. Finally, each designed envirotype × time interval frequency was used as a qualitative marker of environmental relatedness (the hereafter **T** matrix, from typologies).

4.2.5. W matrix of quantitative environmental covariables

The quantitative descriptors of environmental relatedness are the most common method to include environmental information in GP studies considering reaction-norms. Jarquín et al. (2014) proposed the creation of the so-called environmental relatedness kinship ($\mathbf{K_E}$) carried out with a matrix of quantitative environmental covariables (**W** matrix, thus we refer to this environment kinship as $\mathbf{K_{E,W}}$). Here, this pattern of similarity in $\mathbf{K_{E,W}}$ was captured using percentile values (25%; 50%, and 75%) at each of the five-time intervals of development, as suggested by Morais-Júnior et al. (2018) and expanded by Costa-Neto et al., (2021a, Chapter 2 of this thesis). We found 255 and 307 quantitative descriptors for the Multi-Regional and N-level sets at the end of the process, respectively. In this study, we used this $\mathbf{K_{E,W}}$ as a benchmark method (section 4.2.9) to test the effectiveness of the $\mathbf{K_{E,T}}$ matrix (4.2.8) and the total absence of environmental information (baseline genomic model without environmental information, see section 4.2.7).

4.2.6. Statistical models

From a baseline additive-dominant multi-environment GBLUP (section 4.2.7), we tested four other models, created with the inclusion of two types of enviromic assembly (T or W) and structures for G×E effects. More details about each statistical model are provided in the next subsections. All kernel models were fitted using the BGGE R package (Granato et al., 2018) using 15,000 iterations, with 2,000 used as burn-in and using a thinning of 10. This package was used due to the following aspects: (1) is an accurate open-source software and; (2) can accommodate many kernels in a computation-efficient way.

4.2.7. Baseline additive-dominant multi-environment GBLUP

The baseline model includes a fixed intercept for each environment and random genetic variations (additive and dominance). We will refer to this model as GBLUP, which was modeled as an overall main effect plus a genomic-by-environment deviation (the so-called G+GE model, Bandeira e Souza *et al.*, 2017), as follows:

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \boldsymbol{Z}_E\boldsymbol{\beta} + \boldsymbol{Z}_A\boldsymbol{u}_A + \boldsymbol{Z}_D\boldsymbol{u}_D + \boldsymbol{u}_{AE} + \boldsymbol{u}_{DE} + \boldsymbol{\varepsilon} \qquad (Eq.1)$$

where $\boldsymbol{y} = [\boldsymbol{y_1}, \cdots, \boldsymbol{y_n}]'$ is the vector of observations collected in each of the q environments with hybrids and $\mathbf{1}\boldsymbol{\mu} + \boldsymbol{Z}_\mathrm{E}\boldsymbol{\beta}$ is the general mean and the fixed effect of the environments with incidence matrix $\boldsymbol{Z}_E$. Genetic variations are modeled using the main additive effects ($\boldsymbol{u}_\mathrm{A}$), with $\boldsymbol{u}_\mathrm{A} \sim N(\mathbf{0}, \boldsymbol{J}_q \otimes \boldsymbol{K}_\mathrm{A}\boldsymbol{\sigma}^2_\mathrm{A})$, plus a random dominance variation ($\boldsymbol{u}_\mathrm{D}$), with $\boldsymbol{u}_\mathrm{D} \sim N(\mathbf{0}, \boldsymbol{J}_q \otimes \boldsymbol{K}_\mathrm{D}\boldsymbol{\sigma}^2_\mathrm{D})$, where $\boldsymbol{\sigma}^2_\mathrm{A}$ and $\boldsymbol{\sigma}^2_\mathrm{D}$ are the variance component for additive and dominance deviation effects; $\boldsymbol{Z}_\mathrm{A}$ and $\boldsymbol{Z}_\mathrm{D}$ are the incidence matrix for the same effects (absence=0, presence=1), $\boldsymbol{J}_q$ is a q×q matrix of 1s and $\otimes$ denotes the Kronecker Product. G×E effects are modeled using a block diagonal (BD) matrix of the genomic effects, built using $\boldsymbol{u}_\mathrm{AE} \sim N(\mathbf{0}, \boldsymbol{I}_q \otimes \boldsymbol{K}_\mathrm{A}\boldsymbol{\sigma}^2_\mathrm{A})$ and $\boldsymbol{u}_\mathrm{DE} \sim N(\mathbf{0}, \boldsymbol{I}_q \otimes \boldsymbol{K}_\mathrm{D}\boldsymbol{\sigma}^2_\mathrm{D})$, in which $\boldsymbol{I}_q$ is a diagonal matrix of q×q dimension. Residual deviations ($\boldsymbol{\varepsilon}$) were assumed as $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{I}_n\boldsymbol{\sigma}^2)$, where n is the number

of genotype-environment observations. Furthermore, the genotyping data were processed in two matrices of additive and dominance effects, modeled by:

$\mathbf{A} = \{0 = A^2A^2; 1 = A^1A^2; 2 = A^1A^1\}$ and

$\mathbf{D} = \{-2f_l^2 = A^2A^2; 2f(1-f_l) = A^1A^2; -2f(1-f_l)^2 = A^1A^1\}$,

where f_l is the frequency of the favorable allele at locus l. Thus, the genomic-related kinships were estimated as follows:

$$\boldsymbol{K} = \frac{\boldsymbol{XX'}}{\text{trace}(\boldsymbol{XX'})/nrow(\boldsymbol{X})} \quad (Eq.2)$$

where $\boldsymbol{K}$ is a generic representation of the genomic kinship ($\boldsymbol{K}_\text{A}$, $\boldsymbol{K}_\text{D}$), $\boldsymbol{X}$ is a generic representation of the molecular matrix (**A** or **D**), and $nrow(\boldsymbol{X})$ denotes the number of rows in $\boldsymbol{X}$ matrix.

4.2.8. GBLUP with enviromic main effects from T matrix (E-GP)

Eq (2) was also used to shape the environmental relatedness kernels using **T** or **W** matrix. This linear kernel for $\mathbf{K}_E$ was described by Jarquín *et al.* (2014), which some other authors named it after "**Ω**". Thus, here we only tested the difference between the enviromic source considered for building it and not the merit of the kernel method as was done in previous works (Costa-Neto *et al.,* 2021a). Thus, from baseline equation (1), we include a main environmental relatedness effect carried out with the **T** matrix ($\boldsymbol{u}_{\text{E,T}}$), as follows (Costa-Neto *et al.*, 2021a):

$$\boldsymbol{y} = \boldsymbol{1\mu} + \boldsymbol{Z}_A\boldsymbol{u}_A + \boldsymbol{Z}_D\boldsymbol{u}_D + \boldsymbol{u}_{AE} + \boldsymbol{u}_{DE} + \boldsymbol{u}_{E,T} + \boldsymbol{\varepsilon} \quad (Eq.3)$$

with $\boldsymbol{u}_{\text{E,T}} \sim N(\boldsymbol{Z}_E\boldsymbol{\beta}, \boldsymbol{K}_{\text{E,T}} \otimes \boldsymbol{J}_p\boldsymbol{\sigma}^2_{\text{E,T}})$, where $\boldsymbol{J}_q$ is a $p \times p$ matrix of 1s, is $\boldsymbol{K}_{\text{E,T}}$ the environmental relatedness created and variance component from the **T** matrix. If non-enviromic sources are considered, the expected value for environments is given by $\boldsymbol{Z}_E\boldsymbol{\beta}$ as the baseline model. In this model, the G×E effects are also modeled as the BD genomic matrix. Thus, we refer to this model as "E-GP (BD)". The kernel of enviromic assembly ($\boldsymbol{K}_{E,T}$) was built using the panel of envirotype descriptors (**T**) in the same way as described in equation (2).

From model (3), we substitute the BD for a reaction-norm (RN, Jarquín et al., 2014) based on the Kronecker product between the enviromic and genomic kinships (Martini *et al.*, 2020) for additive ($\boldsymbol{u}_{AE,T}$) and dominance effects ($\boldsymbol{u}_{DE,T}$):

$$\boldsymbol{y} = \boldsymbol{1\mu} + \boldsymbol{Z}_A\boldsymbol{u}_A + \boldsymbol{Z}_D\boldsymbol{u}_D + \boldsymbol{u}_T + \boldsymbol{u}_{A,T} + \boldsymbol{u}_{D,T} + \boldsymbol{\varepsilon} \quad (Eq.4)$$

with $\boldsymbol{u}_{\text{A,ET}} \sim N(\boldsymbol{0}, \boldsymbol{K}_{\text{E,T}} \otimes \boldsymbol{K}_A\boldsymbol{\sigma}^2_{\text{AE,T}})$ and $\boldsymbol{u}_{\text{D,ET}} \sim N(\boldsymbol{0}, \boldsymbol{K}_{\text{E,T}} \otimes \boldsymbol{K}_D\boldsymbol{\sigma}^2_{\text{DE,T}})$ where $\boldsymbol{\sigma}^2_{\text{AE,T}}$ and $\boldsymbol{\sigma}^2_{\text{DE,T}}$ are the variance components for enviromic × additive and enviromic × dominance effects performed as reaction-norms (Costa-Neto *et al.,* 2021a; Rogers *et al.*, 2021), respectively. For short, this model will be named "E-GP (RN)".

4.2.9. GBLUP with enviromic main effects from W matrix (W-GP)

Finally, in models (4) and (5), we substitute the enviromic assembly derived from **T** by the same kernel size derived from **W**, that is, an environmental relatedness with $\boldsymbol{u}_{\text{E,W}} \sim N(\boldsymbol{Z}_E\boldsymbol{\beta}, \boldsymbol{K}_{\text{E,W}} \otimes \boldsymbol{J}_p\boldsymbol{\sigma}^2_{\text{E,W}})$, creating two more models:

$$\boldsymbol{y} = \boldsymbol{1\mu} + \boldsymbol{Z}_A\boldsymbol{u}_A + \boldsymbol{Z}_D\boldsymbol{u}_D + \boldsymbol{u}_{AE} + \boldsymbol{u}_{DE} + \boldsymbol{u}_{E,w} + \boldsymbol{\varepsilon} \quad (Eq.5)$$

and

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \boldsymbol{Z}_A\boldsymbol{u}_A + \boldsymbol{Z}_D\boldsymbol{u}_D + \boldsymbol{u}_{E,w} + \boldsymbol{u}_{AE,w} + \boldsymbol{u}_{DE,w} + \boldsymbol{\varepsilon} \qquad (Eq.6)$$

$\boldsymbol{u}_{\mathrm{AE,W}} \sim N(\mathbf{0}, \boldsymbol{K}_{\mathrm{E,W}} \otimes \boldsymbol{K}_A\boldsymbol{\sigma}^2_{\mathrm{AE,W}})$ and $\boldsymbol{u}_{\mathrm{DE,W}} \sim N(\mathbf{0}, \boldsymbol{K}_{\mathrm{E,W}} \otimes \boldsymbol{K}_D\boldsymbol{\sigma}^2_{\mathrm{ED,W}})$, where $\boldsymbol{K}_{\mathrm{E,W}}$ and $\boldsymbol{\sigma}^2_{\mathrm{E,W}}$ are the resulting kinship and the variance components estimated for enviromic assembly from the **W** matrix, respectively. Thus, for short, models (5) and (6) will be referred to as "W-GP (BD)" and "W-GP (RN)", respectively.

4.2.10. Study cases for the E-GP platform

In this study, we conceived two cases to highlight the benefits of E-GP to boost efficiency in prediction-based platforms for hybrid development in maize breeding (Figure 2). The first case (*Case 1*) involves predicting the single-crosses from different theoretical existing experimental network setups, where we dissect the predictive ability over four prediction scenarios. In the second case (Case 2), we explore a theoretical conception of a super-optimized experimental network using the most representative combination of genotypes-environments selected using genomics, enviromic assembly, and genetic algorithms. Below we detail each case studied.

Case 1: expanding the existing field trials: In the first case (*Case 1*), we design a novel cross-validation scheme to split the global available phenotypic information (n), from p genotypes and q environments, into different training setups. Consequently, four prediction scenarios were created based on the simultaneous sampling of the phenotypic information for S genotypes and R environments.

