



Enviromics in breeding: applications and perspectives on envirotypic-assisted selection

Rafael T. Resende¹ · Hans-Peter Piepho² · Guilherme J. M. Rosa³ · Orzenil B. Silva-Junior⁴ · Fabyano F. e Silva⁵ · Marcos Deon V. de Resende^{6,7} · Dario Grattapaglia^{4,8}

Received: 25 June 2019 / Accepted: 10 September 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

Key message We propose the application of enviromics to breeding practice, by which the similarity among sites assessed on an “omics” scale of environmental attributes drives the prediction of unobserved genotype performances.

Abstract Genotype by environment interaction (GEI) studies in plant breeding have focused mainly on estimating genetic parameters over a limited number of experimental trials. However, recent geographic information system (GIS) techniques have opened new frontiers for better understanding and dealing with GEI. These advances allow increasing selection accuracy across all sites of interest, including those where experimental trials have not yet been deployed. Here, we introduce the term enviromics, within an envirotypic-assisted breeding framework. In summary, likewise genotypes at DNA markers, any particular site is characterized by a set of “enviotypes” at multiple “enviomic” markers corresponding to environmental variables that may interact with the genetic background, thus providing informative breeding re-rankings for optimized decisions over different environments. Based on simulated data, we illustrate an index-based enviromics method (the “GIS–GEI”) which, due to its higher granular resolution than standard methods, allows for: (1) accurate matching of sites to their most appropriate genotypes; (2) better definition of breeding areas that have high genetic correlation to ensure selection gains across environments; and (3) efficient determination of the best sites to carry out experiments for further analyses. Environmental scenarios can also be optimized for productivity improvement and genetic resources management, especially in the current outlook of dynamic climate change. Envirotyping provides a new class of markers for genetic studies, which are fairly inexpensive, increasingly available and transferable across species. We envision a promising future for the integration of enviromics approaches into plant breeding when coupled with next-generation genotyping/phenotyping and powerful statistical modeling of genetic diversity.

Introduction

One of the greatest challenges of modern agriculture is dealing with an accelerated growth of the human population worldwide, together with limited prospects of significantly

expanding farmed land. Tailoring highly adapted genetic material to the available environments becomes a key element to increase agricultural yields without the conversion of additional land and losses due to adverse environmental impact (Garnett et al. 2013). The differential response of genotypes across variable environments, known as genotype

Communicated by Martin Boer.

Rafael T. Resende
rafael.tassinari@ufg.br

Dario Grattapaglia
dario.grattapaglia@embrapa.br

¹ School of Agronomy, University of Goiás (UFG), Goiânia, GO 74690-900, Brazil

² Biostatistics Unit, University of Hohenheim, 70593 Stuttgart, Germany

³ Department of Animal and Dairy Sciences, University of Wisconsin-Madison, 53706 Madison, USA

⁴ EMBRAPA Genetic Resources and Biotechnology – EPqB, Brasília, DF 70770-910, Brazil

⁵ Department of Animal Science, University of Viçosa (UFV), Viçosa, MG 36570-900, Brazil

⁶ Department of Statistics, University of Viçosa (UFV), Viçosa, MG 36570-900, Brazil

⁷ EMBRAPA Coffee Research, Brasília, DF 70770-901, Brazil

⁸ Genomic Sciences and Biotechnology Program, Catholic University of Brasília, Brasília, DF 70790-160, Brazil

by environment interaction (GEI), represents one of the major challenges faced by essentially all animal and plant breeding programs.

Traditional GEI studies are based on the evaluation of trials designed to estimate parameters that describe the interaction of tested genetic materials with a restricted set of environments or environmental conditions (Elias et al. 2016). However, to maximize genetic gains over a range of local conditions, it is desirable to collect data from all measurable environmental factors affecting the performance of genotypes in any particular site (Des Marais et al. 2013). Patterns of relationships between environmental variables and the expression of genotypes can be investigated through modern concepts of phenomics and genomics, which rely on high-throughput, large-scale data collection and evaluation techniques (Houle et al. 2010; van Eeuwijk et al. 2018). Such changes in the physical environment can then be incorporated into genetic analyses to better understand the interrelationships between environmental factors and yield or performance variables for detection of the best combinations of genetic and environmental conditions.

In contemporary genetic analyses, the pattern of expression of a trait for a particular genotype across environments, or for various genotypes along descriptors of the environment, can be analyzed using reaction norm models (Ribeiro et al. 2015), which rely on the definition of an environmental variable that satisfactorily describes diverse environmental conditions. In breeding programs, an environmental variable for a reaction norm model of selection is usually calculated as the mean phenotypic performance of a trait in a restricted range of environments (Finlay and Wilkinson 1963; Eberhart and Russell 1966), and genetic covariance structures are proposed to re-rank genotypes by calculating breeding values taking into account GEI (Calus et al. 2004). The effect of describing an environmental variable when building genetic covariance structures for reaction norm models of selection has been investigated, for instance, in animal breeding (for a review, see Rauw and Gomez-Raya, 2015).

The possibility of using experimental data from a particular environment to anchor the prediction of performance and subsequent recommendation of potentially successful genotypes in other untested sites has been a topic of great interest in plant and animal breeding. This task has been explored by a number of authors by addressing environmental similarity based on multiple environmental attributes. To be able to predict the performance of individuals in an untested site, environmental covariates need to be used (Piepho et al. 1998; Malosetti et al. 2016). The combination of environmental covariates with geographic information systems (GIS) has been proposed also by Annicchiarico et al. (2006), and the use of extensive environmental information in reaction norm models has been recommended by Jarquín et al. (2014). In Jarquín et al. (2014), 68 environmental covariates

were used to evaluate grain yield of commercial wheat lines, allowing the prediction of unobserved lines in untested environments based on high-dimensional environmental and genomic data. Several other authors have implemented and expanded this idea, such as Pérez-Rodríguez et al. (2015), who used 76 environmental covariables and pedigree information in nine cotton trials. Additionally, to improve predictive accuracy through the exploration of environmental co-variation, the concept of on-farm trials has been used, which allows the generation of large volumes of data to tap into hundreds or even thousands of environments (Schmidt et al. 2018). The inclusion of environmental information in the genetic models has resulted in significant gains in prediction accuracy of genotype performance, often improving the predictive ability by around 20% and, in some cases, by up to 34% (Jarquín et al. 2014; Acosta-Pech et al. 2017). Performance prediction using environmental data also opens possibilities to expand recommendations of genetic materials across countries that lack experiments for particular crops of interest, thus saving resources, avoiding sanitary barriers and reducing phenotyping costs (Pérez-Rodríguez et al. 2017; Sukumaran et al. 2017).

Various modeling techniques have shown their usefulness in the analysis of data from multiple environments, including mixed-effects models, linear-bilinear, crop growth and Bayesian approaches (van Eeuwijk et al. 2016). Despite the impact that such approaches have already had in understanding and exploiting GEI for prediction of yet-to-be-observed phenotypes, there is still room for expanding and improving their use in applications not yet explored. Understanding the sources of environmental variation has increasingly become a key element for the assessment and recommendation of genotypes under probabilistic scenarios of global climate change and rapid landscape modification by human action (Raza et al. 2019). Such models can be used along the various stages of a breeding cycle (Annicchiarico and Iannucci 2008) to identify loci related to phenotypic trait expression (Ferrero-Serrano and Assmann 2019) and to determine sets of environmental factors underlying phenotypic plasticity (Piepho 2000; Nicotra et al. 2010). For example, traditional genomic selection (GS) and genome-wide association studies (GWAS) can also be developed under environmental gradient models (Acosta-Pech et al. 2017; Velazco et al. 2017; Mota et al. 2020).

Integrating enviromics with breeding

To more thoughtfully explore the effects of the various environmental factors on selective breeding, we borrow the term “enviromics” from human medicine, i.e., the study of the environmental conditions that affect human health in the context of precision medicine (Gad 2008; Teixeira et al. 2011; Riggs et al. 2018). Here, however, we extend

this concept and apply it to the environment-dependent part of reaction norm models for genetic selection with the goal of exploiting patterns of GEI in local environments, mainly in plant, but readily extendable to animal breeding. This work is inspired by exciting developments in the field of population genomics and epidemiology in which a new type of analysis of phenotypic and genotypic data and environmental variables, termed “phenome-wide association studies” has been used to investigate the effects of environmental variables on clinical outcomes using data from large and diverse populations such as the PAGE consortium (Matise et al. 2011). Along the same lines, a recent study involving an extensive analysis of the local environments described by 204 geoclimatic variables of *Arabidopsis* accessions and 131 phenotypes revealed candidate adaptive genetic variation, such as cold tolerance associated with high-dimensional environmental variables (Ferrero-Serrano and Assmann 2019).

In our conceptualization of enviromics in breeding, a particular land area is a geoprocessing environment corresponding to a grid of pixels, just like those of a digital image, and for any single environmental variable a value can be assigned to every pixel. The distribution of values of a particular environmental variable in this collection of pixels constitutes the range of envirotypes (Beckers et al. 2009), that have arisen in a specific land area (Fig. 1). In this context, for each set of pixel coordinates it is possible to assess an effect on the evaluated trait. Thus, connections between pixels allow making feasible predictions in the absence of genomic and/or genealogical relationship information between evaluated genotypes (see Fig. 2, parts “a” and “b”). In crop plants that allow obtaining large numbers of identical copies of the same genotype (e.g., inbred lines, hybrids or clonal varieties), any individual genotype can be tested in multiple sites and over time, so that the notion of a reaction norm of the genotypes distributed across a wide environmental range is straightforward. In animal breeding, or in outbred plant species, the reaction norm of an individual genotype needs to be inferred based on the performance of genetically related individuals across a range of environments and moments (Rauw and Gomez-Raya 2015; Fernandes et al. 2019). In any case, if phenotypic data from an individual genotype in existing field trials is available in some pixels, one can relate the distribution of a particular phenotypic value with the envirotypes using statistical modeling (Hyman et al. 2013; van Eeuwijk et al. 2018).

Beckers et al. (2009) and Xu (2016) first proposed the term envirotypes (environment + types) as all potential environmental factors that affect plant growth and yield, together with the definition of envirotyping as the process for determining and measuring all these environmental factors. In our conceptual framework, the environmental variables that

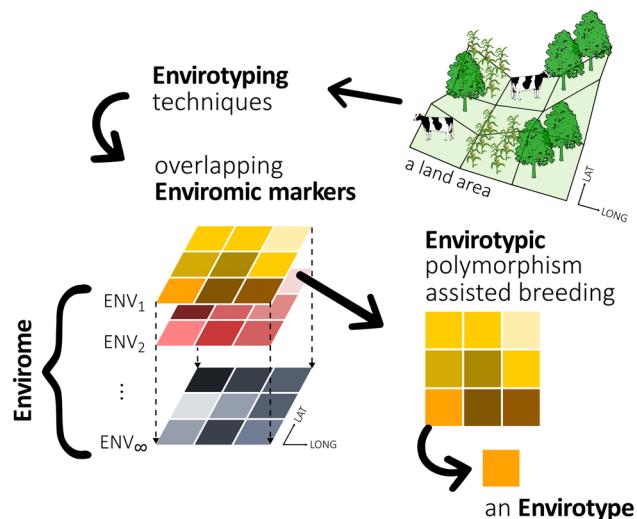


Fig. 1 Enviromics terms anchored into a geoprocessing environment (a land area with nine pixels). *Enviromic markers* (e.g., time-trend climate, landscape or management treatments) are massively achieved by means of modern *envirotyping* techniques; thus, an exhaustive set of such markers composes the *envirome*. A possible value/feature that an enviromic marker can assume is called an *enviotype*, and the combined envirotypes compose the *envirotypic polymorphic variation*

can be obtained by envirotyping can be termed “enviromic markers” (Fig. 1). The potential features that an enviromic marker presents within each site/pixel are the envirotypes. In turn, the combination of all the envirotypes at these markers corresponds to the envirome and the marker polymorphism is the envirotypic variation. As proposed by Xu (2016), envirotyping is therefore a third “typing” technology, alongside genotyping and phenotyping, and enviromics would then be a corresponding third “omics” technology, alongside with genomics and phenomics technologies. Reaction norm models are then built for the whole evaluated area and used to predict the performance of any tested genotype for the trait under evaluation in any pixel (tested and non-tested) in the geoprocessing environment.

Integrating breeding and environmental data relies on the increasing worldwide availability of geoprocessing technologies, such as GIS in the scope of precision agriculture (PA) (Lindblom et al. 2017). The collection and processing of spatiotemporal data on weather, water, soil and yield variables is rapidly increasing due to the societal need for food security and technological advances. The *Arabidopsis CLIMtools* repository (<https://github.com/CLIMtools/>) is, to our knowledge, the first repository of data and tools specifically tailored for the exploitation of environmental variation associated with any gene or variant of interest in plants. In crop plants, the international DivSeek initiative is a remarkable example (Nature Genetics Editorial 2015), and we can envisage in the near future a rapid expansion of similar efforts toward the organization and availability of

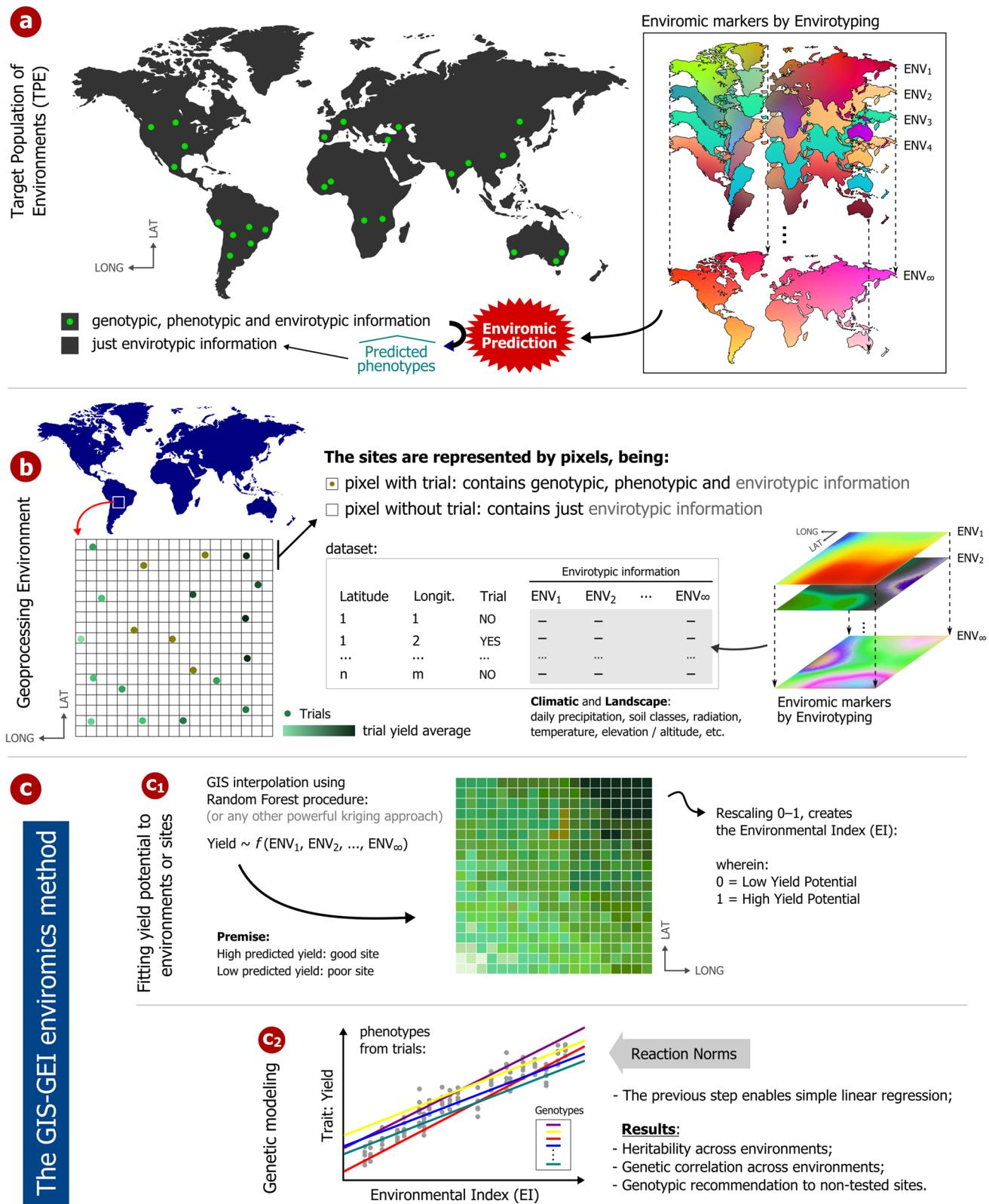


Fig. 2 Schematic conceptualization. **a** Target population of environments (TPE) and global enviromic markers, **b** illustrative example of an enviromics data set, and **c** hypothetical GIS-GEI method application containing just 20 trials and 5 genotypes, applying the “part b” example

very large collections of genetics and genomics data linked to physical resources, germplasm curators, breeders and researchers alike.

The main challenge in the implementation of enviromics approaches is that only a few of the pixels representing a land area contain breeding trials having phenotypic and genotypic information (Fig. 2b). However, all pixels have environmental information since affordable meteorological stations have been increasingly installed in multiple and diverse locations, such that environmental data can be interpolated by kriging across any desired area (Oliver and Webster 2015). Because geospatial information is now easily accessible, data-driven approaches supported by GIS can be exploited for breeding practices. Challenges of data analysis in agricultural research are changing as the features of available data improve (Wolfert et al. 2017), especially through the use of modern geotechnologies (Xu 2016). Remote-sensing technologies can also provide data that can be used with specific spectral bands for the aimed purposes (Kasampalis et al. 2018). Precipitation, temperature, terrain altitude (in the form of digital elevation models), some cultural treatment, solar radiation, soil properties and water deficit information are some of the variables that relate well to most phenotypic traits evaluated in agricultural production (Xu 2016; Chang 2017). Indeed, the combination of GIS and breeding has been gradually showing its potential (Hyman et al. 2013; Haghishattalab et al. 2017; Marcatti et al. 2017; Costa-Neto et al. 2020). For example, the spatial positioning of trials or even experimental plots linked to the measurement of multiple environmental information by envirotyping, either at macro- or micro-scales, suggests that the prospect of enviromic techniques is especially promising in association with breeding cycles and genotypic recommendations.

In this study, we explore the concept of enviromics in the context of phenotypic breeding by presenting and applying a method based on geographic information systems coupled with genetics (GIS–GEI) to a case study based on advanced environmental interpolation techniques. We show that the main advantages brought by this method are an improved matching of genotypes to their most appropriate sites (either tested or non-tested), a detailed zoning of breeding areas with high genetic correlation among sites within zones, and the identification of the best sites to carry out experiments for further analysis based on regions that maximize trait heritability.

Methods

Environmental and genetic data simulation

Two sequential algorithms were adopted in the simulation process. The first algorithm generated and characterized

the land area composition within a geoprocessing environment and formulated the envirotypic data, independently of genotypic and phenotypic information. Figure 2b shows a schematic example containing some trials, and the arrangement of the envirotypic data of each enviromic marker. The second algorithm entailed generating trials and genotypes, while the plant phenotypes were simulated to provide four sources of variation: (i) genotypic values normally distributed with known mean and variance; (ii) an infinitesimal genotypic relationship with each envirotypic information of the previous simulation stage; (iii) a particular trial effect; and (iv) an overall random error. For a better understanding of the simulation process and algorithms, please refer to the commented simulation code “code_data-simulation.R” provided at [https://figshare.com/articles/Enviroomics_data/8264132].

Envirotypic algorithm

A set of 50 trials was randomly allocated to a square area covering $100 \times 100 = 10,000$ pixels. One hundred overlapping rasters containing envirotypic information of a single environmental variable at their corresponding sites/pixels were simulated. Overlapping sites/pixels across all the rasters constitute an enviromic marker for the land area. We then simulated data for each enviromic marker corresponding to the values of some exogenous environmental variable that could potentially affect yield in the area, by adopting a simple matrix multiplication between two smoothed vectors of dimensions 100×1 and 1×100 , thus composing a 100×100 squared matrix of a gradient of values. In our simulation, the values attributed to the enviromic markers were merely random, but in practice they would correspond, for example, to the historical annual temperature, historical annual precipitation, any cultural treatments, terrain altitude, nutritional or physical soil characteristics, radiation and vegetation indices, among others, as described by Xu (2016).

The formulation of the envirotypic data based on purely random values of some environment variables should be seen as a simplification for the purpose of our simulations. In fact, this formulation could be particularly challenging for the genetic modeling in enviromics, which may have important implications for the assessment of the model performance on relating, for instance, the influence of time-dependent trends of climate or cropping management treatments to the phenotypic trait of interest.