- G, E: refers to the predictions of the tested genotypes within the experimental network (known genotypes in known environmental conditions). The size of this set is $\mathrm{n}_{[\mathrm{G,E}]} = \mathrm{n} \times \left(\frac{\mathrm{S}}{\mathrm{p}}\right) \times \left(\frac{\mathrm{R}}{\mathrm{q}}\right)$;
- *n*G,E: refers to predictions of untested (new) genotypes within the experimental network (known environmental conditions). The size of this set is $\mathrm{n}_{[\mathrm{nG,E}]} = \mathrm{n} \times \left(1 - \frac{\mathrm{S}}{\mathrm{p}}\right) \times \left(\frac{\mathrm{R}}{\mathrm{q}}\right)$;
- G,*n*E: in this scenario, predictions are made under environmental conditions external to those found within the experimental network. However, there is phenotypic information available within the experimental network. The size of this set is $n_{[G,nE]} = n \times \left(\frac{S}{p}\right) \times \left(1 - \frac{R}{q}\right)$;
- *n*G,*n*E: refers to predicting untested (new) genotypes and untested (new) environmental conditions. This set's size is $n_{[nG,nE]} = n \times \left(1 - \frac{S}{p}\right) \times \left(1 - \frac{R}{q}\right)$.

Theoretically, if $R/q = 1$, then $n_{[G,nE]} = n_{[nG,nE]} = 0$, equal to the commonly used CV1 scheme (prediction of novel genotypes in known environments). Different intensities of R/q can be sampled, which permits the testing of different sets of experimental networks. Here we simulated three different experimental network setups for each tropical maize data set. For the *N-level set*, we made 3/8, 5/8, and 7/8; for the *Multi-local set* 2/5, 3/5, and 4/5. We assumed the same level of genotype sampling as the training set for all experimental setups, equal to a fraction of $\frac{S}{p} = 0.7$. Each training setup was randomly sampled 50 times in order to compute the prediction quality statistics. For this purpose, two statistics were used to assess the statistical models' performance over these training setups. We calculated Pearson's moment correlation (r) between observed (y) and predicted ($\hat{y}$) values and used the average value for each model and training setup as a predictive ability statistic. To check the GP's ability to replace field trials, we

then computed the coincidence (CS, in %) between the field-based selection and the selection-based selection of the top 5% best-performing hybrids in each environment.

Case 2: designing super-optimized field trials: The second case (*Case 2*) was performed on the optimized training set described below. The first step was to compute a full-entry G×E kernel, based on the Kronecker product ($\otimes$) between the enviromic assembly-based relatedness kernel ($\boldsymbol{K}_{E,T}$, $q \times q$ environments) and genomic kinship ($\boldsymbol{K}_G$, $p \times p$ genotypes), thus$\boldsymbol{K}_{GE,T} = \boldsymbol{K}_{E,T} \otimes \boldsymbol{K}_G$, with an $n \times n$ dimension, in which $n = pq$. Here we adopted the kernel made up for additive effects ($\boldsymbol{K}_G = \boldsymbol{K}_A$) as the genomic kinship, despite the benefits of dominance effects in modeling G×E. We chose to use only $\boldsymbol{K}_A$ for simplicity and since additive effects seems to be a major genomic-related driver of G×E for grain yield in tropical maize (Dias *et al.*, 2018; Alves *et al.*, 2019; Costa-Neto *et al.*, 2021a; Roger *et al.*, 2021), a fact that was also observed for *Case 1* (Supplementary Tables 1-2). Later, we applied a single-value decomposition in $\boldsymbol{K}_{GE,T}$, following $\boldsymbol{K}_{GE,T} = \boldsymbol{UVU^T}$ where $\boldsymbol{U}$ is a total of eigenvalues and $\boldsymbol{V}$ the respective eigenvectors. The number of eigenvalues that explains 98% of the variance present in $\boldsymbol{K}_{GE,T}$ indicate the number of effective SNPs by envirotype-marker interactions, which is also the minimum core of genotype-environment combinations ($\boldsymbol{N}_{GE}$). Thus, the reduced phenotypic information of some genotypes in some environments ($\boldsymbol{N}_{GE}$) was used to predict a virtual experimental network ($\boldsymbol{N}_{test}$), involving all remaining single-crosses in all available environments, thus given by $\boldsymbol{N}_{test} = n - \boldsymbol{N}_{GE}$,

Following this step, a genetic algorithm scheme using the design criteria PEV_{MEAN} was used to identify the $\boldsymbol{N}_{GE}$ combinations of genotypes in environments within the $\boldsymbol{K}_{GE,T}$ entries that must be phenotyped (Misztal, 2016). This optimization was implemented using the *SPTGA* R package (Akdemir and Isidro-Sánchez, 2019) using 100 iterations: five solutions selected as elite parents were used to generating the next set of solutions and mutations of 80% for each solution generated.

4.2.11. Virtual screening for yield plasticity

Finally, we tested each GP model's potentials to predict the genotypes' phenotypic plasticity and stability across environments using only the $\boldsymbol{N}_{GE}$ phenotypic information. First, the prediction ability was computed for genotypes by correlating the predicted and observed grain yield values across environments (Costa-Neto *et al.*, 2021a). The second measure was based on the Finlay-Wilkinson adaptability model's regression slope (Finlay and Wilkinson, 1963). The GP predicted values were regressed to the observed environmental deviations, as follows:

$$M_{ij} = \bar{y}_{i.} + b_i I_j + \varepsilon_{ij} \qquad \text{(Eq. 7)}$$

where M_{ij} is the expected GP-based mean value of grain yield for i^{th} genotype at j^{th} environment, $\bar{y}_{i.}$ is the mean genotypic value for i^{th} genotype, b_i is the genotype plastic response across the mean-centered standardized environmental score (I_j) and ε_{ij} is the variety of residual deviation sources not accounted in the model. After this step, the Pearson's product-moment correlation between GP-based ($\hat{b}_i$) and phenotypic-enabled estimates were computed as an indicator of the ability to reproduce plastic responses *in silico* for the p genotypes. For this, the mean squared error is also calculated as:

$$MSE = \sum_{i=1}^{p} \frac{(b_i - \hat{b}_i)^2}{p} \qquad (Eq. 8)$$

All statistics were computed using the entire data sets and only the top 5% of genotypes selected for each environment. The latter aimed to check the efficiency of the E-GP method to produce high-quality virtual screenings for plasticity.

4.2.12. Software and Data Availability

All analyses were conducted using R statistical software (R Core Team, 2019). Data and codes are available at Supplementary Codes and Data into a web format as https://github.com/gcostaneto/EGP [verified 25th May 2021].

4.3. RESULTS

4.3.1. Case 1: Accuracy in predicting diverse G×E scenarios

A cross-validation scheme was designed to assess the predictive ability of the enviromic-aided approaches in the face of traditional GBLUP. For that, sample genotypes (70%) and environments were used to compose a drastically sparse training set for MET (training environments/total of environments). This helped assess the efficiency of E-GP for *Case 1*, in which we were able to dissect the predictive ability in different scenarios of a scarcity of phenotypic records: novel genotypes in tested environments (*n*G,E); tested genotypes in untested environments (G,*n*E), and novel genotype and environment conditions (*n*G, *n*E). Tables 1 and 2 present the N-level and Multi-Regional sets results, respectively. Then, these results were gathered for both data sets and four prediction scenarios in order to check for the joint predictive ability analysis (Figure 3).

The predictions within known environmental conditions of a certain experimental network involve scenarios *G,E* and *nG,E*. For the first scenario, all models outperformed the GBLUP in all experimental setups in both data sets. The highest values of predictive ability were observed for enviromic-aided GP models using the block-diagonal matrix for G×E effects (BD), that is, the E-GP (BD) and W-GP (BD), respectively. Two general trends were observed: the size of the experimental setup has a small effect on GP models' accuracy. Secondly, higher accuracy gains were observed for the N-level set (Table 1), with a higher number of entries (more genotypes and more environments). The accuracy gains in this N level set ranged from +8% (r = 0.83 for E-GP RN at 7/8 experimental setup), in relation to r = 0.77 (GBLUP), to +24% (r = 0.92 for W-GP RN at 3/8 experimental setup), in relation to r = 0.74 (GBLUP). In contrast, for the Multi-Regional set (Table 2), both RN-G×E models reduced the accuracy (on average, -3%). For the BD-G×E models, small gains in accuracy (from +4% to +8%) were observed.

That is also a trend for the second prediction scenario (*n*G,E), in which the Multi-Regional set presented an average gain of 10% for all enviromic-aided GP models with BD-G×E, and a reduction of 10% for all RN-G×E models. Conversely to the previous scenario (G,E, within the experimental network, using known genotypes), the *n*G,E is one of the most important plant breeding scenarios. Within the experimental network, it focuses on predicting novel genotypes). For the N-level set, gains up to 100% were observed for all enviromic-aided models using any G×E structure. No differences were observed between enviromic-aided models and experimental setups. On average, all enviromic-aided models achieved a predictive ability of approximately r = 0.66 across all experimental setups (3/8,

5/8, and 7/8, Table 1). In contrast, the GBLUP model has been impacted with reduced accuracy and a lack of phenotypic records. The highest gains in predictive ability were observed for scenario 3/8, average +118% for BD-G×E models, and +119% for RN-G×E models.

The predictions within new environmental conditions across the experimental network involve *G,nE* and *nG,nE*. For the former, the E-GP models outperformed W-GP and GBLUP across most experimental setups, despite small differences between the enviromic-aided approaches. For the E-GP BD at the N-level set (Table 1), the gains in predictive ability ranged from +24% (r = 0.49 at 7/8 setup, Table 1), in relation to r = 0.40 (GBLUP), to +35% (r = 0.57 at 5/8 setup), in relation to r = 0.43 (GBLUP). However, for scenario 3/8, these gains were equal to +10% (r = 0.57) in relation to the +13% archived by the benchmark W-GP RN (r = 0.58), both over the r = 0.53 from GBLUP. In scenario 7/8, W-GP was outperformed by GBLUP, with a reduction in accuracy between -18% and -16%, where the E-GP made better use of the large phenotypic information available for training GP models (gains from +20% to +24% over GBLUP). A similar pattern was observed for the Multi-Regional set (Table 2), in which the gains of E-GP ranged from +4% to +6% across all setups, and W-GP ranged from -3% to +6% under the same conditions.

The second scenario involving novel growing conditions also predicts novel genotypes (*nG,nE*) into account. Thus, all predictions were based on the phenotypic records from reassembled genotypes and considering the environmental similarity conceived from enviromics. With a large experimental network and genomics, the E-GP models outperformed W-GP and GBLUP when predicting new G×E. Observed accuracy gains ranged from 33% (r = 0.39 for E-GP RN) to 40% (r = 0.42 for E-GP BD), in experimental setup 7/8 (Table 1), where GBLUP achieved r = 0.30, and from 47% (r = 0.46 for E-GP BD) to 51% (r = 0.48 for E-GP BD), at the experimental setup 5/8, where GBLUP achieved r = 0.32. Unlike observations in the other prediction scenarios, the models RN-G×E outperformed BD-G×E in experimental setups 3/8 (N-Level set) and 2/5 (Multi-Regional set).

4.3.2. Accuracy trends across diverse experimental setups

This section highlights the main target of our *Case 1* study, in which the predictive ability was achieved using the merged information of scarce genotypes at some environments. Joint accuracy trends showed that E-GP was useful at increasing GP accuracy (Fig.3a) and explaining the phenotypic variation sources in both maize data sets (Supplementary Table 1). For scenarios with reduced phenotypic information (e.g., 3/5, 3/8, and 4/8), any model with some degree of environmental information outperformed the GBLUP for all scenarios. The E-GP approach (purple colors in Figure 3a) better captured envirotype-phenotype relations and converted them into accuracy gains among these models. This is also reflected in the E-GP efficiency as a predictive breeding tool capable of reproducing field-based trials (Fig.3b).

Regarding the G×E structures, the contribution of RN-G×E is significant only for drastically lacking phenotypic records (training setup 3/8), leading to the conclusion that the use of a main-effect is substantial for most cases E-GP is enough to increase accuracy in GBLUP. For setup 2/5 (Multi-Regional Set), no differences were observed between all the GP models.

The coincidence between the GP-based selection and the in-field selection (CS, %) ranged from ~35% to ~50%, in models with some environmental information, while it ranged between 30% and 40% for GBLUP (without environmental information). For the E-GP approach accounting for a wide number of phenotypic records in the training set (7/8, 3/5, and 4/5), values of CS up to 55% were found. Among these models, it seems that the RN-G×E reduces the CS estimates concerning the BD-G×E based models. Considering both figures 3a and 3b, it is possible to

suggest that predictive ability does not imply an increase of CS, that is, in the power of selecting the best performing genotypes in certain environments. However, the drastic increase in the E-GP accuracy in relation to the other models leads us to infer that despite the lower rise in CS, the E-GP models are useful when predicting GE for a vast number of single-crosses.