Genotypic and phenotypic algorithms

Values for a particular phenotype collected from 100 plant genotypes (coded as G001, G002, up to G100) in multiple

trials were simulated, mimicking a response trait, e.g., agricultural yield. It could correspond, for instance, to the yield of an annual crop, timber volume of a planted forest tree, forage biomass or fruit crop yield. The enviromes of the overlapping enviromic markers can influence (either positively or negatively) the phenotypic trait expression of a plant genotype at the particular site/pixel, with hypothetically known latent-effect magnitudes, similar to allele types of a molecular marker at a genomic locus.

More specifically, the raw genetic effect of each genotype was considered as normally distributed with known trait mean and variance. In addition, the primary relationship between enviromic markers and the trait expression was modeled as linear, and embedded in phenotypic values, weighted by the latent importance of each marker. This enviromic weighted effect on the trait was also used to order trials from the least to the most favorable one over the trait expression.

The simulation of GEI effects was performed based on a general first-order autoregressive process to describe the spatially variable genotypic interactions across the ordered trials. This simulation process specified the behavior of the genotype across environments depending on its previous values combined with a non-predictable stochastic term (random noise) to set the next value. This process was performed until all trials characterized suitable information in terms of nonlinear predictable trends (please, see Fig. 3b to check GEI levels over different environments).

To establish sets of phenotypic, environmental and genetic data suitable for enviromic marker analysis, the simulation process was repeated with varying parameter values targeting a relative genetic variance of approximately 10% ($h^2 \approx 0.10$), a typical value for a considerable proportion of quantitative traits in plants. Note that for simplicity, we used the term heritability throughout the study to express the relative genetic variance of a trait. For a more in-depth treatment of the precise definition and estimates of heritability in the context of plant breeding, see Schmidt et al. (2019a). The following model was used for the simulations:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zg} + \mathbf{Wt} + \boldsymbol{\varepsilon}, \quad (1)$$

where \mathbf{y} is the vector of phenotypic means per genotype and trial; $\boldsymbol{\beta}$ represents the vector of fixed effects (overall intercept); \mathbf{g} represents the vector of random effects of genotypes, assumed $\mathbf{g} \sim N(0, \mathbf{K}\sigma_g^2)$; \mathbf{K} is a kinship matrix built from pedigree or genomic information; \mathbf{t} represents the vector of random effects of trials, assumed $\mathbf{t} \sim N(0, \mathbf{I}\sigma_t^2)$; and \mathbf{X} , \mathbf{Z} and \mathbf{W} are known incidence matrices for $\boldsymbol{\beta}$, \mathbf{g} and \mathbf{t} , respectively. The residual vector $\boldsymbol{\varepsilon}$ was assumed as $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma_\varepsilon^2)$. The relative genetic variance, herein termed trait heritability, is given by $h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_t^2 + \sigma_\varepsilon^2)$, where σ_g^2 , σ_t^2 and σ_ε^2 are the variance components related to genotypes, trials and residuals, respectively.

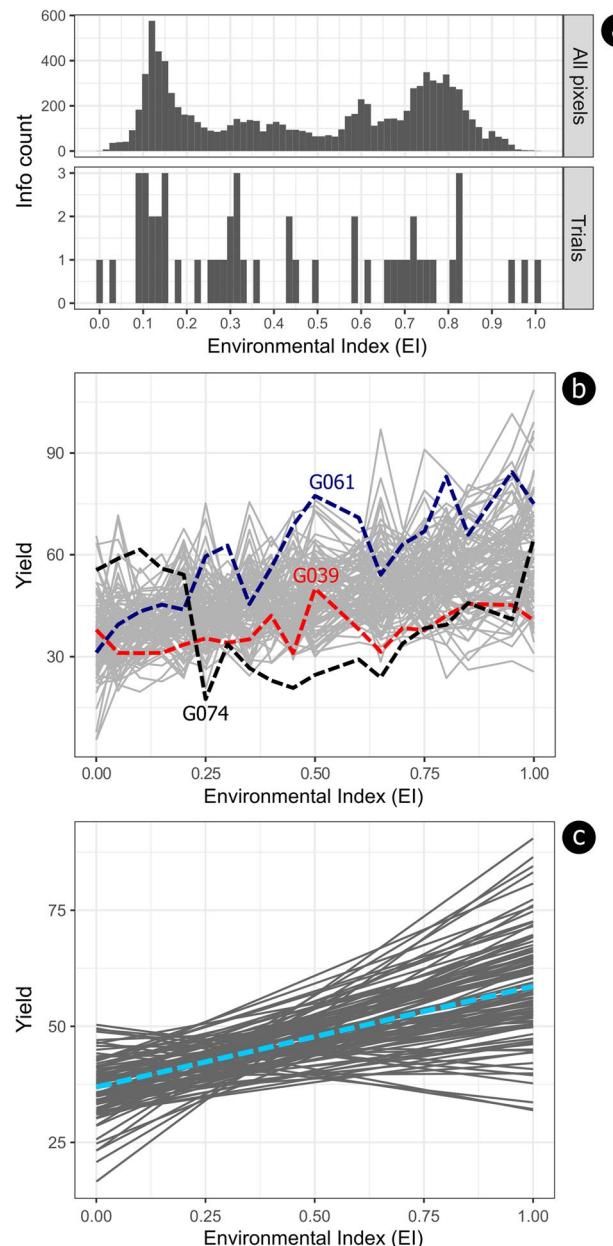


Fig. 3 Features of the environmental index (EI) obtained in the simulations. **a** Comparison between EI of all area pixels and EI of the 50 trials; **b** phenotypic means across environments for the 100 simulated genotypes: genotypes with the highest (G061), lowest (G039) and intermediate (G074) yield are highlighted; **c** reaction norm from the linear random regression model addressing the 100 genotypes. The blue dotted line is the overall fit tendency

Case study: GIS–GEI method for enviromics

To establish a case study for enviromics for the joint analysis of an experimental setting accounting for phenotypic, genotypic and enviromic data, we propose the geospatial (geographic information system) genetics–environment interaction (GIS–GEI) method. GIS–GEI proceeds by introducing a

new approach for evaluating a phenotypic trait by converting the land area into a set of pixels in a geoprocessing framework, making full use of the envirotyping implementation for breeding purposes. The general idea of the method is to generate maps to be utilized within an optimal genotype recommendation framework. Thus, it is possible to identify geographic zones with high genetic correlation between them, i.e., a mega-environment (ME) in which genotype classification changes are minimized (Gauch and Zobel 1997). Additionally, it provides maximization of selection accuracy by using GEI in the model, which often implies a substantial reduction of the residual variance.

The analytical procedures used in GIS–GEI consist of two steps. The first one refers to the development of the enviromic markers relating the trait with the environmental variables, from which an environmental index (EI) is built. The second one refers to the genetic modeling, which consists in fitting reaction norm mixed models assuming the EI as explanatory variable. This second step also considers the estimation of genetic-environmental parameters such as trait heritability based on the enviromic markers. These procedures are summarized in Fig. 2c (parts c₁ and c₂) and detailed below.

Step 1: Environmental index (EI)

For computing the EI, the phenotypic mean within the 50 trials deployed in the area was calculated and subsequently EI values were generated for the entire range of pixels raster by using a random forest (RF) regression in R software (Liaw and Wiener 2002). Random forest is a nonparametric multivariate modeling technique that is well suited to capture nonlinear dependencies among variables and uses a common machine learning algorithm based on an enhanced utilization of *regression trees*. Several studies from different geoscience fields have found a superiority of RF over other machine learning techniques (Koch et al. 2019). In our study, five hundred decision trees (default arguments of the *randomForest* R function) were built to establish the relationships between the 100 enviromic markers and the mean performance of the genotypes within trial for the evaluated yield trait. To this end, the RF models obtained from the phenotypic mean data were used to predict the EI across all pixels in the area, using kriging interpolation (see Fig. 2c₁). We assumed that a higher predicted phenotypic mean of a specific site indicates higher local adaptive fitness or productivity of the genotype and, consequently, better site quality. In this context, if there is interest in establishing any specific EI threshold value (e.g., to discriminate good sites from poor ones), breeders can use information from previous experiments, such as meta-analysis, or exploit future perspectives

according to expected phenotypic gains to ensure profits from a given breeding program. Finally, the EI values were rescaled to a 0–1 interval, with 0 being the worst site, and 1 the best one. In addition to composing the genetic model elucidated below, the EI has an important role in imputing the envirotypes for all pixels of the area, which may then be used for further breeding inferences.

As any other data-based modeling technique, the RF algorithm requires training and validation. The assessment of EI quality for the whole area was carried out using leave-one-out cross-validation (Kohavi 1995), so that the model was trained with data of 49 trials to predict the 50th. The validation procedure was repeated until all environments had a predicted value, and subsequently, the correlation between the observed and predicted values was calculated.

Step 2: Genetic modeling

To represent the association of the evaluated trait with EI, the following linear model (for the single-trait case) can be adopted (Resende et al. 2001):

$$y_{ij} = \beta_0 + \beta_1 EI + (a_{0j} + a_{1j} EI) + e_{ij}$$

where y_{ij} is the measurement of the genotype j associated with the trial i ($i = 1, 2, \dots, 50$; $j = 1, 2, \dots, 100$); β_0 and β_1 are fixed (population-level) intercept and slope for EI, the environmental index; a_{0j} and a_{1j} are random intercept and slope coefficients (individual-level) of genotype j , both jointly forming the random effects of additive genetic value; and e_{ij} is the random error. A bivariate normal distribution needs to be assumed here to make the model invariant with respect to re-scaling of the EI covariate. In matrix notation, this model can be written as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_0\mathbf{a}_0 + \mathbf{Z}_1\mathbf{a}_1 + \mathbf{e}, \quad (2)$$

where \mathbf{y} is the vector of phenotypic observations; $\boldsymbol{\beta}$ is the vector of fixed effects; \mathbf{e} is the vector of residual terms, assumed as $N(\mathbf{0}, \mathbf{I}\sigma_{e_i}^2)$, with $\sigma_{e_i}^2$ being the residual variance in the i^{th} EI; and \mathbf{a}_0 and \mathbf{a}_1 are the vectors of random regression coefficients, i.e., the intercepts and EI slopes, respectively. In addition, \mathbf{X} is the incidence matrix for the fixed effects; \mathbf{Z}_0 is the incidence matrix for \mathbf{a}_0 , containing zeros and ones; \mathbf{Z}_1 is the incidence matrix for \mathbf{a}_1 , containing zeros and EI values. Notice that heterogeneity of residual variances was assumed in this model, since EI is randomly generated with different noises. With such model specifications, the expected value of \mathbf{y} across trials (i.e., model averages) is $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, meaning that $E(y_{ij}) = \beta_0 + \beta_1 EI$ for all y_{ij} , which refers to the overall mean in EI. The structure of variances and covariances is given by:

$$\text{Var} \begin{pmatrix} \mathbf{a}_0 \\ \mathbf{a}_1 \\ \mathbf{e} \end{pmatrix} = \begin{bmatrix} \mathbf{K}\sigma_{a_0}^2 & \mathbf{K}\sigma_{a_0 a_1} & 0 \\ \mathbf{K}\sigma_{a_1 a_0} & \mathbf{K}\sigma_{a_1}^2 & 0 \\ 0 & 0 & \mathbf{I}\sigma_{e_i}^2 \end{bmatrix},$$

where \mathbf{K} is the kinship matrix; $\sigma_{a_0}^2$, $\sigma_{a_1}^2$ and $\sigma_{a_0 a_1}$ (or $\sigma_{a_1 a_0}$) are variances of and covariances between the random regression coefficients, i.e., they are covariance functions that continuously describe the covariance structure for the trait, in the range of EI covered by the data; \mathbf{I} is an identity matrix; and $\sigma_{e_i}^2$ is the residual variance across environments, which is described in more detail below. The covariance matrix (Σ_a) between random genetic effects for an individual genotype is:

$$\Sigma_a = \begin{bmatrix} \sigma_{a_0}^2 & \sigma_{a_0 a_1} \\ \sigma_{a_0 a_1} & \sigma_{a_1}^2 \end{bmatrix}, \text{ and } \Sigma_a^{-1} \sigma_{e_i}^2 = \begin{bmatrix} \alpha_{00} & \alpha_{10} \\ \alpha_{10} & \alpha_{11} \end{bmatrix}.$$

The estimation and prediction of fixed and random effects according to the linear model specified above can be obtained through the mixed model equations (MME):

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_0 & \mathbf{X}'\mathbf{Z}_1 \\ \mathbf{Z}_0'\mathbf{X} & \mathbf{Z}_0'\mathbf{Z}_0 + \mathbf{K}^{-1}\alpha_{00} & \mathbf{Z}_0'\mathbf{Z}_1 + \mathbf{K}^{-1}\alpha_{10} \\ \mathbf{Z}_1'\mathbf{X} & \mathbf{Z}_1'\mathbf{Z}_0 + \mathbf{K}^{-1}\alpha_{10} & \mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{K}^{-1}\alpha_{11} \end{bmatrix} \begin{bmatrix} \beta \\ \mathbf{a}_0 \\ \mathbf{a}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}_0'\mathbf{y} \\ \mathbf{Z}_1'\mathbf{y} \end{bmatrix}$$

Genetic and residual variances are dependent on EI, that is, they can increase or decrease throughout sites. The genetic variance in the i th EI can be written as: $\sigma_{g_i}^2 = H_i' \Sigma_a H_i = \sigma_{a_0}^2 + 2EI_i \sigma_{a_0 a_1} + EI_i^2 \sigma_{a_1}^2$, where $H_i = \{1, EI_i\}$; $h_{g_i}^2 = \sigma_{g_i}^2 / (\sigma_{g_i}^2 + \sigma_{e_i}^2)$ is the relative genetic variance, herein termed heritability, for a particular EI_i ; $\sigma_{g_{ii^*}}^2 = H_i' \Sigma_a H_{i^*} = \sigma_{a_0}^2 + (EI_i + EI_{i^*}) \sigma_{a_0 a_1} + EI_i EI_{i^*} \sigma_{a_1}^2$ is the genetic covariance between sites i and i^* (dependent on EI information); $r_{g_{ii^*}} = \sigma_{g_{ii^*}} / (\sigma_{g_i}^2 \sigma_{g_{i^*}})^{1/2}$ is the genetic correlation between sites i and i^* (dependent on EI information), and $H_{i^*} = \{1, EI_{i^*}\}$.

For the use of MME, estimates of Σ_a and $\sigma_{e_i}^2$ are required. Estimates of the genetic parameters for spatial data were obtained by restricted maximum likelihood (REML) using the EM algorithm. At each EM iteration, the mixed model equations were solved marginally for the assumed EI. This process was implemented as a generalization of genetic parameters estimation for spatial data under heterogeneous residual variances. In summary, the EM algorithm alternates between calculating conditional expected values and maximizing the restricted likelihoods. In this context, the term \mathbf{y} (observed phenotypic values) are defined as “incomplete data”, whereas the “complete data” is defined as \mathbf{y} and the unobservable random effects (\mathbf{a}_0 and \mathbf{a}_1). The convergence

criterion was 10^{-6} , and the number of iterations varied around 100. The implemented EM algorithm was compared with the package “sommer” (Covarrubias-Pazaran 2016) with equivalent results. The variance component estimators under an EM algorithm are given by the following equations:

$$\hat{\sigma}_{e_i}^2 = \frac{y'y - \hat{\beta}' X'y - \hat{a}_0' Z_0'y - \hat{a}_1' Z_1'y}{N - \text{rank}(X)}, \text{ calculated marginally according to EI}_i \text{ values; and}$$

$$\hat{\Sigma}_{aij} = \frac{\hat{a}_{0i} K^{-1} \hat{a}_{1j} + \text{tr}(K^{-1}) C_{ij} \hat{\sigma}_{e_i}^2}{q},$$

where N is the total number of observations; $\text{rank}(X)$ is the rank of X (or number of linearly independent columns of X); q is the number of random elements (i.e., number of individuals or genetic values to be predicted); $\text{tr}()$ is the matrix trace operation or sum of matrix diagonal elements; C_{ij} is a partition of the inverse (C) of the coefficients matrix of the mixed model equations:

$$C = \begin{bmatrix} C_{XX} & C_{XZ_0} & C_{XZ_1} \\ C_{Z_0 X} & C_{Z_0 Z_0} & C_{Z_0 Z_1} \\ C_{Z_1 X} & C_{Z_1 Z_0} & C_{Z_1 Z_1} \end{bmatrix}$$

Starting from initial values for Σ_a and $\sigma_{e_i}^2$, we obtain β , a and λ from the mixed model equations, which are used to update the estimates of $\sigma_{e_i}^2$, $\sigma_{a_0}^2$, $\sigma_{a_0 a_1}$, $\sigma_{a_1 a_0}$ and $\sigma_{a_1}^2$ which are then returned to the mixed model equations. This is performed successively until convergence. Initial values of $\sigma_{a_0}^2$, $\sigma_{a_0 a_1}$ and $\sigma_{a_1}^2$ can be obtained using REML estimates $\hat{\sigma}_{g_i}^2$ for some EI.

The estimated genetic value (EGV) for any genotype j can be obtained for various sites (indicated by the EI) through the random linear regression: $EGV_{ij} = \hat{a}_{0j} + \hat{a}_{1j} EI_i$. The obtained EGV allows the re-ordering of the candidates to selection, according to the desired EI. Within a real breeding program, the proposed methodology can be applied to candidate genotypes to be tested into a next-stage recurrent breeding strategy, or even to evaluate cultivars in a final stage of a breeding cycle. These features are discussed in more detail in the Discussion section.

Based on the genetic correlation matrix among grouped EI (i.e., pixels addressing $0 \leq EI < 0.1, 0.1 \leq EI < 0.2, \dots, 0.9 \leq EI \leq 1$), we used a popular distance-based method, the Unweighted Pair Group Method with Arithmetic mean (UPGMA), to define *breeding zones*, which are locations within which genetic correlations between genotype and EI were optimized. In addition, a recommendation map for potential yield of the best-ranked genotypes was also provided as additional visual information derived from GIS–GEI.

Validation on unbalanced data scenarios

Finally, in order to evaluate the proposed methodology under unbalanced data conditions, the random regression model was tested for the following two situations described below. Inferences were carried out based on phenotypic averages of the selected genotypes (i.e., the best ranked genotype for each pixel).

- Random reduction of the number of trials (down to a minimum of three). The data from trials that were removed were assigned as validation groups, and those from the remaining trials as training groups. Thus, more than one draw was made in each reduction stage accounting for approximately a total of 1000 iterations;
- Different levels of genotype imbalance per trial with the constraint that all genotypes were always present in the analysis, i.e., for each genotype, the number of experiments was reduced (approximately 500 iterations were performed). The data from genotypes that were removed were assigned as validation groups, and those from the remaining genotypes as training groups.

Results

Data simulation and building of the environmental index (EI)

Simulated data sets are available at [https://figshare.com/articles/Enviromics_data/8264132]. The final phenotypic mean of the simulated yield trait was 46.10 units with standard deviation equal to 12.90 and minimum and maximum values equal to 5.50 and 117.10, respectively. To illustrate the simulation framework, three enviromic markers as well as the spatial distribution of the genotypic trials within the target area are illustrated in Fig. 2b.

The EI was constructed by extrapolating yield for the whole area, and subsequently rescaling it to a 0–1 interval. The EI mean and standard deviation for all pixels were 0.53 ± 0.27 , and for the 50 trials, they were 0.43 ± 0.29 . The EI distribution for the entire area presented an irregular shape with a higher density around 0.10 and 0.75. Lower densities were observed in the lower and upper tails (Fig. 3a). Figure 3b shows the behavior of the 100 genotypes along the range of EI values in terms of the reaction norm for all EIs. The EI distribution for the 50 trials showed a well-distributed range between 0 and 1 (Fig. 3c).

A great overlap was observed between the EI of the whole area and EI subset to the 50 trials (Fig. 3a), which is a desired feature for the application of the GIS–GEI approach. Moreover, the EI is expected to show a positive correlation with the simulated trait yield. The correlation between the

average yield of the 50 trials and the EI was 0.98, whereas it was 0.54 when considering both trials and genotypes. The decreased correlation is explained by the within-trial genotypic variability. The cross-validation predictive ability (predictive correlation) of the RF model was 0.87. These metrics together highlight the suitability of the GIS–GEI approach, even when there are no yield records from existing trials in the target area. Finally, the general mean and standard deviation of yield considering all 100 genotypes and interpolating to the whole area (10,000 pixels) were equal to 47.23 ± 5.46 .