4.3.3. Case 2: enviromic assembly with optimized training sets

Those results lead us to investigate *Case 2* (Fig.2), where we checked the possibility of training efficient and biologically accurate GP scenarios from super-optimized training sets to predict virtual experimental networks. Thus, we combine two selective phenotyping approaches (Misztal, 2016; Akdemir and Isidro-Sánchez, 2019) to identify combinations of genotypes and environments using in-silico representations of the enviromic assembly × genomic kinships.

The process of designing virtual networks involved two steps. First, we used a single-value decomposition (SVD)-based algorithm to select the effective number of individuals (NGE) (Misztal, 2016) representing at least 98% of the variation of $\boldsymbol{K}_{G,ET}$. It was done in $\boldsymbol{K}_{G,ET}$ because this kernel represents an in-silico representation of envirotypes and genotypes (Akdemir and Isidro-Sánchez, 2019). Under sparse MET conditions, it led to a training size equal to NGE = 67 and NGE = 49 for the N-level set (n = 4,560) and Multi-Regional set (n = 1,235), respectively. It represents only 1.5% and 4% of the whole experimental network; Supplementary Fig. 1-2. For didactic purposes, from here onwards, we will represent the values of NGE as the training set size/number of genotypes.

We also checked the use of all environments, although the accuracy differences were tiny in relation to this sparse MET scenario (Table 3). Furthermore, small differences were achieved by E-GP and W-GP models with BD-G×E, but both higher than RN-G×E and GBLUP (Fig 4). Major differences were highlighted as follows. For within-field trials, predictive ability ranged from r = 0.76 (W-GP) to r = 0.87 (E-GP). Then, for virtual-networks, it ranged from r = 0.14±0.11 (GBLUP) to r = 0.60±0.06 (E-GP). In virtual-networks, the predictive ability of models trained with drastically reduced phenotypic records ranged from r = 0.10 (GBLUP, NGE = 67/4560) to r = 0.58 (E-GP, NGE = 67/4560) and r=0.18 (GBLUP, NGE = 49/1235) to r=0.81 (E-GP, NGE = 49/1235).

The predictive ability was computed considering only the top 5% of genotypes in each environment and data set. The objective was to verify if the GP approaches could adequately predict the performance of the best-evaluated genotypes in the field. For the Multi-Regional set, the predictive ability ranged from r = 0.098 (GBLUP, NGE = 210/1235) to r = 0.579 (W-GP BD, NGE = 49/1235) and r = 0.578 (E-GP BD, NGE = 49/1235); For the N-level set, W-GP outperformed E-GP, leading to r = 0.554 (W-GP BD, NGE = 536/4560) in front of r = 0.554 (E-GP RN, NGE = 67/4560) but with less phenotyping data. In contrast, the best E-GP model at the higher number of genotypes and environments evaluated in the field r = 0.484 (E-GP RN, NGE = 536/4560) were outperformed by the same model, yet with less phenotyping data r = 0.554 (E-GP RN, NGE = 67/4560). For GBLUP, the effective size of the training set was important, ranging in predictive ability from r = 0.070 (NGE = 67/4560) to r = 0.152 (NGE = 536/4560). The result of both sets suggests that when using enviromic-aided approaches, the use of fewer amounts of, but more representative, phenotyping information is better than more amounts of, yet less representative, phenotyping data.

Figure 4 was created using the average values of Table 3. This figure shows that the optimization was more effective for growing conditions contrasting across macro-regions (Fig. 4a) than for experimental networks involving fewer locations (Fig. 4b). Notably, it is possible to drastically reduce field costs for experimental networks conducted across diverse locations. However, for screening management conditions, greater precautions must be considered with the use of E-GP.

4.3.4. Predicting genotype-specific plasticity and environmental quality

In this step, we checked these models' ability to produce virtual screenings for yield plasticity (Fig.5). We used the Finlay-Wilkinson method (FW, Eq. 7) over the predicted GY means of each genotype *i* in environment *j* (M_{ij}). Hence, we compared the ability of E-GP in the prediction of: (1) individual genotypic responses across environments, (2) the gradient of environmental quality (h_j), and (3) the plasticity coefficient (b_1) of the FW model describing the rate of responsiveness to *h*. The results in Fig 5 involve both data sets (further details about each data set are given in Supplementary Fig 2-3).

All models that included some degree of enviromic assembly outperformed the GBLUP-based approach when predicting individual genotype responses across the MET (Fig 5a). The median values of *r* ranged from *r* =0.17 (GBLUP), in which 45% of the genotypes were not well predicted (red colors), to *r* = 0.83 (E-GP), in which up to 60% of the genotypes were very well predicted (purple colors). The inclusion of any enviromic assembly and G×E structure led to drastic gains in accuracy for a particular genotype response across contrasting (and unknown) G×E conditions (gains up to ~378%). However, the BD structure outperformed RN in terms of resolution (many purple colors in Fig 5a). A major part of the accurately predicted performance of genotypes across environments ranged from *r* = 0.75 to *r* =1.0. Due to this, for the next figures, we plotted only the E-GP considering the BD-G×E structure.

GBLUP was unable to correctly reproduce h_j for an in-silico study using the FW model (Fig.5b). We observe that E-GP better describes the h_j gradient (mean-centered average values of GY for each environment), with *r* near to 1 (correlation between observed and predicted environmental quality) also suggesting a low bias (*slope* = 0.924 between observed and predicted values). Consequently, this was reflected in the quality of yield plasticity predictions (Fig.5c-e), as yield plasticity was represented as linear responsiveness over the environmental variation. The graphical representation of genotype-specific linear reaction norms dictated by the linear regression slope (b_1) was likely more similar to E-GP than GBLUP about those observed in field-based testing (Fig 5b). The accuracy for b_1 ranged from *r* = 0.08 (GBLUP) to *r* = 0.43 (E-GP), an increase of 437%.

4.4. DISCUSSION

In this study, we presented the first report on (1) the use of Shelford's Law to guide the assembly of the enviromics for predictive breeding purposes over experimental networks; (2) the integration of enviromic assembly-based kernels with genomic kinship into optimization algorithms capable of designing selective phenotyping strategies and (3) a break of the paradigm relying on the fact that large phenotype data do not always contribute to increasing the accuracy of GP for contrasting G×E scenarios, but enviromics do. We report that the process of deriving markers

of environmental relatedness, here named 'enviromic assembly', is crucial for the implementation of low-cost GP platforms over multi-environmental conditions.

We suggest that the process of enviromic assembly is supported by a strong theoretical background in ecophysiology, illustrating the potential uses of environmental information to increase the accuracy of predictive breeding for yield and plasticity. Our results indicate that the E-GP platform (Figure 2) can fit two types of prediction scenarios in plant breeding programs: (1) better use of the available phenotypic records to train more accurate GP models capable of aiding the selection of genotypes across multi-environmental conditions and (2) a method that reduces costs for field-based testing and enables an early screening for yield plasticity under crossover G×E conditions. Furthermore, we show that any model with some degree of enviromic assembly (by typology or quantitative descriptors) is always better to reproduce the genotypes' environmental quality of field trials and phenotypic plasticity.

Below we discuss the aspects that support the use of E-GP for multi-environment predictions, involving the importance of breaking the paradigm that states that enviromics are not necessary to predict G×E accurately. We then discuss how the genomic and enviromic sources are linked in the phenotypic records collected from the fields and how this type of knowledge can improve the quality of the prediction-based pipelines for crop improvement. Finally, we envisage possible environmental-assembly applications supporting other predictive breeding fields, such as optimizing crop modeling calibration and how it can couple a novel level of climate-smart solutions for crop improvement as anticipating the plasticity of a large number of genotypes using reduced phenotypic data.

4.4.1. Importance of enviromics for multi-environment genomic prediction

Genomic prediction (GP) platforms were first designed to model the *genotype-to-phenotype* relations under single environment conditions, e.g., in a breeding program nursery (Lorenzana and Bernardo, 2009; Windhausen *et al.*, 2012; Zhao *et al.*, 2012; Zhang *et al.*, 2015). Under these conditions, the micro-environmental variations within breeding trials (e.g., spatial gradients in soil properties) are minimized in the phenotypic correction step by separating useful genetic patterns and experimental noises (non-genetic patterns) (Galli *et al.*, 2018). However, those phenotypic records carry the indissoluble effects of macro-environmental fluctuations of certain weather and soil factors that occurred during crop growth and development (Li *et al.*, 2018; Vidotti *et al.*, 2019; Millet *et al.*, 2019; Guo *et al.*, 2020; Jarquín *et al.*, 2020). That seems to be of no concern when predicting novel genotypes under these same growth conditions (the CV1 scheme for single-environment models) yet becomes noise for multi-environment prediction scenarios. It is a consequence of the macro-environment fluctuations in the lifetime of the crops (Allard and Bradshaw, 1964; Bradshaw, 1965; Arnold *et al.*, 2019), responsible for modulating the rate of gene expression (e.g., Jończyk *et al.*, 2017; Liu *et al.*, 2020) and fine-tuning epigenetic variations related to transcriptional responses (Vendramin *et al.*, 2020).

For each unit that we call "environment" (field trial at the specific year, location, planting date, and crop management), there are various environmental factors such as water availability, canopy temperature, global solar radiation, and nutrient content in the soil. Shelford suggests that a population's fitness is given by the amount and distribution of resources available for its establishment and adaptation (Shelford, 1931). Thus, we reinterpret this concept by assuming that the relation between input availability (deficit, optimum amount, or excess) across different crop development stages drives the amount of the genetic potential expressed in phenotypes produced by the same genotype for a given environment. Therefore, there is also an indissoluble *envirotype-phenotype covariance* in the phenotypic

records that is interpreted as a G×E interaction for each environment. Thus, the pioneer approaches to measuring crop adaptability use the average value of a given trait in a given environment as an environmental quality index (e.g., Finlay and Wilkinson, 1963). However, the problem with this approach is that it explains the quality of the environment realized by the genotypes evaluated in it, making it inefficient to explain the drivers of environmental quality and incapable of predicting untested growing conditions, as observed in our results for *Case 2* using GBLUP without enviromic data. In addition, our results for *Case 1* highlight that it is a limit in accuracy for traditional GBLUP across MET, in which the accuracy remains almost the same, regardless of the number of phenotypic records available.

A second intrinsic covariance can interpret this last result within the phenotypic records, which is the *genotype-envirotype covariance*. By adapting the Quantitative Genetics theory to the terminology used here, we can infer that each genotype reacts differently to each envirotype, resulting in a given phenotype. This phenotype is then used to provide small crop phenology differences (genetically determined window sizes for each development stage). Pioneer works have been carried out to understand the genetic and environmental determinants of flowering time in sorghum (Li *et al.*, 2018) and rice (Guo *et al.*, 2020). That can be indirectly interpreted as cardinal differential thresholds for temperature response. Jarquín *et al* (2020) proved that it is possible to increase the ability of GP in predictive novel G×E by coupling information of day-length in the benchmark GP models. For all these examples reported above, we can infer that, when trying to predict a novel genotype, by borrowing genotypic information from the relatives at different environments, it is impossible to reproduce the genotype-envirotype covariance without adding any enviromic information into the model.

The presence of both *genotype-envirotype* and *envirotype-phenotype* covariances might explain the gains in the predictive ability due to the use of multi-environment GP models in contrast to single-environment GP models (Bandeira e Souza *et al.*, 2017; de Oliveira *et al.*, 2020) and why deep learning approaches have successfully captured intrinsic G×E patterns and translated them into gains in accuracy (Montesinos-López *et al.*, 2018; Crossa *et al.*, 2019; Cuevas *et al.*, 2019). Conversely, this also might explain the need to incorporate secondary sources of information in the prediction of grain yields across multiple environments (Westhues *et al.,* 2017; Ly *et al.,* 2018; Millet *et al.,* 2019; Costa-Neto *et al.*, 2021a; 2021b; Jarquín *et al.*, 2020), as well as the possible limitations of CGM approaches contrasting scenarios differing from those targeted near-iso conditions of CGM calibration (e.g., Cooper *et al.*, 2016; Messina *et al.*, 2018). Thus, an alternative can be supervised approaches to describe the environmental relatedness, such as in this paper, and perhaps unsupervised algorithms capable of taking advantage of the covariances related to the genotype-phenotype, genotype-envirotype, and envirotype-phenotype dynamics.