Genetic modeling

Throughout the pixels across the area, and considering the assumed range of EI, the trait heritabilities varied between 0.41 (EI = 0.00) and 0.47 (EI = 1.00), with the lowest valley value equal to 0.19 when EI = 0.32. The two highest heritability values were observed at the EI distribution extremes (Fig. 4a). The average estimated heritability was 0.32, thus larger than the parameter value used in the simulation (without the GEI factor). In terms of the worst (EI = 0) and the best (EI = 1) environments, the genetic variances were 39.18 and 138.29, respectively. The lowest value was equal to 16.95 at EI = 0.30. The residual variance was 56.75 at EI = 0 (the lowest value) and 154.50 at EI = 1. The model without EI, which considers the trials as experimental blocks, yielded an overall $\hat{\sigma}_g^2 = 20.00$; $\hat{\sigma}_t^2 = 40.75$ and $\hat{\sigma}_e^2 = 106.23$, resulting in an overall trait heritability equal to 0.12.

Higher genetic correlations (r_g) were observed between EI = 1.00 and EI = 0.50, with r_g values ranging between 0.99 and 0.84, suggesting low re-ranking between genotypes in these locations (Fig. 4b). The lowest genetic correlation ($r_g = -0.49$) occurred between the extremes EI = 0.00 and EI = 1.00. The UPGMA procedure for grouping the locations with high genetic correlation between EI and genotypes resulted in the definition of three breeding zones: red (EI between 0.00 and 0.32); khaki (EI between 0.32 and 0.44) and blue (EI between 0.44 and 1.00) (Fig. 5a). Among the set of 50 trials, 25 belonged to the red zone, five experiments were within the khaki zone, and the blue zone encompassed 20 trials. Within the red zone, the genetic correlation (r_g) was, on average, equal to 0.92; within the khaki zone the average correlation was equal to 0.98; and for the blue zone it was 0.96. The genetic correlations between the red and khaki zones, red and blue and khaki and blue were 0.39, –0.05 and 0.70, respectively, indicating a substantial genetic re-ranking shared between the khaki and the blue breeding zones. The red zone is the one with the lowest yield potential, with an average EI of 0.16, the khaki zone has an intermediate potential (average EI of 0.38), and the blue zone has the highest potential (average EI of 0.72).

Of the 100 genotypes evaluated through GIS–GEI, there were five (namely G050, G061, G062, G065 and G098)

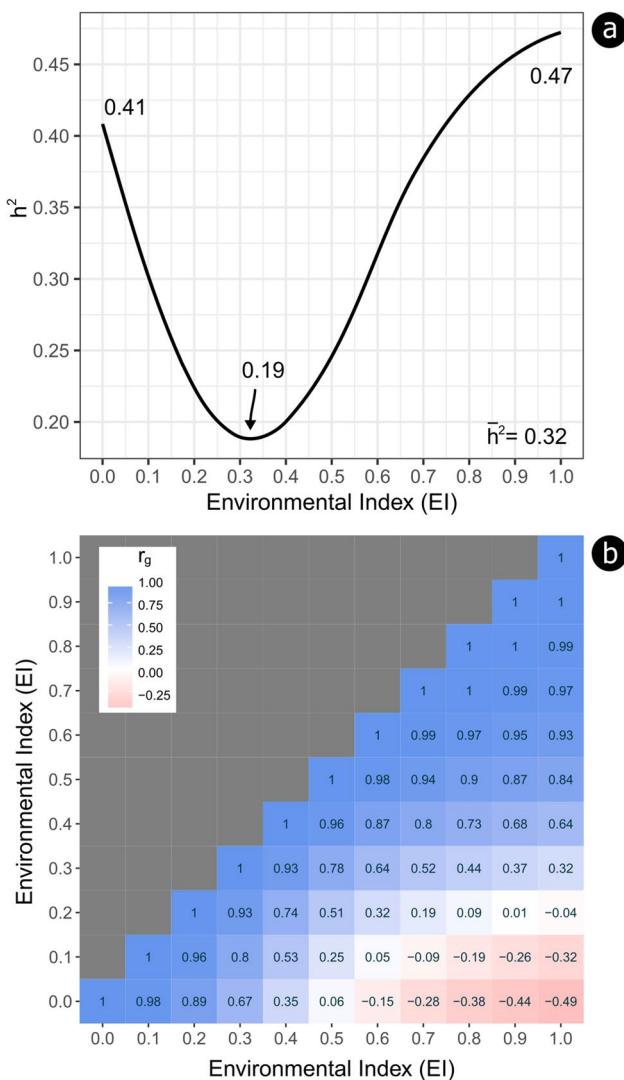


Fig. 4 Breeding parameters across the environmental index (EI). **a** Heritability estimate across the entire EI range. **b** Heatmap depicting the genetic correlation between different EI values

ranking first in at least one pixel, which were the best in 14.1, 30.2, 1.85, 15.3 and 38.6% of the whole area, respectively (Fig. 5b). When looking for the best genotype for each pixel, the tendency was that neighboring pixels shared the same genotype. The expected yield potential when deploying these five recommended genotypes in the field conditions of the evaluated envirotype is shown in Fig. 5c with an average yield equal to 62.24. It is important to notice that each pixel in the area has a particular genotypic ranking. For example, when removing the best one (first selected), the second-best genotypes were G006, G032, G041, G074 and G089 in 4.3, 44.9, 25.7, 0.1 and 25.0% of the area, respectively. Thus, different genotypic recommendation panels can be allocated to the area according to the different rankings of genotypes in the environmental gradient.

Validation in unbalanced data scenarios

Considering the two different conditions of unbalanced data tested, the reduction in the number of experimental trials resulted in a constrained EI predictive ability to recommend genotypes to locations in the area. In terms of yield potential, a reduction was observed for the selected genotypes, but the study indicates that an unbalancing of up to 40% of the total number of trials (a reduction from 50 to ~30 trials in our simulated case study) did not result in significant loss of predictive ability for genotype recommendation. However, using a very small number of experiments should be avoided, because this condition can, by chance, result in biased yield potential estimates for the area given the use of inadequate genotypes (Fig. 6a). In contrast, imbalance of genotypes within the experiments did not present major risks for the application GIS–GEI. It is a good practice, however, to have genotypes allocated in at least 20 of the 50 evaluated experiments (Fig. 6b), to ensure an appropriate level of representativeness of the genotypes across the land area.

Discussion

Enviromics in the quantitative genetics' framework

Finding an environmental benchmark that represents agricultural productivity is one way of characterizing the quality of the environment, independent of the in-depth knowledge of the sources of the observed variation in phenotypic expression. Individually, a particular environmental factor that affects plant growth and yield (an envirotype) may not significantly affect large groups of genotypes that share diverse levels of genetic relationship across locations in a wide geographical area. For a single envirotype, repeated observation of an individual genotype (or its progeny) may show that its performance has a spatially dependent pattern across the locations within the area. In the concept underlying our study, the geographical area is a virtual image so that each location is a multiple overlapped pixel space in a coordinate system. Each pixel has intensity in some range, so that the envirotypes are now an image consisting of a big collection of pixels (Fig. 1). It is noteworthy that only a few pixels contain genetic trials with phenotypic and possibly genotypic information, but all pixels should have environmental information.

Grouping together images with similar pixels is analogous to the concept of linkage disequilibrium (LD) among genetic markers. It is known that many loci may not be directly involved in the expression of the phenotype itself, but their association to the causal loci due to LD increases the accuracy of predictive models (Jannink et al. 2010).

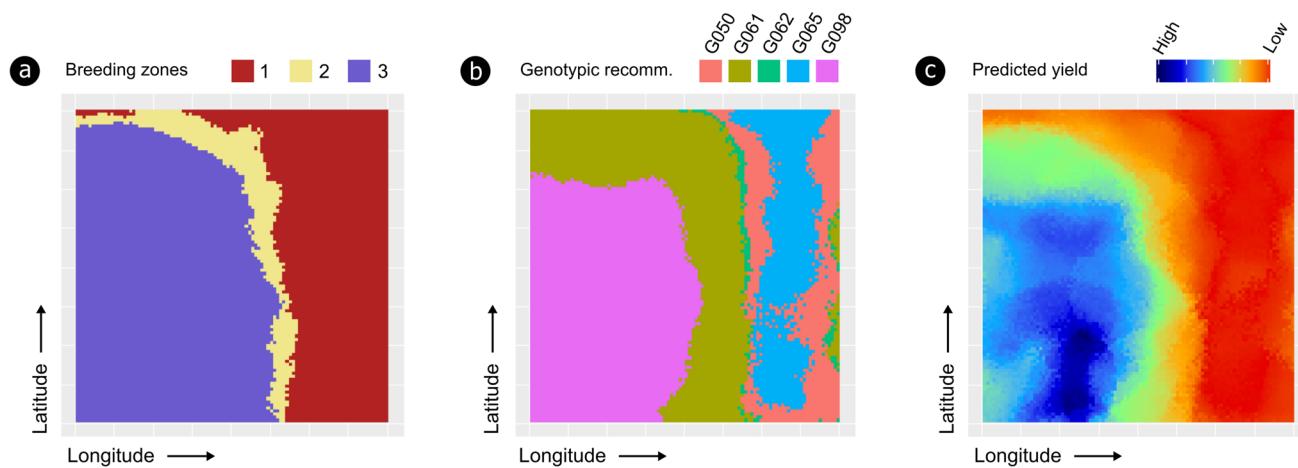


Fig. 5 Genetic extrapolations. **a** Breeding zones map depicting the three zones fitted by the UPGMA clustering procedure. A higher genetic correlation is observed within zones, as opposed to across zones, indicating fewer genotypic ranking changes are expected with

zones. **b** Recommendation map depicting the distribution of the top five genotypes for yield. **(c)** Yield extrapolation for the top five genotypes across the land area

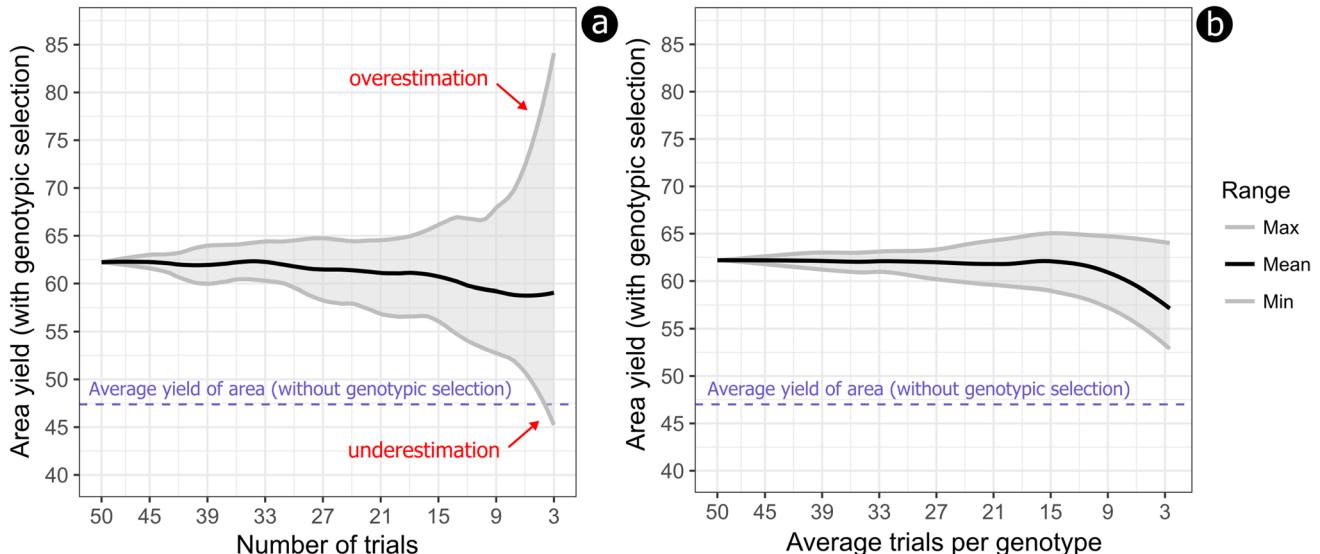


Fig. 6 Enviromic prediction under two unbalanced data scenarios. **a** Impact of reducing the number of trials on yield projection of the selected genotypes. **b** Impact of reducing the average number of trials per genotype

als that include a specific genotype on yield projection of the selected genotypes. The dotted blue line is the average yield for the entire area without genotypic selection/recommendation

The conceptualization of enviromics is built on this same principle. If measurable environmental variables are missing or neglected (equivalent to missing molecular marker genotype data at genomic regions), the combination of enviromic markers may still represent the explanation of a representative fraction of the phenotypic variation, either for all genotypes or for just some of them. Additionally, under a structured environmental and genetic (pedigree- and/or marker-based) dataset, the model's predictive ability may be enhanced using the general concept of distance based on latitude/longitude (for example) and/or LD between markers. In

this case, cross-validation analysis can be performed considering specific spatial patterns (for instance, testing extreme environments) and genetic relationships (for instance, testing subpopulations) in order to account for possible structures in the dataset.

The resolution of an enviromic model will depend primarily on the size of the available sample unit (pixel). For example, when using geographic information, the pixel size will delimit the refinement level of the genotypic evaluation (Marcatti et al. 2017). Environmental information with high spatial resolution, particularly adequate to contemplate

forest stands, crop plantations and livestock farms, should be preferred to improve the accuracy of the models.

Equivalent to the metrics of Call Rate and MAF (Minor Allele Frequency), when using DNA marker data, enviromic markers also require quality control measures. The number of missing values described by the Call Rate parameter can be solved by adopting two strategies. The first would be to increase the size of all pixels in the area, and use the average information available in the neighborhood. Although this will increase the unit-of-handling area, thereby decreasing the accuracy of the recommendation or prediction, it may still be a suitable alternative. The second, probably more feasible strategy, would be to impute the values of the missing pixels by means of kriging, for which both the neighborhood values and the values of other enviromic markers can be used. For example, similar temperature values between close locations are more likely than between distant ones, and on a local scale, the temperature correlates well with terrain elevation, or with global latitude. Like MAF for a genetic marker, an enviromic marker with small variance would have a low SEV (Scale of Environmental Variation), such that environmental covariates would have a low variance in under-sampled pixels compared to all pixels in the area. In this case, enviromic markers with low SEV, i.e., low information content, could be discarded. In the present study, markers were simulated without missing data. Nonetheless, additional follow-up studies should be performed to define optimal thresholds for quality control metrics of enviromic markers.

Enviromics applied to general breeding practice

Within the context of maximizing genetic gains through selective breeding, phenotypic traits are thought to be affected by several environmental variables. An environmental variable that is favorable for the field deployment of a particular plant species is not always the same for another species. For example, while the presence of exchangeable aluminum in the soil is unfavorable to most annual crops, for some forest tree crops it will be irrelevant (Poschenrieder et al. 2008). In addition, such environmental conditions in a land area may interact at different levels to determine the final genotype yield (Resende et al. 2018). Additive Main-effects and Multiplicative Interaction (AMMI) and Genotype \times Genotype \times Environment interaction (GGE) models have been successfully applied to quantify individual site effects and their interactions on the expression of complex traits (Yan et al. 2000; Gauch 2006) under a multiple latent variables framework.

Along these lines, enviromic models can be easily generalized with multiple regressors to address specific objectives, such as to relate productivity of annual grain crops to flowering time and environmental markers based on photoperiods

at different photometric scales (Millet et al. 2019). Moreover, such model can be used to conjecture on rotation yield of perennial crops, in which the stands are planted and harvested at different times, by using enviromic markers allowing differentiation of particular growth periods in each stand. Lastly, they can aid in proposing environmental-based strategies for integrated management of biotic stresses, thus providing model ability to capture effects related to resistance to different pathogens in the field. Another alternative is to group the environmental markers into sets by some sort of similarity, such as temperatures measured daily throughout the entire period of growth and development of the crop.

Several types of environmental variables can be used in enviromics, such as temporal climatic information (Fick and Hijmans 2017) and vegetation indices obtained from remote sensing based canopies (Xue and Su 2017). Xu (2016) lists a number of environmental factors affecting plant growth and yield. Categorical indices can also be used, such as the climatic index of *Köppen* (Kottek et al. 2006), or even soil classes (Hartemink 2015), since each class is properly represented by one or more experimental points. Although enviromics could be expected to be more suitable for continuously distributed traits, given the intimate relationship between the GEI and quantitative traits, it can also be applied to more discretely distributed traits related to resistance or tolerance to biotic and abiotic stressors, provided that the model for generation of EI is fed with data from environmental variables that trigger the targeted physiological stresses. For instance, if the focus of the study is to improve drought tolerance, it is extremely important that data on water availability (or a highly correlated variable used as proxy) is included in the enviotyping routine. Similarly, if the focus is resistance to tick-borne diseases in animals, it is important to feed the model with data on the occurrence and density of the arachnids responsible for the disease (Giles et al. 2014). Thus, it is possible to adopt the enviromics approach to traits with different levels of heritability, including disease escape in regions with endemic pests for example (Shakoor et al. 2017).

Enviromics may be applied in conjunction with the breeding program strategy specific to each crop or even livestock species or at any specific stage of the program. Annichiarico and Iannucci (2008) suggested genetic improvement pathways specific to each geoclimatic area based on distinct genetic basis and selection environments. In maize breeding, for example, differences in the capture of genetic variances can be observed along the trials carried throughout the breeding program (Cooper et al. 2014). Enviromics may also have applications in recurrent selection programs, directing preferred crosses and selecting suitable parents to specific sites. On the other hand, in annual crop improvement programs, propagated exclusively via crossing, the recommended genetic materials can be found from the parents'

performance, even if these parents have never been tested in the target environment. One can take advantage of the predicted pixel-by-pixel genetic values by coupling kinship structures to the enviromics models, according to the stage of the breeding program, and simply devise the strategy that best suits the specific situation or application. One can also perform it in a mixed model approach, by using the direct product of the kinship matrix (K) with the reaction norm covariance (Σ_a), i.e., as $K \otimes \Sigma_a$, where \otimes is the Kronecker product.

Additionally, the generation of phenotypic variability is fundamental to any breeding program, as a result of harnessing the genetic resources in germplasm banks. The enviromics models can also direct the rescue of materials with location-specific phenotypes, or according to the local consumer profile or preference, directing the trials to target population of environments (TPE) (Chapman et al. 2000). The TPE considering all pixels with a rectangular area (a naïve strategy) is just an illustration, but it can be much more general, such as combining on-farm data from an entire country, sets of countries or data collected across countries, by which economically underdeveloped countries could also benefit (Pérez-Rodríguez et al. 2017). In other words, a TPE can be quite large, depending on data availability (Fig. 2a), as envirotypic data should only be linked to the crop growth period. It is important to notice that while breeders do not need to have an in depth understanding of geotechnologies, a reasonable assumption is that they should get acquainted with the use of spatial data.

Selection and recommendation via enviromics can also be addressed at multiple times. In a big data context, it is perfectly feasible to assume that one or more environments sampled today will behave similarly to others in the future. Nevertheless, considering a specific location, the value of environmental covariates cannot be known with certainty for the coming year. In order to get around this issue, the models must be implemented considering the local environmental data in times closer to the moment of the crop implementation, or alternatively implemented using environmental data from future weather forecasts. This premise is also valid for climate change scenarios, in which one or more environments already represent sites with warming weather or even rain scarcity that another environment may have in the future. It is prudent nonetheless to mention that other studies should be done to validate the operation of enviromic models for environmental scenarios with a high degree of uncertainty.

Still from the point of view of selection, a great benefit of enviromics models is the possibility of quantifying residual variances inherent to the whole geographic environment, and possibly maximizing the genetic variance captured, especially in those breeding program stages that eventually display low heritabilities. Thus, in addition to maximizing the

genetic variance components for the selection of individuals, families or parents (Gomez-Raya and Burnside 1990), it is possible to optimize selection by identifying sites that provide a higher ratio between genetic and residual variances. Especially in populations with a high level of improvement, or even in the stages of the program when trait heritabilities are lower, such high-heritability environments can help increasing the accuracy of genotypic selection.

Enviromics to improve cultivar recommendation

In the routine of a crop breeding program, it is common practice to allocate experiments to several sites throughout a land area, to better cover the spectrum of possible sites for genotype recommendation (Annicchiarico et al. 2006). This procedure aims to identify those particular genotypes that may be recommended for the largest number of sites in terms of their phenotypic means in comparison to their competitors. Underlying this practice is the assumption that the top genotypes in the evaluated trials may perform better across the whole area in which the breeding experiments were deployed. The extrapolation arguments are mostly *ad-hoc* as they rely on parameter estimates that may be unrealistic for a new location. It can be noted, thus, that these practices often assume that the evaluated and non-evaluated sites across the land area show a high level of correlation in terms of the effect of their environmental variables, mainly on growth and yield. The high correlations among environmental variables across the area are indeed a necessary condition to preclude changes in genotypic ranking.

Models based on reaction norms are interesting because they show the stability of genotypes to environmental changes and their adaptive ability. Although stable genotypes may be attractive, as they generally do not result in unexpected surprises to the breeder, adaptive genotypes may respond better to crop management procedures such as irrigation and fertilization (Cobb et al. 2013), control of biotic agents (Shakoor et al. 2017) or even improvements in animal comfort in the case of livestock. Negative predicted yields at lower EI may indicate that the genotype would not survive in such environments, or would have exceedingly low yields. On the other hand, genotypes with positive predicted yields in extreme environments may indicate genotypes with high resilience potential. These resilient genotypes usually do not have high regression slopes and are considered to be less adaptive and very stable in their response to different environmental variables (Fig. 3c).