4.4.2. Sometimes main-effect enviromics is better than reaction-norm models

Our results from *Case 1* show that the inclusion of enviromic sources (for main-effects or explicitly incorporated in the RN-G×E structure) led to a better description of the envirotype-phenotype covariances, which was reflected in accuracy gains. It is worth highlighting that incorporating enviromic sources does not replace the incorporation of a design matrix for environments (here used as fixed effects) as it is commonly associated in previous studies of GP reaction-norm. Here we show that enviromic sources came up as tentative to capture the envirotype-phenotype covariances. The cross-validation scheme used in *Case 1* allowed us to observe that the joint prediction of different genotype-environment conditions (Fig 3) might better highlight how enviromic sources can contribute to increasing the predictive ability of GP, mostly due to its usefulness in approaching the environmental correlation

among field trials. It shows more transparency for the influence of the scenarios G,*n*E and *n*G*n*E, in which we had a considerable lack of phenotypic information in training GP. We can infer that schemes such as CV1 (only *n*G,E) are the least adequate option to show the benefits of coupling enviromics in GBLUP. However, looking at a drastically sparse MET condition (joint prediction scenarios) shows that enviromics improves the accuracy of GP as the size of the MET also increases. Predictions are made up of tiny experimental networks.

4.4.3. Differences in using environmental covariables (W) and typologies (T)

Regarding the enviromic assembly approaches used in this study, there was evidence that using typologies as envirotype descriptors (**T** matrix) is more biologically accurate in representing environmental relatedness than quantitative descriptors (**W** matrix) based on quantile covariables. This increase in biological accuracy was reflected in the statistical accuracy and then boosted plant breeders' ability to carry out selections across multi-environment conditions. Further efforts in this sense must be devoted to increasing the level of explanation of the genotype-envirotype covariances, which can also take advantage of Shelford's Law to refine the limits of tolerance for particular genotypes. Thus, different genotypes will be under the influence of a diverse set of envirotypes, which can be realized for the same environmental factor (e.g., solar radiation, air temperature, soil moisture) according to its occurrence across crop lifetime (e.g., vegetative stage) and the adaptation zone designed from ecophysiology concepts (e.g., temperature cardinals defining which temperature level results in stress and optimum growing condition).

A second difference may be explained by the fact that quantitative environmental covariates are not an additive effect to compose an environment variation. Despite this, we agree with Resende et al. (2020), and we adapted the idea of envirotypes as markers of environment relatedness in a different manner. For example, the common use of mean values of covariates such as rainfall, solar radiation, and air temperature, in reality, represents a non-additive between each other; yet, they are very well correlated for a given site-planting date condition, even when using strategies to deal with collinearity, such as partial least squares (e.g., Vargas *et al.*, 2006; Porker *et al.,* 2020;). We can use an example as a given day of crop growing in which a large amount of rainfall has occurred. We can suppose that the sky is cloudy, with less radiation and lower temperature. Thus, using such G-BLUP inspired approach is not an ideal solution to estimate the environmental variance. Conversely, the environmental typologies (**T**) are based on frequencies (ranging from 0 to 1), where the sum of all frequencies are equal to 1 (100% of the variation). In addition, those typologies can be built for a given site using historical weather data, adapting the approach of Gillberg et al. (2019) and de los Campos et al. (2020). If no typologies are considered, the expected environment effect is given for a fixed-environment intercept (with 0 variance within and between environments). Despite this fact, another option is using nonlinear kernel methods to estimate only the environment-relatedness, as this approach takes advantage of nonlinear relationships among covariates (Costa-Neto et al., 2021a, b).

4.4.4. Balance between size of the experimental network and model accuracy

This study shows that environmental information can break the paradigm that claims that more phenotype information leads to greater accuracy of GP models over MET. Our results highlight that the traditional GBLUP models assume that the variation due to G×E is purely genomic-based across field trials, leading to an implicit

conclusion that the yield plasticity is constant (slope ~ 0) for all genotypes, which is unrealistic. It also reflects that G×E patterns are non-crossover (scale changes in performance across different variations), that is, a well-performing genotype will always be good across environments, and a poorly performing genotype has the same trend for all environments. Despite the gains achieved in predicting the quality of a novel environment and the plasticity for tested and untested genotypes, we noticed that the inclusion of enviromic sources also leads to the unrealistic conclusion that all genotypes respond in the same way the gradient of climate and soil quality. Our results show a reasonable accuracy in predicting yield plasticity, but further efforts must be made to improve this approach's explanation of the yield plasticity as a nonlinear variation across the gradient of environmental factors.

The use of selective phenotyping strategies made up with enviromic assembly × genomic kinships showed a drastic reduction of in-field efforts. Combined with enviromic-aided GBLUP models, it led to almost the same predictive ability achieved using a wide number of genotypes and environments for a large experimental network. Thus, we can enumerate the benefits of the enviromic approaches tested in this study as (1) the possibility of training prediction models for yield plasticity with reduced phenotyping efforts, (2) a consequence of the assembly of enviromics with genomics allowing the selection of the genotype-environment combinations that best represents the main inner covariances among phenotypes produced by different environments (the genotype-phenotype, envirotype-phenotype dynamics mentioned above).

Considering both enviromics approached, we conclude that the advantages of E-GP over W-GP can be enumerated as (1) the flexibility to design a wide number of environment-types assuming different frequencies of occurrence of key stressful factors in crop development; (2) it allows the use of historical weather and in-field records to compute trends of certain envirotypes at certain environments, which can be coupled into (3) the definition of TPE and characterization of mega-environments, as the main approach used for this relies on the study of the frequency of occurrence of the main environment-types (e.g., Heinemann *et al.*, 2019). For the latter, for example, the **T** matrix proposed here is just an arrangement of an environment × typology matrix, in which each entry represents its frequency of occurrence at a particular time interval of the crop lifetime. Conversely, the advantages of W-GP over E-GP rely on plasticity in creating large-scale envirotype descriptors with reasonable biological accuracy.

4.4.5. Climate-smart solutions from enviromics with genomics

Modern plant breeding programs must deliver climate-smart solutions cost-effectively and time-reduced (Crossa et al., 2021). By climate-smart solutions, we mean (1) adopting cost-effective approaches capable of providing fast and cheap solutions to face climate change (2) a better resource allocation for field trial efforts to collect representative phenotype information to feed prediction-based platforms for crop improvement, such as training accurate GP models and CGM-based approaches capable of guiding several breeding decisions, (3) a better understanding of which envirotypes most limit the adaptation of crops across the breeding TPE, revising historical trends and expecting future scenarios (e.g., Ramirez-Villegas et al., 2018; 2020; Heinemann et al., 2019) (4) understanding the relationship between secondary traits and their importance in explaining the plant-environment dynamics for given germplasm at given TPE (e.g., Cooper et al., 2021). However, most of those objectives will be hampered if the MET-GP platforms do not consider models with a higher biological meaning (Hammer et al., 2019) and reliable environmental information. A cost-effective solution for that, if the breeder has no access to sensor network tools, relies on the use of remote sensing tools to collect and process basic weather and soil data, such as those available in the EnvRtype R package (Costa-Neto et al., 2021b).

If selective phenotyping is added in the enviromics-aided pipeline for GP, additional traits and the possibility of screening genotypes across a wide number of managed environments will increase. It can support field trials' training for CGM approaches, which demands phenotyping of traits across crop life, such as biomass accumulation and partitioning among different plant organs. Finally, using models considering an explicit environmental gradient of key-environmental factors is a second alternative for this approach. It can be done to discover the genetic determinants of the interplay between plant plasticity and environment variation. As a wide range of genes reacts to each gradient of environmental factors, the use of whole-genome regressions of reaction-norm for each environmental factor must be useful to screen potential genotypes (in our case, single-crosses) for a diverse set of scenarios (e.g., increased heat stress). Pioneer works used this methodology in wheat breeding (Heslot *et al.*, 2014; Ly *et al.*, 2018) inspired other cereal crop applications.

For example, Millet *et al.* (2019) fine-tuned the methodology by creating a two-stage analysis of factorial regression (FR) involving environmental data, followed by a GP based on the genotypic-specific sensibility for key environmental factors found in the FR step. In general, studies involving FR analysis found that the effect of high temperatures at grain-filling and maturation (Epinat-Le Signor *et al.*, 2001; Romay *et al.*, 2010), water balance at flowering (Epinat-Le Signor *et al.*, 2001; Millet *et al.*, 2019) and intercept radiation at the vegetative phase (Millet *et al.*, 2019) are the main drivers of G×E for yield components in maize. Thus, Millet *et al.* (2019) explores this opportunity offered by FR to use genotypic-specific regressions, which coupled with genomic data, led to an increase of the accuracy of MET-GP by 55% concerning the benchmark environmental similarity model made up of mean values of environmental factors, as proposed by Jarquín *et al.* (2014).

From the aspects mentioned above, we envisage that the use of GP for multi-environment predictions must account for some degree of ecophysiological reality while also considering the balance and the relation between parsimony and accuracy (Hammer *et al.*, 2019; Costa-Neto *et al.*, 2021b; Cooper *et al.*, 2021). Here we also highlight in our literature review that multi-environment GP must account for the impact of (1) resource availability in the creation of biologically accurate platforms in training CGM-based approaches and delivering reliable envirotyping information for those purposes, (2) availability of the knowledge of experts in training CGM approaches. Thus, ecophysiology concepts to provide solutions for raw environmental data processing in enviromic assembly information for predictive purposes seem to be a cost-effective alternative to leverage accuracy involving parsimony and biological reality.

References

Allard, R. W., and Bradshaw, A. D. (1964). Implications of genotype-environmental interactions in applied plant breeding. *Crop Sci.* 4, 503–508.

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M. (1998). *Crop Evapotranspiration – guidelines for computing crop water requirements*. 56th ed. Rome: FAO Irrigation and Drainage Paper Nº 56.

Antolin, L. A. S., and Heinemann, A. B. (2021). Impact assessment of common bean availability in Brazil under climate change scenarios. 191. doi:10.1016/j.agsy.2021.103174.

Arnold, P. A., Kruuk, L. E. B., and Nicotra, A. B. (2019). How to analyse plant phenotypic plasticity in response to a changing climate. *New Phytol.* 222, 1235–1241. doi:10.1111/nph.15656.

Alves, F. C., Granato, Í. S. C., Galli, G., Lyra, D. H., Fritsche-Neto, R., and De Los Campos, G. (2019). Bayesian analysis and prediction of hybrid performance. *Plant Methods* 15, 1–18. doi:10.1186/s13007-019-0388-x.

Akdemir, D., and Isidro-Sánchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. *Sci. Rep.* 9, 1–15. doi:10.1038/s41598-018-38081-6.

Bandeira e Souza, M., Cuevas, J., Couto, E. G. de O., Pérez-Rodríguez, P., Jarquín, D., Fritsche-Neto, R., et al. (2017). Genomic-Enabled Prediction in Maize Using Kernel Models with Genotype × Environment Interaction. *G3* 7, g3.117.042341. doi:10.1534/g3.117.042341.

Bournan, B. A. M., Keulen, H. Van, and Rabbingeh, R. (1996). The 'School of de Wit' Crop Growth Simulation Models : A Pedigree and Historical Overview. 52.

Bradshaw, A. D. (1965). Evolutionary significance of phenotypic plasticity in plants. *Adv. Genet.* 13, 115–155.

Cooper, M., Messina, C. D., Podlich, D., Totir, L. R., Baumgarten, A., Hausmann, N. J., et al. (2014). Predicting the future of plant breeding: Complementing empirical evaluation with genetic prediction. *Crop Pasture Sci.* 65, 311–336. doi:10.1071/CP14007.

Cooper, M., Technow, F., Messina, C., Gho, C., and Radu Totir, L. (2016). Use of crop growth models with whole-genome prediction: Application to a maize multienvironment trial. *Crop Sci.* 56, 2141–2156. doi:10.2135/cropsci2015.08.0512.

Cooper, M., Powell, O., Voss-Fels, K. P., Messina, C. D., Gho, C., Podlich, D. W., et al. (2021). Modelling selection response in plant-breeding programs using crop models as mechanistic gene-to-phenotype (CGM-G2P) multi-trait link functions. *in silico Plants* 3, 1–21. doi:10.1093/insilicoplants/diaa016.

Cortés, A. J., Restrepo-Montoya, M., and Bedoya-Canas, L. E. (2020). Modern Strategies to Assess and Breed Forest Tree Adaptation to Changing Climate. *Front. Plant Sci.* 11. doi:10.3389/fpls.2020.583323.

Costa-Neto, G., Fritsche-Neto, R., and Crossa, J. (2021a). Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. *Heredity (Edinb).* 126, 92–106. doi:10.1038/s41437-020-00353-1.

Costa-Neto, G., Galli, G., Carvalho, H. F., Crossa, J., and Fritsche-Neto, R. (2021b). EnvRtype: a software to interplay enviromics and quantitative genomics in agriculture. *G3 Genes|Genomes|Genetics*. doi:10.1093/g3journal/jkab040.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends Plant Sci.* 22, 961–975. doi:10.1016/j.tplants.2017.08.011.