Although simulations were carried out with a relatively low heritability trait, results showed that it is possible to capture greater heritabilities in extreme locations (i.e., EI close to 0 or to 1) (Fig. 4a). Although Bänziger et al. (2000) argue that differences between genotypes are generally smaller under stress and larger differences are, therefore,

more difficult to detect, sites with limiting characteristics to plant development may lead to selection pressure on the most adapted individuals (McKown et al. 2014), favoring the manifestation of more productive genotypes. Sites with better conditions may also provide similar results, since some genotypes may demonstrate better ability to take advantage of the available resources. This behavior is usually demonstrated in reaction norm studies in animal breeding (Ribeiro et al. 2015), forest tree improvement (Resende et al. 2018) and agricultural grain crops (Jarquín et al. 2014).

Identifying the spatial boundaries in which a selected genotype may be recommended would be an important advance (Annicchiarico et al. 2005), considering that it is usually impossible to deploy a large number of trials in a certain area due to limited resources. Across the EI built with the GIS–GEI method, our results showed genetic correlations between -0.49 and 1.00 , indicating that depending on the relationship between selection and recommendation site, it is possible to accurately detect superior genotypes avoiding make a major mistake (Fig. 4b). The efficacy of a system of evaluation of cultivars depends largely on the genetic correlation between genotype performance in multi-environmental trials (MET) (Löffler et al. 2005). Pixels can then be grouped in sets that produce a similar ranking of the genotypes, i.e., the breeding zones (also known as mega-environments). To explore different scenarios of genetic correlations between sites, a full range of variance–covariance matrices could be considered in GEI models, in particular addressing Genotypes \times Mega-environments, e.g., *compound symmetry*, to access a single covariance between site groups, either *unstructured* or *factor analytical* matrices, allowing heterogeneous residues per site group (Kleinknecht et al. 2013).

A map of breeding zones is therefore indicated, precisely linked to the size of the pixel used (Fig. 5a). In our simulations, three breeding zones were assumed, with high genetic correlations within each one, indicating high agreement between the selection and recommendation site. The definition of fewer breeding zones provided a lower genetic correlation within them (results not shown), and adopting more breeding zones caused the opposite. In this context, three zones provided reasonable amount of information to be exploited under a breeding framework. When no trials are available in any particular breeding zone, one can use selection decisions from a zone with the highest average correlation. In our simulation, the genetic correlation between the khaki and blue zones was 0.70 , indicating that no drastic GEI would be expected between these two macro-environments. Furthermore, within a breeding area, priority should be given to allocate trials where a better ability to capture missing heritability is expected, as sites with higher heritability are expected to provide higher selection accuracy.

With a recommendation map in hands, the breeder or the agricultural extension service may indicate better genotypes for very specific boundaries in the area (Fig. 5b). More than one genotype should be recommended per site, contemplating two points highlighted by Annicchiarico et al. (2006): (i) mitigate the risk of unexpected susceptibility to a biotic or abiotic stress by a single recommended cultivar; and (ii) take into account genotypic differences that were not statistically significant during the recommendation analysis. Schedule and logistic issues can also be incorporated into enviromics models, such as the availability of improved seeds for grain crops or seedlings for forest trees. If the recommended genotypes are effectively deployed in the area, the potential yield may increase by 32% when compared to the average yield of the area planted with unselected genotypes (Fig. 6). This is a key point as it addresses the challenge of increasing production with the same land area.

The proposed genetic modeling adopted in the GIS–GEI methodology is based on mixed models, which deliver a powerful statistical framework for dealing with unbalanced data (Gianola and Rosa 2015). However, when tested with an unbalanced dataset, by radically reducing the number of experiments in the field, even such models were not able to satisfactorily select genotypes that maintained the average yield. Although mixed models have the ability to work with unbalanced data, like any other statistical procedure, they are unable to correct for exceedingly inefficient sampling (Schmidt et al. 2019). On the other hand, genotypic imbalance was not detrimental to the recommendation ability of the GIS–GEI model. Using information from mega-environments (here denoted as breeding zones—Fig. 5a), González-Barrios et al. (2019) demonstrated that genotypic imbalance in different environments can be circumvented with special designs, corroborated by our inferences in the enviromics context. In practice, these results indicate that if a choice has to be made, it is better to establish more and smaller trials with less replication of genotypes than deploying a smaller number of more highly replicated trials (Moehring et al. 2014).

An insufficient number of trials may sometimes over- or underestimate the yield of selected genotypes. This happens because when the trials show yields above the average of the area, the yield of the recommended genotypes will be overestimated in relation to the expected future yield. The reverse is also true with an inadequate sampling of trial environments, resulting in below-average yields. While Pérez-Rodríguez et al. (2015) achieved predictive capacity gains of approximately $4\text{--}6\%$ when evaluating only nine experiments, Jarquín et al. (2014) achieved up to 34% gain by evaluating on-farmer trials addressing 134 locations.

Finally, when there is no phenotypic information from the target environment, predictive ability measures can be achieved through individual prediction error variance (PEV)

because the proposed estimation method is based on mixed model equations. When using Bayesian methods, an equivalent measure of uncertainty can be obtained by the posterior standard deviation (Sorensen and Gianola 2002).

Future perspectives of enviromics in breeding

New genomic methods have paved the way for predicting phenotypes of unobserved genotypes in untested environments (Malosetti et al. 2016; Voss-Fels et al. 2019). Our work provides a glimpse into the promising area of further including enviromics in this context to optimize the ability to predict the performance of breeding material, especially for species subject to complex response patterns across environments and time. The enviromics models presented may seamlessly incorporate any kinship structure (K) by changing the existing relationship matrix between genotypes from the traditional numerator relationship matrix A , to the genomic relationship matrix G . In the context of using G matrix, improved genomic predictions of complex phenotypes are expected across environments by simultaneously taking into account information from molecular (e.g., SNP data) and enviromic markers, especially for late-expressing or difficult to measure phenotypes. Additionally, given the very high genotyping density possible with current SNP panels, the enviromics approach may also be used to generate a catalog of SNP markers for each micro-region in a target area, naturally respecting the strong GEI typically displayed by traits of low heritability. Additionally, genomic-based enviromics models can also exploit precise field-level information of the trial, such as competition between plants (Cappa et al. 2017). Finally, environmental information can be incorporated into prediction models via envirotyping (Xu 2016) combining genomics, controlled crosses, germplasm data and next-generation phenotyping (Cobb et al. 2013). We can also mention that epigenetic effects triggered by changes in the environment can also be better captured with enviromic models, by the incorporation of epigenetic matrices T into the breeding analysis (Varona et al. 2015).

The construction of environmental indexes can also benefit from the application of artificial intelligence, eco-physiological process models (Asseng et al. 2013) or biogeographical similarity approaches (Vilhena and Antonelli 2015), in order to obtain indexes that relate more closely to the target trait, avoiding the inclusion of trait-irrelevant enviromic markers. As the methodology maximizes the number of recommended genotypes based on the potential yield of the area, one should also be concerned with the maintenance of genetic variance throughout the selection cycles. This can be accomplished by using optimization models that maximize selection gains with genetic diversity of the selected individuals (El-Kassaby and Lstiburek 2009; Mullin 2017), and also provide multi-trait gains for all pixels in the geoprocessing area, in scenarios of

different environmental conditions (Bustos-Korts et al. 2019). Another interesting strategy would be the recommendation of specific parents and crosses for specific environments, generating progeny based on the combination of the best individuals by environment (van Ginkel and Ortiz 2017).

A possible limitation of the proposed GIS-GEI method is the assumption of a common GEI profile for two sites with similar environment index. Different factors that make the two sites similar for productive potential may in fact result in different GEI patterns. To overcome this issue, a single-step enviromics model would improve the prediction accuracy over environments. These models can be frequentist or purely Bayesian, since associated covariates are treated as unknown, thereby allowing inference for all unknowns together within a single-step linear random regression as follows: $y = X\beta + Pw + Z_0a_0 + Z_1a_1 + e$, where y is the vector of observations, β is the fixed effects vector of order p , w is the vector of environmental effects, a_0 is the vector of random genetic intercepts, a_1 is the vector of random genetic slopes. It can be assumed that $[a_0, a_1] \sim N(0, G \otimes \Sigma_a)$, with G being the genomic relationship matrix under a GBLUP framework. Furthermore, X , P , Z_0 and Z_1 are the respective known incidence matrices, whereas each row of Z_1 has exactly one element equal to the environmental covariate (w_i or an estimate of w_i), with all other elements in that row equal to zero, and e is the vector of random residuals. To infer environmental sensitivities, three stages are required. The first stage defines the distribution of the phenotypic data conditional on all other parameters; the second stage is represented by the prior distributions of the location parameters (β , w , a_0 and a_1); and the third one is based on specifying prior distributions for the (co) variance components. In addition, the previously described model is based on GBLUP, which is not suitable for variable selection (shrinkage estimates) in the presence of many correlated covariates (enviromic or molecular markers). However, we believe that other specific genomic prediction models like Bayes A, Bayes B or Bayesian LASSO (Gianola et al. 2009) can be adapted to infer on GEI. Furthermore, new combinations and synergism between enviromics and modern GEI approaches, using the mixed models explored here, as well as crop growth models, are envisioned.

Concluding remarks

In the context of breeding practice, the term enviromics involves the application of envirotyping techniques to describe the performance of a plant or animal along the different gradients of a large number of environmental variables. To account for the environmental effect on a phenotype, we have developed an infinitesimal-like approach taking into account additive and non-additive contributions from enviromic markers in an analogous

fashion as traditional quantitative genetics models. Multivariate models, such as principal components analysis and modern approaches from artificial intelligence will likely allow better definition of enviromic markers improving the computation of EI. Enviroomics models are flexible and can be easily adjusted according to changes in the environment, a particularly useful tool in the context of climate change scenarios. Additionally, enviromic markers are climatic or landscape-based variables. As such, they are not only universally applicable to any animal or plant species, but more importantly they can be obtained and used jointly with omics marker data such as DNA, RNA, proteomic, metabolomic and epigenomics. Finally, we have also proposed a methodology called GIS–GEI, which is a remake of classical GEI approaches derived from the enviroomics conceptualization. It can be useful for recommending genotypes for specific areas, for defining optimal breeding zones (i.e., mega-environments), for understanding the spatial boundaries in which a genetic trial can be used for selecting breeding material, and for identifying sites that provide better capabilities for the genetic expression of a phenotypic trait. We believe that the concept presented here should represent a relevant advance in the existing approaches and be a useful addition to the toolbox of modern breeding programs, especially with the increasing availability of genomic and environmental big data.