Crossa, J., Martini, J. W. R., Gianola, D., Pérez-Rodríguez, P., Jarquin, D., Juliana, P., et al. (2019). Deep Kernel and Deep Learning for Genome-Based Prediction of Single Traits in Multienvironment Breeding Trials. *Front. Genet.* 10, 1–13. doi:10.3389/fgene.2019.01168

Crossa, J., Fritsche-Neto, R., Montesinos-lopez, O. A., Costa-Neto, G., Dreisigacker, S., Montesinos-lopez, A., et al. (2021). The Modern Plant Breeding Triangle : Optimizing the Use of Genomics, Phenomics, and Enviromics Data. *Front. Plant Sci.* 12, 1–6. doi:10.3389/fpls.2021.651480.

Cuevas, J., Montesinos-López, O., Juliana, P., Guzmán, C., Pérez-Rodríguez, P., González-Bucio, J., et al. (2019). Deep Kernel for genomic and near infrared predictions in multi-environment breeding trials. *G3 Genes, Genomes, Genet.* 9, 2913–2924. doi:10.1534/g3.119.400493.

de los Campos, G., Pérez-Rodríguez, P., Bogard, M., Gouache, D., and Crossa, J. (2020). A data-driven simulation platform to predict cultivars' performances under uncertain weather conditions. *Nat. Commun.* 11. doi:10.1038/s41467-020-18480-y.

de Oliveira, A. A., Resende, M. F. R., Ferrão, L. F. V., Amadeu, R. R., Guimarães, L. J. M., Guimarães, C. T., et al. (2020). Genomic prediction applied to multiple traits and environments in second season maize hybrids. *Heredity (Edinb).* 125, 60–72. doi:10.1038/s41437-020-0321-0.

Dias, K. O. D. G., Gezan, S. A., Guimarães, C. T., Nazarian, A., Da Costa E Silva, L., Parentoni, S. N., et al. (2018). Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. *Heredity (Edinb).* 121, 24–37. doi:10.1038/s41437-018-0053-6.

Epinat-Le Signor, C., Dousse, S., Lorgeou, J., Denis, J.-B., Bonhomme, R., Carolo, P., et al. (2001). Interpretation of genotype x environment interactions for early maize hybrids over 12 years. *Crop Sci.* 41, 663–669. doi:10.2135/cropsci2001.413663x.

Finlay, K. W., and Wilkinson, G. N. (1963). The analysis of adaptation in a plant breeding programme. *J. Agric. Res.* 14, 742–754.

Gage, J. L., Jarquin, D., Romay, C., Lorenz, A., Buckler, E. S., Kaeppler, S., et al. (2017). The effect of artificial selection on phenotypic plasticity in maize. *Nat. Commun.* 8. doi:10.1038/s41467-017-01450-2.

Galli, G., Lyra, D. H., Alves, F. C., Granato, Í. S. C., e Sousa, M. B., and Fritsche-Neto, R. (2018). Impact of phenotypic correction method and missing phenotypic data on genomic prediction of maize hybrids. *Crop Sci.* 58, 1481–1491. doi:10.2135/cropsci2017.07.0459.

Galli, G., Sabadin, F., Costa-Neto, G. M. F., and Fritsche-Neto, R. (2021). A novel way to validate UAS-based high-throughput phenotyping protocols using in silico experiments for plant breeding purposes. *Theor. Appl. Genet.* 134, 715–730. doi:10.1007/s00122-020-03726-6.

Gillberg, J., Marttinen, P., Mamitsuka, H., and Kaski, S. (2019). Modelling G×E with historical weather information improves genomic prediction in new environments. *Bioinformatics* 35, 4045–4052. doi:10.1093/bioinformatics/btz197.

Granato, I., Cuevas, J., Luna-Vázquez, F., Crossa, J., Montesinos-López, O., Burgueño, J., et al. (2018). BGGE: A new package for genomic-enabled prediction incorporating genotype × environment interaction models. *G3 Genes, Genomes, Genet.* 8, 3039–3047. doi:10.1534/g3.118.200435.

Guo, T., Mu, Q., Wang, J., Vanous, A. E., Onogi, A., Iwata, H., et al. (2020). Dynamic effects of interacting genes underlying rice flowering-time phenotypic plasticity and global adaptation. *Genome Res.* 30, 673–683. doi:10.1101/gr.255703.119.

Hammer, G., Messina, C., Wu, A., and Cooper, M. (2019). Biological reality and parsimony in crop models — why we need both in crop improvement ! 1–21. doi:10.1093/insilicoplants/diz010.

Heinemann, A. B., Ramirez-Villegas, J., Rebolledo, M. C., Costa Neto, G. M. F., and Castro, A. P. (2019). Upland rice breeding led to increased drought sensitivity in Brazil. *F. Crop. Res.* 231, 57–67. doi:10.1016/j.fcr.2018.11.009.

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, 463–480.

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., et al. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet.* 127, 595–607. doi:10.1007/s00122-013-2243-1.

Jarquin, D., Kajiya-Kanegae, H., Taishen, C., Yabe, S., Persa, R., Yu, J., et al. (2020). Coupling day length data and genomic prediction tools for predicting time-related traits under complex scenarios. *Sci. Rep.* 10, 1–12. doi:10.1038/s41598-020-70267-9.

Jończyk, M., Sobkowiak, A., Trzcinska-Danielewicz, J., Skoneczny, M., Solecka, D., Fronk, J., et al. (2017). Global analysis of gene expression in maize leaves treated with low temperature. II. Combined effect of severe cold (8 °C) and circadian rhythm. *Plant Mol. Biol.* 95, 279–302. doi:10.1007/s11103-017-0651-3.

Lorenzana, R. E., and Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theor. Appl. Genet.* 120, 151–161. doi:10.1007/s00122-009-1166-3.

Li, X., Guo, T., Mu, Q., Li, X., and Yu, J. (2018). Genomic and environmental determinants and their interplay underlying phenotypic plasticity. *Proc. Natl. Acad. Sci. U. S. A.* 115, 6679–6684. doi:10.1073/pnas.1718326115.

Liu, S., Li, C., Wang, H., Wang, S., Yang, S., Liu, X., et al. (2020). Mapping regulatory variants controlling gene expression in drought response and tolerance in maize. *Genome Biol.* 21, 1–22. doi:10.1186/s13059-020-02069-1.

Ly, D., Huet, S., Gauffreteau, A., Rincent, R., Touzy, G., Mini, A., et al. (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. doi:10.1016/j.fcr.2017.08.020.

Martini, J. W. R., Crossa, J., Toledo, F. H., and Cuevas, J. (2020). On Hadamard and Kronecker products in covariance structures for genotype × environment interaction. *Plant Genome* 13, 1–12. doi:10.1002/tpg2.20033.

Messina, C. D., Technow, F., Tang, T., Totir, R., Gho, C., and Cooper, M. (2018). Leveraging biological insight and environmental variation to improve phenotypic prediction: Integrating crop growth models (CGM) with whole genome prediction (WGP). *Eur. J. Agron.* 100, 151–162.

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

Montesinos-López, A., Montesinos-López, O. A., Gianola, D., Crossa, J., and Hernández-Suárez, C. M. (2018). Multi-environment genomic prediction of plant traits using deep learners with dense architecture. *G3 Genes, Genomes, Genet.* 8, 3813–3828. doi:10.1534/g3.118.200740.

Millet, E. J., Kruijer, W., Coupel-Ledru, A., Alvarez Prado, S., Cabrera-Bosquet, L., Lacube, S., et al. (2019). Genomic prediction of maize yield across European environmental conditions. *Nat. Genet.* 51, pages952–956. doi:10.1038/s41588-019-0414-y.

Misztal, I. (2016). Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics* 202, 401–409. doi:10.1534/genetics.115.182089.

Morais Júnior, O. P., Duarte, J. B., Breseghello, F., Coelho, A. S. G., and Magalhães Júnior, A. M. (2018). Single-step reaction norm models for genomic prediction in multienvironment recurrent selection trials. *Crop Sci.* 58, 592–607. doi:10.2135/cropsci2017.06.0366.

Porker, K., Coventry, S., Fettell, N. A., Cozzolino, D., and Eglinton, J. (2020). Using a novel PLS approach for envirotyping of barley phenology and adaptation. *F. Crop. Res.* 246, 1–11. doi:10.1016/j.fcr.2019.107697.

Ramirez-Villegas, J., Heinemann, A. B., Castro, A. P., Breseghello, F., Navarro-Racines, C., Li, T., et al. (2018). Breeding implications of drought stress under future climate for upland rice in Brazil. *Glob. Chang. Biol.* 1, 1–16. doi:10.1111/ijlh.12426.

Ramirez-Villegas, J., Molero Milan, A., Alexandrov, N., Asseng, S., Challinor, A. J., Crossa, J., et al. (2020). CGIAR modeling approaches for resource-constrained scenarios: I. Accelerating crop breeding for a changing climate. *Crop Sci.* 60, 547–567. doi:10.1002/csc2.20048.

Resende, R. T., Piepho, H. P., Rosa, G. J. M., Silva-Junior, O. B., e Silva, F. F., de Resende, M. D. V., et al. (2020). Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. *Theor. Appl. Genet.* doi:10.1007/s00122-020-03684-z.

Robert, P., Le Gouis, J., and Rincent, R. (2020). Combining Crop Growth Modeling With Trait-Assisted Prediction Improved the Prediction of Genotype by Environment Interactions. *Front. Plant Sci.* 11, 1–11. doi:10.3389/fpls.2020.00827.

Rogers, A. R., Dunne, J. C., Romay, C., Bohn, M., Buckler, E. S., Ciampitti, I. A., et al. (2021). The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. *G3 Genes|Genomes|Genetics* 11. doi:10.1093/g3journal/jkaa050.

Romay, M. C., Malvar, R. A., Campo, L., Alvarez, A., Moreno-González, J., Ordás, A., et al. (2010). Climatic and genotypic effects for grain yield in maize under stress conditions. *Crop Sci.* 50, 51–58.

Shelford, V. . E. . (1931). Some Concepts of Bioecology. *Ecology* 12, 455–467.

Sparks, A. (2018). nasapower: A NASA POWER Global Meteorology, Surface Solar Energy and Climatology Data Client for R. *J. Open Source Softw.* 3, 1035. doi:10.21105/joss.01035.

Tigchelaar, M., Battisti, D. S., Naylor, R. L., and Ray, D. K. (2018). Future warming increases probability of globally synchronized maize production shocks. *Proc. Natl. Acad. Sci. U. S. A.* 115, 6644–6649. doi:10.1073/pnas.1718031115.

Toda, Y., Wakatsuki, H., Aoike, T., Kajiya-Kanegae, H., Yamasaki, M., Yoshioka, T., et al. (2020). Predicting biomass of rice with intermediate traits: Modeling method combining crop growth models and genomic prediction models. *PLoS One* 15, 1–21. doi:10.1371/journal.pone.0233951.

Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., et al. (2014). A powerful tool for genome analysis in maize: Development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics* 15, 1–15. doi:10.1186/1471-2164-15-823.

Vargas, M., Van Eeuwijk, F. A., Crossa, J., and Ribaut, J. M. (2006). Mapping QTLs and QTL x environment interaction for CIMMYT maize drought stress program using factorial regression and partial least squares methods. *Theor. Appl. Genet.* 112, 1009–1023. doi:10.1007/s00122-005-0204-z.

Voss-Fels, K. P., Cooper, M., and Hayes, B. J. (2019). Accelerating crop genetic gains with genomic selection. *Theor. Appl. Genet.* 132, 669–686. doi:10.1007/s00122-018-3270-8.

Vendramin, S., Huang, J., Crisp, P. A., Madzima, T. F., and McGinnis, K. M. (2020). Epigenetic regulation of ABA-induced transcriptional responses in maize. *G3 Genes, Genomes, Genet.* 10, 1727–1743. doi:10.1534/g3.119.400993.

Vidotti, M. S., Matias, F. I., Alves, F. C., Rodríguez, P. P., Beltran, G. A., Burguenõ, J., et al. (2019). Maize responsiveness to Azospirillum brasilense: Insights into genetic control, heterosis and genomic prediction. *PLoS One* 14, 1–22. doi:10.1371/journal.pone.0217571.

Westhues, M., Schrag, T. A., Heuer, C., Thaller, G., Utz, H. F., Schipprack, W., et al. (2017). Omics-based hybrid prediction in maize. *Theor. Appl. Genet.* doi:10.1007/s00122-017-2934-0.

Windhausen, V. S., Atlin, G. N., Hickey, J. M., Crossa, J., Jannink, J.-L., Sorrells, M. E., et al. (2012). Effectiveness of Genomic Prediction of Maize Hybrid Performance in Different Breeding Populations and Environments. *G3:Genes|Genomes|Genetics* 2, 1427–1436. doi:10.1534/g3.112.003699.

Xu, Y. (2016). Envirotyping for deciphering environmental impacts on crop plants. *Theor. Appl. Genet.* 129, 653–673. doi:10.1007/s00122-016-2691-5.

Zhao, Y., Gowda, M., Liu, W., Würschum, T., Maurer, H. P., Longin, F. H., et al. (2012). Accuracy of genomic selection in European maize elite breeding populations. *Theor. Appl. Genet.* 124, 769–776. doi:10.1007/s00122-011-1745-y.

Zhang, X., Pérez-Rodríguez, P., Semagn, K., Beyene, Y., Babu, R., López-Cruz, M. A., et al. (2015). Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. *Heredity (Edinb).* 114, 291–299. doi:10.1038/hdy.2014.99.

TABLES

Tabela 1. Predictive ability (± standard error) of the genome-based prediction models (GP) for the N-level set of tropical maize hybrids (570 hybrids × 2 locations × two years × two nitrogen managements). Bold values denote higher predictive ability values for each scenario: G,E (known genotypes at known growing conditions), G,*n*E (known genotypes at new growing conditions), *n*G, E (new genotypes at known growing conditions), and *n*G, *n*E (new genotypes at new growing conditions).

Training Setup	Model	Prediction Scenario			
		G, E	**G, *n*E**	***n*G, E**	***n*G, *n*E**
7/8 Environments	GBLUP	0.771± 0.064	0.397±0.046	0.310±0.054	0.297±0.029
	E-GP (BD)	**0.903**±0.115	**0.493**±0.169	**0.615**±0.022	**0.416**±0.153
	E-GP (RN)	0.833±0.118	**0.477**±0.199	**0.613**±0.040	0.394±0.193
	W-GP (BD)	**0.915**±0.115	0.333±0.208	**0.614**±0.025	0.242±0.189
	W-GP (RN)	0.885±0.117	0.327±0.210	0.613±0.031	0.23±0.196
5/8 Environments	GBLUP	0.747±0.049	0.432±0.046	0.294±0.026	0.323±0.04
	E-GP (BD)	0.905±0.056	**0.554**±0.144	**0.659**±0.015	**0.464**±0.113
	E-GP (RN)	0.833±0.056	**0.570**±0.132	**0.660**±0.025	**0.475**±0.104
	W-GP (BD)	**0.931**±0.057	0.449±0.286	0.659±0.019	0.347±0.253
	W-GP (RN)	0.897±0.056	0.501±0.229	**0.660**±0.026	0.395±0.198
3/8 Environments	GBLUP	0.739±0.040	0.527±0.080	0.295±0.015	0.394±0.044
	E-GP (BD)	0.899±0.026	0.534±0.081	0.660±0.012	0.388±0.038
	E-GP (RN)	0.823±0.026	**0.566**±0.086	**0.663**±0.015	**0.420**±0.041
	W-GP (BD)	**0.924**±0.026	0.532±0.08	0.660±0.015	0.384±0.038
	W-GP (RN)	0.886±0.025	**0.579**±0.088	**0.663**±0.020	**0.424**±0.041

Tabela 2. Predictive ability (± standard error) of the genome-based prediction models (GP) for the Multi-Local set of tropical maize hybrids (247 hybrids × 5 locations in different regions of Brazil). Bold values denote the higher predictive ability values for each scenario: G,E (known genotypes at known growing conditions), G,*n*E (known genotypes at new growing conditions), *n*G, E (new genotypes at known growing conditions), and *n*G, *n*E (new genotypes at new growing conditions).

Training Setup	Model	Prediction Scenario			
		G, E	**G, *n*E**	***n*G, E**	***n*G, *n*E**
4/5 Environments	GBLUP	0.953±0.040	0.497±0.072	0.552±0.171	0.340±0.138
	E-GP (BD)	**0.987**±0.006	**0.526**±0.054	**0.599**±0.097	**0.363**±0.131
	E-GP (RN)	0.873±0.084	0.520±0.064	0.496±0.126	**0.358**±0.143
	W-GP (BD)	**0.989**±0.005	**0.527**±0.056	**0.599**±0.098	0.361±0.131
	W-GP (RN)	0.931±0.057	0.492±0.078	0.501±0.130	**0.366**±0.125
3/5 Environments	GBLUP	0.927±0.045	0.528±0.066	0.543±0.208	0.381±0.142
	E-GP (BD)	**0.984**±0.006	**0.556**±0.052	**0.597**±0.097	**0.400**±0.131
	E-GP (RN)	0.845±0.073	0.550±0.059	0.477±0.120	0.385±0.135
	W-GP (BD)	**0.987**±0.005	**0.555**±0.053	**0.598**±0.095	**0.394**±0.132
	W-GP (RN)	0.915±0.049	0.514±0.072	0.483±0.124	0.392±0.119
2/5 Environments	GBLUP	0.913±0.050	0.552±0.063	0.538±0.223	0.409±0.149
	E-GP (BD)	**0.982**±0.006	**0.574**±0.051	**0.593**±0.095	**0.410**±0.135
	E-GP (RN)	0.831±0.069	0.572±0.060	0.468±0.117	0.394±0.134
	W-GP (BD)	**0.986**±.004	**0.575**±0.051	**0.592**±0.096	**0.411**±0.139
	W-GP (RN)	0.906±0.046	0.539±0.067	0.476±0.119	0.404±0.116

Tabela 3. Predictive ability of the genomic prediction models (GP) for two tropical maize data sets (Multi-Regional and N-level) produced using the effective number of phenotypic records (N_{GE}, genotypes-environments observations) and for the scenarios Field Trials (predicting N_{GE}) and Virtual Network (predicting $n - N_{GE}$, where n is the number of genotypes by environments available in the full data set). The reference "full" and "5%" in parentheses represents the predictive ability produced with all genotypes and using only the top 5%, respectively

Scenario	Models				
	GBLUP	**W-GP (BD)**	**W-GP (RN)**	**E-GP (BD)**	**E-GP (RN)**
Multi-Regional set					
Field Trials					
N_{GE} = 210 (full)	0.698	0.962	0.892	0.964	0.893
N_{GE} = 210 (5%)	0.991	0.995	0.992	0.997	0.998
N_{GE} = 49 (full)	0.738	0.941	0.840	0.942	0.840
N_{GE} = 49 (5%)	0.991	0.991	0.991	1.000	1.000
Virtual Network					
N_{GE} = 210 (full)	0.175	0.794	0.787	0.793	0.787
N_{GE} = 210 (5%)	0.098	0.736	0.750	0.713	0.715
N_{GE} = 49 (full)	0.190	0.810	0.788	0.810	0.789
N_{GE} = 49 (5%)	0.241	0.759	0.755	0.758	0.706
N-level set					
Field Trials					
N_{GE} = 536 (full)	0.982	0.984	0.775	0.991	0.775
N_{GE} = 536 (5%)	0.964	0.861	0.861	0.998	0.999
N_{GE} = 67 (full)	0.983	0.981	0.718	0.989	0.719
N_{GE} = 67 (5%)	0.967	0.833	0.802	0.998	1.000
Virtual Network					
N_{GE} = 536 (full)	0.196	0.608	0.612	0.601	0.612
N_{GE} = 536 (5%)	0.152	0.554	0.545	0.406	0.484
N_{GE} = 67 (full)	0.102	0.574	0.572	0.578	0.573
N_{GE} = 67 (5%)	0.070	0.545	0.539	0.379	0.510

FIGURES

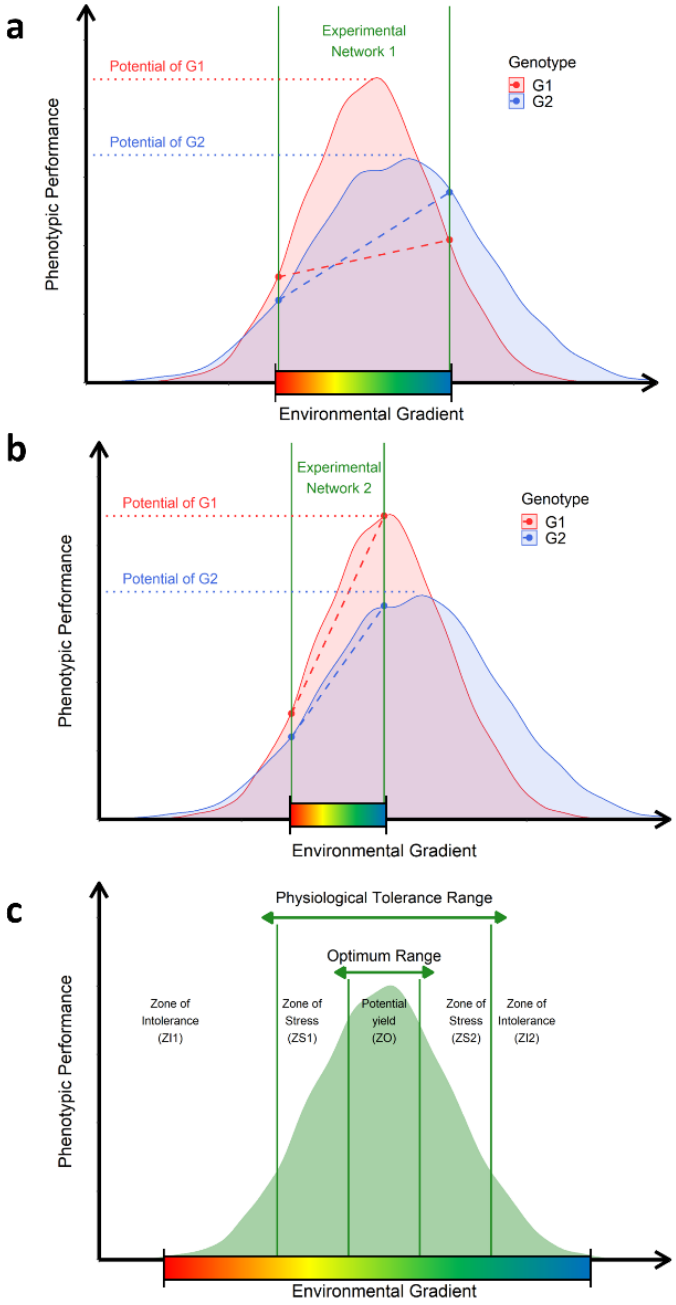

Figura 1. Ecophysiological insights to translate raw-environmental data into enviromic sources. **A**. Representation of an experimental network involving an unknown number of environments from a theoretical TPE and two genotypes (G1 and G2). The range of the environmental gradient is delimited by the space between the two vertical green lines. Each genotype has a nonlinear function describing the genetic limits of their phenotypic plasticity (curves) and genetic potential (horizontal dotted lines) of a given trait. Diagonal dotted lines denote the observed reaction-norm experienced by those genotypes; **B**. representation of a second experimental network involving the same genotypes, but different environments were sampled from the theoretical TPE. **C**. adaptation of Shelford's Law of Tolerance, describing the cardinal (or biological) genetic limits (vertical green lines) to determine the amount of the factor that results in different adaptation zones. Across these zones, crop performance is described by zones of stress caused by deficit or excess (physiological tolerance range) and zones of optimal growing conditions that allow the plants to express the genetic potential for a given trait (optimum range). The core of possible environmental variations contemplated as putative phenotypic plasticity for a given genotype, germplasm, or crop species

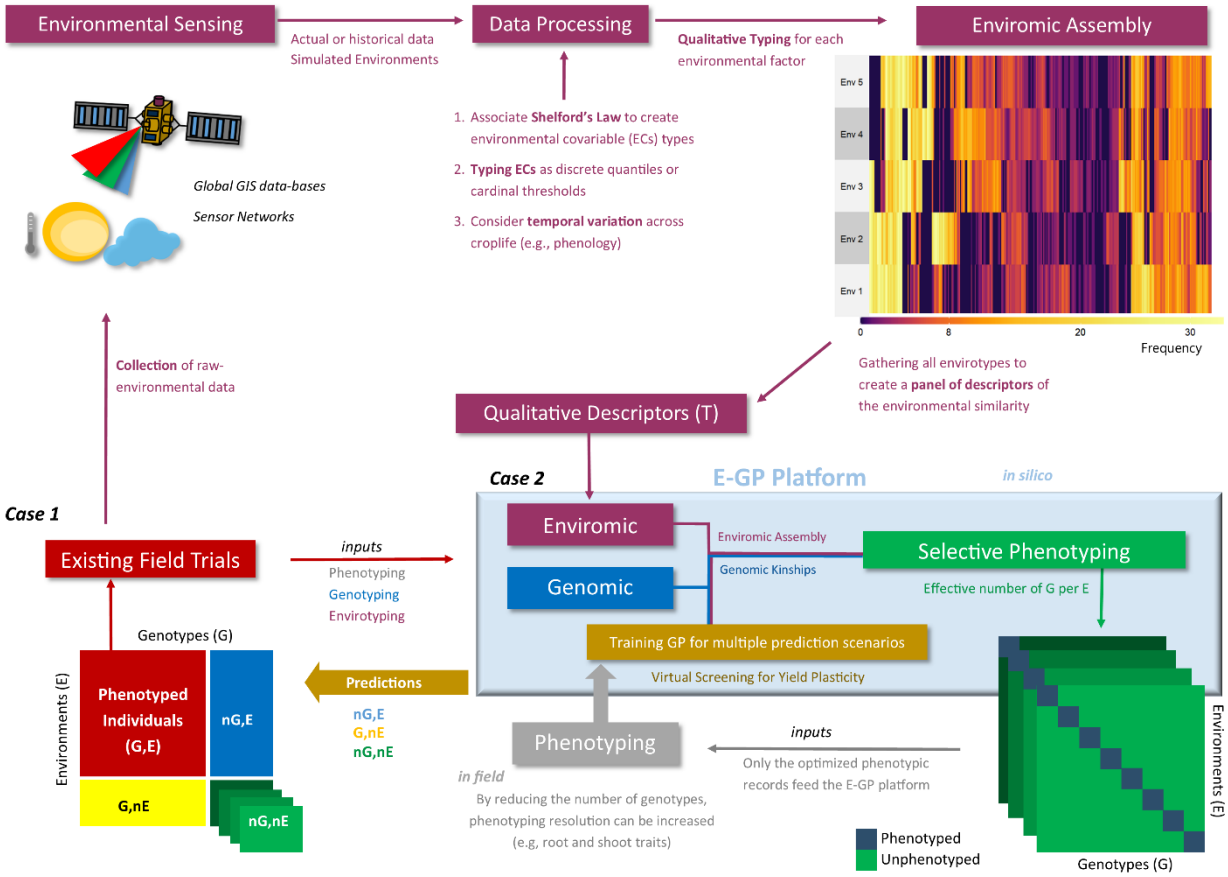

Figura 2. Workflow of the E-GP considering the two study Cases of this study

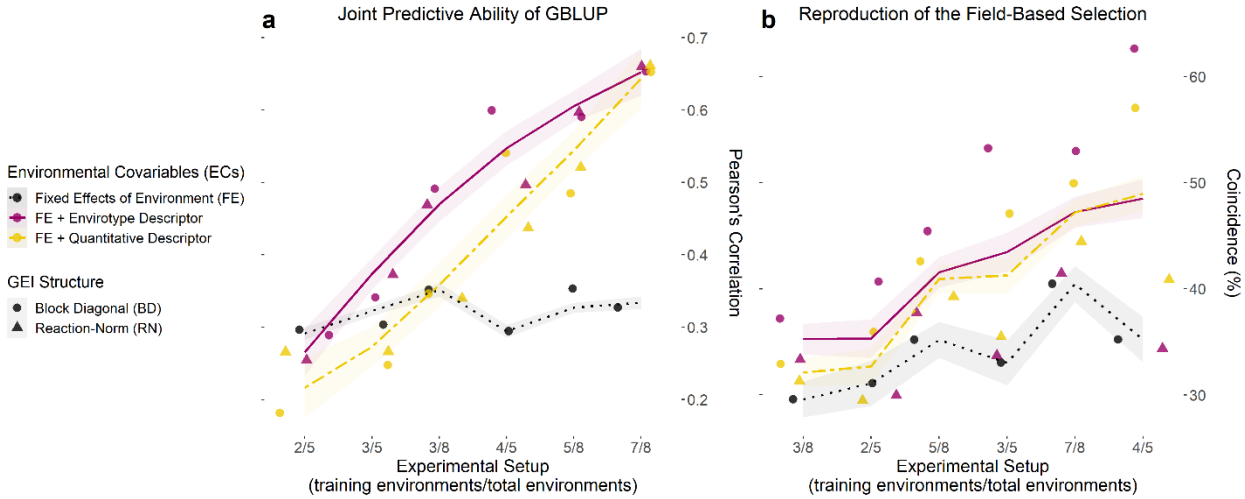

Figura 3. Joint accuracy trends of GP models for each training setup of existing experimental networks. a. Predictive ability computed with the correlation (r) between observed (y) and predicted ($\hat{y}$) values for the grain yield of each genotype in each environment, over three experimental setups (number of environments used/total of environments) for both maize sets (N-level and Multi-local), using 70% of the genotypes as a training set and the remaining 30% as a testing set. **b**. Coincidence index (CS) between the field-based and prediction-based selection of the best 5% genotypes in each environment for the same experimental setups and data sets. Dots and triangles represent the point estimates of predictive ability and CS for models involving a block diagonal genomic matrix for G×E effects (dotted) and an enviromic × genomic reaction-norm G×E effect (triangle). Trend lines were plotted from the partial values of each sample (from 1 to 50) and three prediction scenarios (nG, E; G, nE and nG, nE) by using the gam() integrated with smoothness estimation in R. Black dotted lines represent the benchmark GBLUP method, considering the effect of the environment as a fixed intercept. Yellow two-dash lines represent the GBLUP involving the main effect from quantitative descriptors (**W** matrix). Finally, solid dark pink lines represent the GBLUP involving the main effect of envirotype descriptors (**T** matrix). Thus, the latter represents the E-GP based approach for *Case 1* (predictions under existing experimental networks).

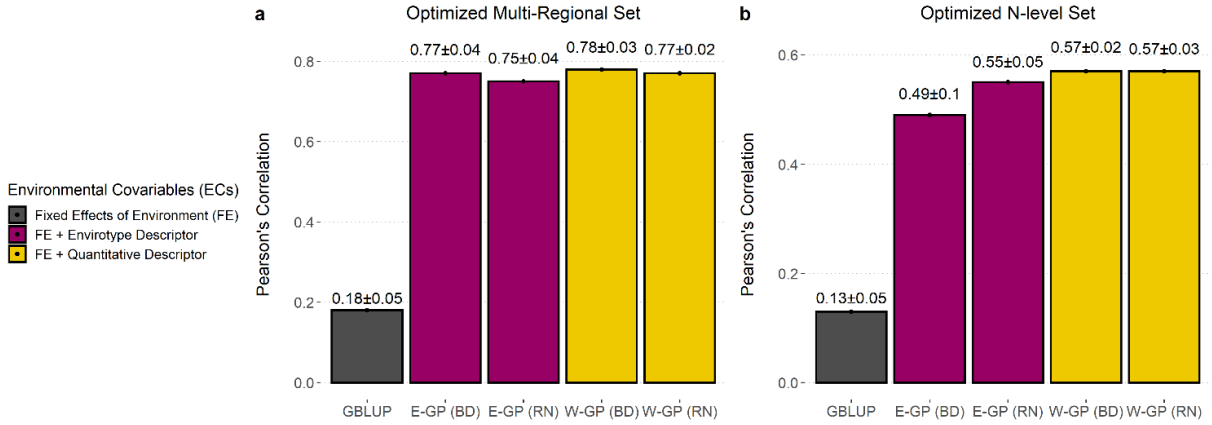

Figura 4. Accuracy of GP models trained with super-optimized experimental networks. Predictive ability (*r*) plus standard deviation measured by the correlation between observed and predicted values for each model in the optimized Multi-Regional Set (**a**); and for the N level Set (**b**). Barplots were colored according to the type of environmental covariable (ECs) used: none (black), envirotype descriptor (**T** matrix, wine), and quantitative descriptor (**W** matrix, yellow)

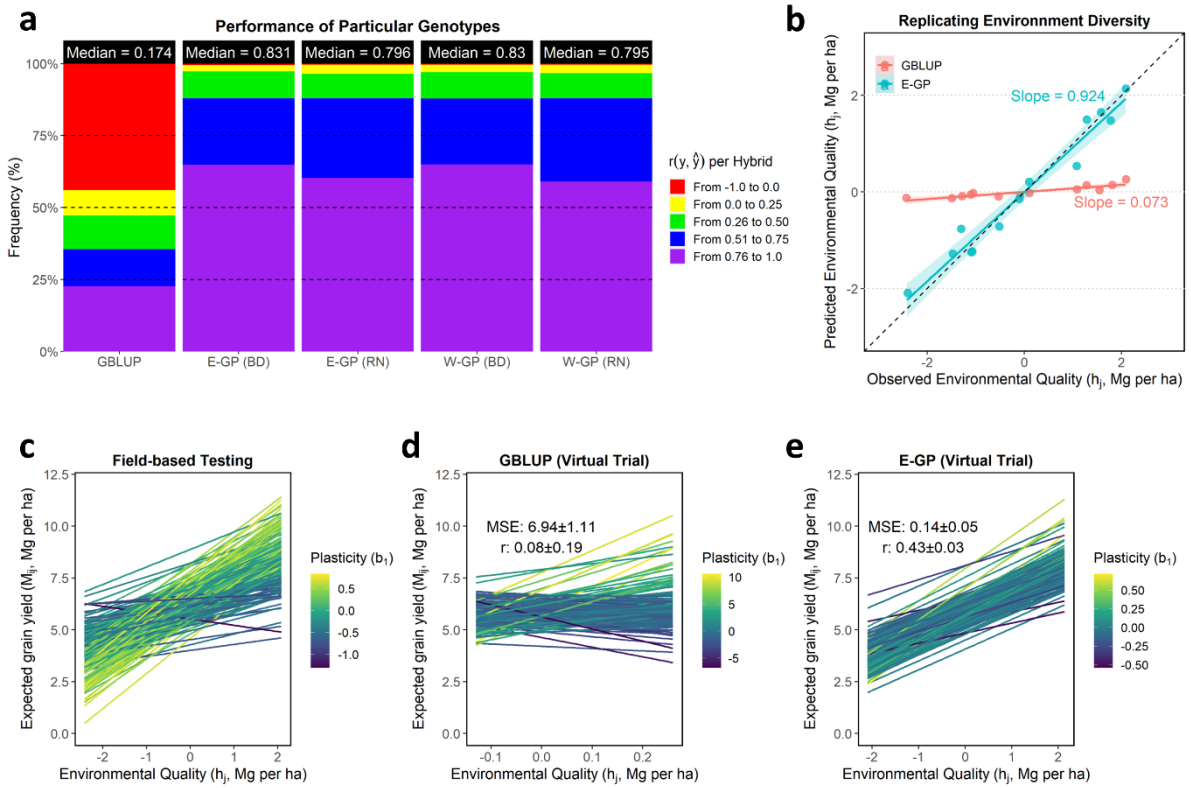

Figura 5. Accuracy of GP models in reproducing the genotype-specific plasticity. **a**. The panel of predictive ability (*r*) explaining the plasticity of genotypes across environments. This statistic was estimated for each individual (hybrid) by correlating observed and predicted values across environments. Individuals with values below 0 were considered unpredictable and marked in red. **b**. ability of the prediction-based tools to reproduce an existing experimental network's environmental quality (h_j). In the X-axis, we find the h_j computed using the phenotypic records of a current experimental network. In the Y-axis, the h_j values are presented considering a virtual experimental network built up using GBLUP and E-GP (with BD) predictions. **c**-**e**. Yield plasticity panels denoting each genotype's G×E effects across the h_j values for observed field-testing screening (**c**) concerning prediction-based (**d**-**e**). Only the 5% best genotypes in each environment were used to create this plot. Each line was colored with the genotype-specific plasticity coefficient (b_1). For the N-level set, the full-optimized set (536 hybrids over eight environments) was used.

SUPPLEMENTARY TABLES

Tabela 1. Supplementary. Estimated variance components (±standard deviation) for Multi-Regional Set.

Effect	Model				
	GBLUP	**W-GP (BD)**	**E-GP (BD)**	**W-GP (RN)**	**E-GP (RN)**
T	-	-	0.157±0.019	-	0.055±0.004
W	-	0.090±0.007	-	0.078±0.007	-
A	0.464±0.012	0.465±0.012	0.464±0.011	0.462±0.011	0.499±0.014
D	0.218±0.004	0.217±0.004	0.216±0.004	0.211±0.004	0.232±0.005
AE	0.197±0.003	0.199±0.003	0.195±0.003	-	-
DE	0.089±0.001	0.088±0.001	0.089±0.001	-	-
AT	-	-	-	-	0.007±0.009
AW	-	-	-	0.004±0.002	-
DT	-	-	-	-	0.002±0.001
DW	-	-	-	0.002±0.003	-
Residual	0.234±0.001	0.235±0.001	0.235±0.001	0.239±0.001	0.237±0.001
Explained variance (%)	81%	82%	83%	76%	77%

Tabela 2. Supplementary. Estimated variance components (±standard deviation) for N-level set.

Effect	Model				
	GBLUP	**W-GP (BD)**	**E-GP (BD)**	**W-GP (RN)**	**E-GP (RN)**
T	-	-	0.195±0.016	-	0.324±0.029
W	-	0.319±0.029	-	0.643±0.06	-
A	1.371±0.031	1.392±0.03	1.350±0.03	1.330±0.029	1.396±0.036
D	0.574±0.005	0.575±0.005	0.575±0.005	0.550±0.005	0.426±0.005
AE	0.363±0.003	0.366±0.004	0.365±0.003	-	-
DE	0.337±0.003	0.338±0.003	0.336±0.003	-	-
AT	-	-	-	-	0.034±0
AW	-	-	-	0.016±0.001	-
DT	-	-	-	-	0.012±0
DW	-	-	-	0.006±0.002	-
Residual	1.154±0.004	1.153±0.004	1.155±0.004	1.45±0.003	1.37±0.003
Explained variance (%)	70%	72%	71%	64%	62%

SUPPLEMENTARY FIGURES

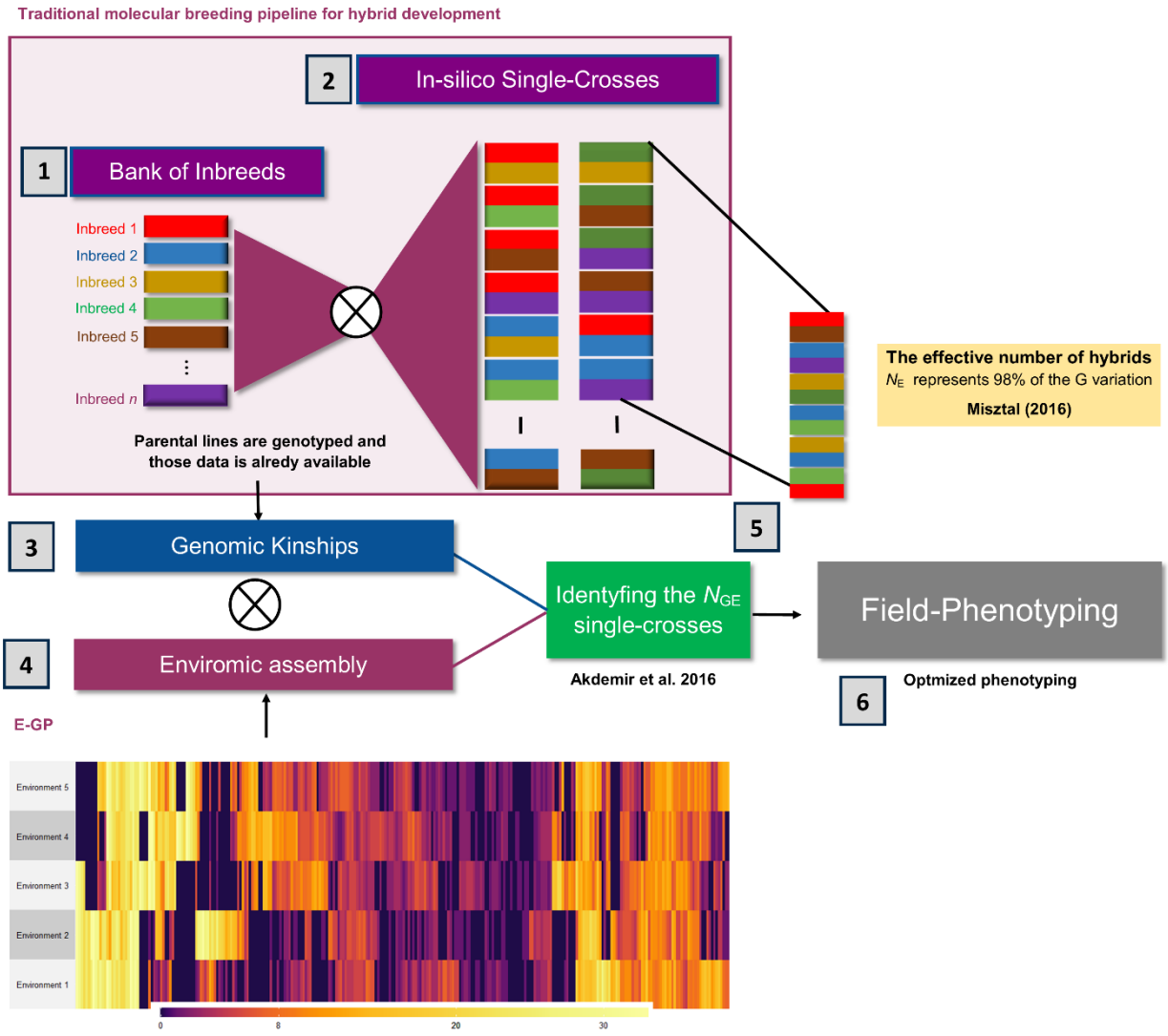

Figura 1. Supplementary. Use of selective phenotyping with genomic kinship and enviromic assembly for boosting the hybrid breeding pipelines. In purple, it is presented the current molecular breeding approaches of hybrid development. (1-2) (1) From a bank of elite inbreeds already genotyped, it is possible to create a large number of in-silico single-crosses (2) using the Kronecker product between each SNP marker from desirable parentals. Then, a single-value decomposition (SVD) of the genomic kinship (G) reveals the effective number of genotypes (N_E) that represents at least 98% of the variation in G (3). However, under the E-GP platform, the use of G plus an enviromic assembly (T) for a target population of environments (TPE) can be used to build up in silico possible growing conditions that crops may experience (4). Then, via the Kronecker product between enviromic x genomic, it is possible to build a matrix accounting for genotypic observations per environment. Then, we apply SVD on that to select the effective number of genotypes per environments (N_{GE}) that represents at least 98% of the variation of the realized experimental network (5). Later, using the SPTGA package, that provides a genetic algorithm, we can define the most relevant combinations of genotype x environment (6). Finally, only these individuals are phenotyped in specific locations. Then, it will be used as a training population for genomic-based prediction or other research purposes, such as training crop growth models or running a factorial regression analysis. This approach allows an optimized training of those models, which may increase efficiency in predicting phenotypic landscapes across novel growing conditions.

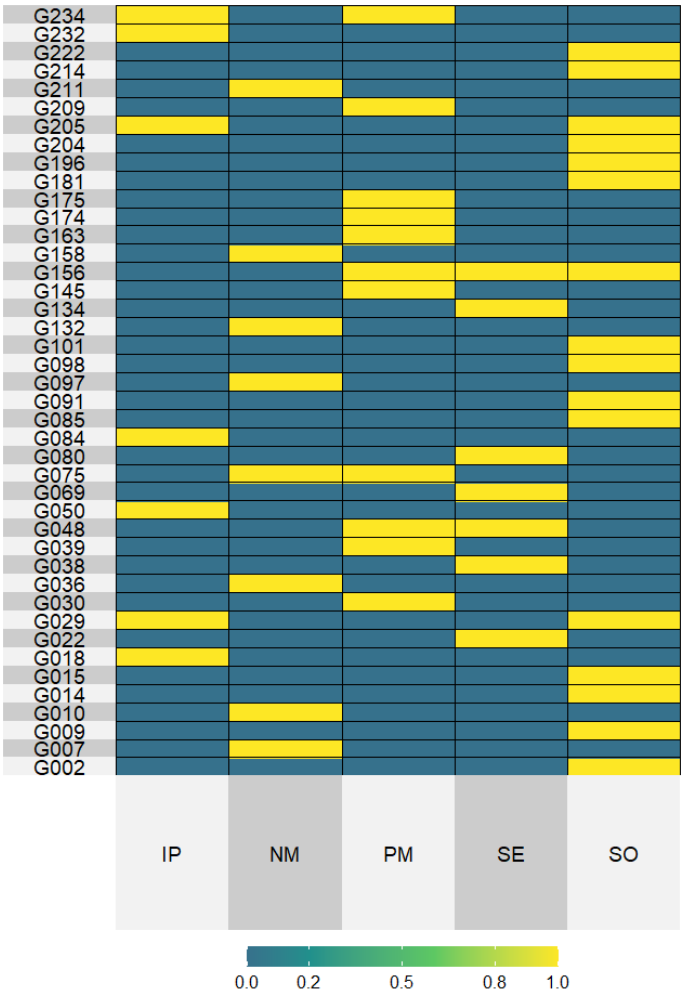

Figura 2. Supplementary. Summary of the selective phenotyping approach drawn by the effective number of observations (N_{GE}) phenotyped in the field for the Multi-Regional Set (247 tropical maize hybrids over 5 locations).The core of 42 maize hybrids (rows) per 5 environments (columns). In yellow is the hybrid-environment combinations phenotyped in the field-based trials. It resulted in N_{GE} = 49, because some genotypes occur in more than one environment. At each environment, the number of genotypes was: IP (7), NM(8), PM (11), SE (7), and SO (16). The remaining 205 hybrids plus the blue cells were considered a testing set (virtual experimental network)

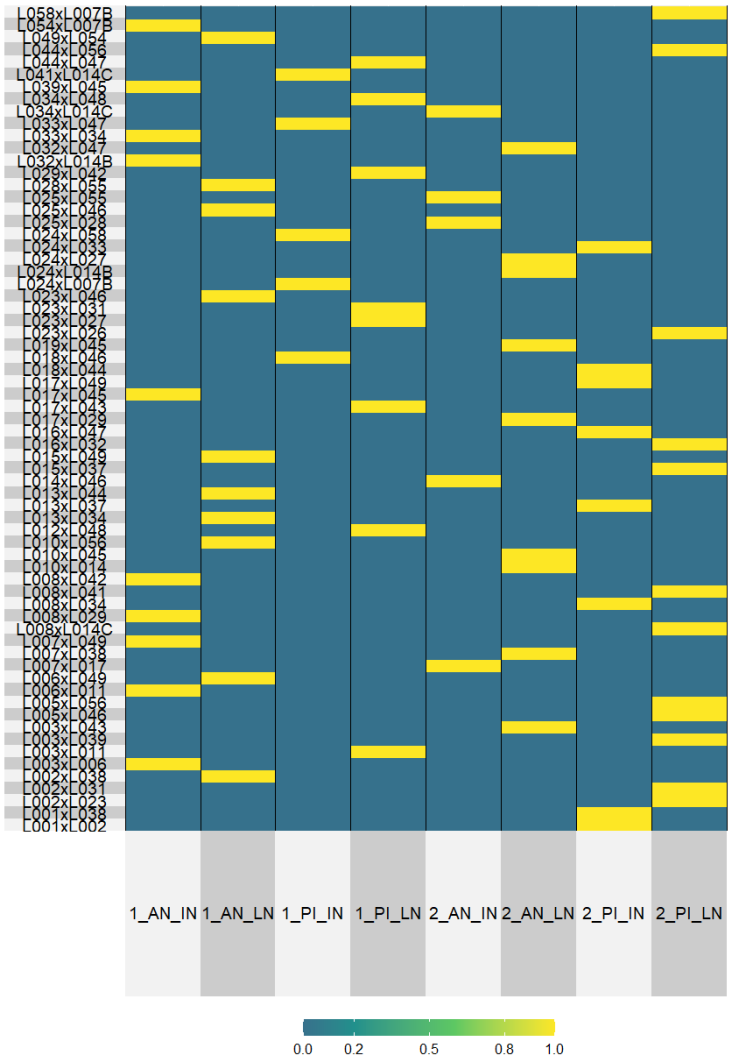

Figura 3. Supplementary. Summary of the selective phenotyping approach drawn by the effective number of observations (N_{GE}) phenotyped in the field for the N-level Set (570 tropical maize hybrids over eight environments). The core of 67 maize hybrids (rows) per 8 environments (columns). In yellow is the hybrid-environment combinations phenotyped in the field-based trials. It resulted in N_{GE} = 67. At each environment, the number of genotypes were: 1_AN_IN (10), 1_AN_LN (10), 1_PI_IN (5), 1_PI_LN (8), 2_AN_IN (5), 2_AN_LN (9), 2_PI_IN (8), 2_PI_LN (12). The remaining 503 hybrids plus the blue cells were considered as testing set (virtual experimental network)