Acknowledgements We would like to acknowledge professors Gustavo E. Marcatti from UFSJ and Helio G. Leite from UFV for their valuable input regarding the use of GIS and landscape level plant management data. We also want to thank the editor Martin P. Boer and the four anonymous reviewers for their considerations and suggestions to improve the manuscript. This work was partly supported by a post-doctoral grant (project FAP-DF 0193.001198/2016) to RTR (grant 4026592013-9), DFG grant (PI 377/20-1) to HPP, and infrastructure funding from the Brazilian Council for Scientific and Technological Development (CNPq/PQ 307096/2018-1).

Author's contribution RTR, HPP, OBSJ, GJMR and FFS were involved in simulation and data analysis. MDVR, RTR, HPP, GJMR and FFS developed mathematical–statistical procedures and notations. RTR, HPP, GJMR, OBSJ, FFS and DG wrote the manuscript. All authors reviewed the manuscript.

Availability of data and materials This article is the peer-reviewed version of the preprint posted august 06, 2019 at: <https://doi.org/10.1101/726513>. Code used to generate the simulated data, and the envirotypic (File S1) and phenotypic data (File S2) are available in the repository: https://figshare.com/articles/Enviroomics_data/8264132.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

- Acosta-Pech R, Crossa J, de los Campos G et al (2017) Genomic models with genotype \times environment interaction for predicting hybrid performance: an application in maize hybrids. *Theor Appl Genet* 130:1431–1440
- Annicchiarico P, Bellah F, Chiari T (2006) Repeatable genotype \times location interaction and its exploitation by conventional and GIS-based cultivar recommendation for durum wheat in Algeria. *Eur J Agron* 24:70–81
- Annicchiarico P, Bellah F, Chiari T (2005) Defining subregions and estimating benefits for a specific-adaptation strategy by breeding programs. *Crop Sci* 45:1741–1749
- Annicchiarico P, Iannucci A (2008) Breeding strategy for faba bean in southern Europe based on cultivar responses across climatically contrasting environments. *Crop Sci* 48:983–991
- Asseng S, Ewert F, Rosenzweig C et al (2013) Uncertainty in simulating wheat yields under climate change. *Nat Clim Chang* 3:827
- Bänziger M, Edmeades GO, Beck D, Bellon M (2000) Breeding for drought and nitrogen stress tolerance in maize: from theory to practice. CIMMYT, Mexico
- Beckers J, Wurst W, de Angelis MH (2009) Towards better mouse models: enhanced genotypes, systemic phenotyping and envirotype modelling. *Nature Rev Genet* 10(6):371–380
- Bustos-Korts D, Boer MP, Malosetti M et al (2019) Combining crop growth modelling and statistical genetic modelling to evaluate phenotyping strategies. *Front Plant Sci* 10:1491
- Calus MPL, Bijma P, Veerkamp RF (2004) Effects of data structure on the estimation of covariance functions to describe genotype by environment interactions in a reaction norm model. *Genet Sel Evol* 36:489
- Cappa EP, El-Kassaby YA, Muñoz F et al (2017) Improving accuracy of breeding values by incorporating genomic information in spatial-competition mixed models. *Mol Breed* 37:125
- Chang J-H (2017) Climate and agriculture: An ecological survey, 1st edn. Routledge, New York, USA
- Chapman SC, Hammer GL, Butler DG, Cooper M (2000) Genotype by environment interactions affecting grain sorghum. III. Temporal sequences and spatial patterns in the target population of environments. *Aust J Agric Res* 51:223–234
- Cobb JN, DeClerck G, Greenberg A et al (2013) Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. *Theor Appl Genet* 126:867–887
- Cooper M, Messina CD, Podlich D et al (2014) Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. *Crop Pasture Sci* 65:311–336
- Costa-Neto GMF, Júnior OPM, Heinemann AB et al (2020) A novel GIS-based tool to reveal spatial trends in reaction norm: upland rice case study. *Euphytica* 216:37
- Covarrubias-Pazaran G (2016) Genome-assisted prediction of quantitative traits using the R package sommer. *PLoS ONE* 11:1–15
- Des Marais DL, Hernandez KM, Juenger TE (2013) Genotype-by-environment interaction and plasticity: exploring genomic responses of plants to the abiotic environment. *Annu Rev Ecol Evol Syst* 44:5–29
- Eberhart SA, Russell WA (1966) Stability parameters for comparing varieties. *Crop Sci* 6:36–40
- Editorial (2015) Growing access to phenotype data. *Nat Genet* 47:99
- El-Kassaby YA, Lstiburek M (2009) Breeding without breeding. *Genet Res (Camb)* 91:111–120
- Elias AA, Robbins KR, Doerge RW, Tuinstra MR (2016) Half a century of studying genotype \times environment interactions in plant breeding experiments. *Crop Sci* 56:2090–2105

- Fernandes AFA, Alvarenga ÉR, Alves GFO et al (2019) Genotype by environment interaction across time for Nile tilapia, from juvenile to finishing stages, reared in different production systems. *Aquaculture* 513:734429
- Ferrero-Serrano Á, Assmann SM (2019) Phenotypic and genome-wide association with the local environment of *Arabidopsis*. *Nat Ecol Evol* 3:274–285
- Fick SE, Hijmans RJ (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int J Climatol* 37:4302–4315
- Finlay KW, Wilkinson GN (1963) The analysis of adaptation in a plant-breeding programme. *Aust J Agric Res* 14:742–754
- Gad SC (2008) Preclinical development handbook: toxicology. Wiley, New York, USA
- Garnett T, Appleby MC, Balmford A et al (2013) Sustainable intensification in agriculture: premises and policies. *Science* 341:33–34. <https://doi.org/10.1126/science.1234485>
- Gauch H, Zobel RW (1997) Identifying mega-environments and targeting genotypes. *Crop Sci* 37:311–326
- Gauch HG (2006) Statistical analysis of yield trials by AMMI and GGE. *Crop Sci* 46:1488–1500
- Gianola D, de los Campos G, Hill WG et al (2009) Additive genetic variability and the Bayesian alphabet. *Genetics* 183:347–363
- Gianola D, Rosa GJM (2015) One hundred years of statistical developments in animal breeding. *Annu Rev Anim Biosci* 3:19–56
- Giles JR, Peterson AT, Busch JD et al (2014) Invasive potential of cattle fever ticks in the southern United States. *Parasit Vectors* 7:1–11
- Gomez-Raya L, Burnside EB (1990) The effect of repeated cycles of selection on genetic variance, heritability, and response. *Theor Appl Genet* 79:568–574
- González-Barrios P, Díaz-García L, Gutiérrez L (2019) Mega-environmental design: using genotype × environment interaction to optimize resources for cultivar testing. *Crop Sci* 59:1899–1915
- Haghghatalab A, Crain J, Mondal S et al (2017) Application of geographically weighted regression to improve grain yield prediction from unmanned aerial system imagery. *Crop Sci* 57:2478–2489
- Hartemink AE (2015) The use of soil classification in journal papers between 1975 and 2014. *Geoderma Reg* 5:127–139
- Houle D, Govindaraju DR, Omholt S (2010) Phenomics: the next challenge. *Nat Rev Genet* 11:855
- Hyman G, Hodson D, Jones P (2013) Spatial analysis to support geographic targeting of genotypes to environments. *Front Physiol* 4:1–13
- Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. *Brief Funct Genomics* 9:166–177
- Jarquín D, Crossa J, Lacaze X et al (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor Appl Genet* 127:595–607
- Kasampalis DA, Alexandridis TK, Deva C et al (2018) Contribution of remote sensing on crop models: a review. *J Imaging* 4:52
- Kleinke K, Möhring J, Singh KP et al (2013) Comparison of the performance of best linear unbiased estimation and best linear unbiased prediction of genotype effects from zoned Indian maize data. *Crop Sci* 53:1384–1391
- Koch J, Stisen S, Refsgaard JC et al (2019) Modeling depth of the redox interface at high resolution at national scale using random forest and residual gaussian simulation. *Water Resour Res* 55:1451–1469
- Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Ijcai*. pp 1137–1145
- Kottek M, Grieser J, Beck C et al (2006) World map of the Köppen-Geiger climate classification updated. *Meteorol Zeitschrift* 15:259–263
- Liaw A, Wiener M (2002) Classification and regression by randomForest. *R news* 2:18–22
- Lindblom J, Lundström C, Ljung M, Jonsson A (2017) Promoting sustainable intensification in precision agriculture: review of decision support systems development and strategies. *Precis Agric* 18:309–331
- Löffler CM, Wei J, Fast T et al (2005) Classification of maize environments using crop simulation and geographic information systems. *Crop Sci* 45:1708–1716
- Malosetti M, Bustos-Korts D, Boer MP, van Eeuwijk FA (2016) Predicting responses in multiple environments: issues in relation to genotype × environment interactions. *Crop Sci* 56:2210–2222
- Marcatti GE, Resende RT, Resende MDV et al (2017) GIS-based approach applied to optimizing recommendations of *Eucalyptus* genotypes. *For Ecol Manage* 392:144–153
- Matise TC, Ambite JL, Buyske S et al (2011) The Next PAGE in understanding complex traits: design for the analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. *Am J Epidemiol* 174:849–859
- McKown AD, Klápková J, Guy RD et al (2014) Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. *New Phytol* 203:535–553
- Millet EJ, Kruijer W, Coupel-Ledru A et al (2019) Genomic prediction of maize yield across European environmental conditions. *Nat Genet* 51:952–956
- Moehring J, Williams ER, Piepho H-P (2014) Efficiency of augmented p-rep designs in multi-environmental trials. *Theor Appl Genet* 127:1049–1060
- Mota LFM, Lopes FB, Júnior GAF et al (2020) Genome-wide scan highlights the role of candidate genes on phenotypic plasticity for age at first calving in nellore heifers. *Sci Rep* 10:1–13
- Mullin TJ (2017) OPSEL 2.0: a computer program for optimal selection in tree breeding, 1st edn. Skogforsk, Uppsala, Sweden
- Nicotra AB, Atkin OK, Bonser SP et al (2010) Plant phenotypic plasticity in a changing climate. *Trends Plant Sci* 15:684–692
- Oliver MA, Webster R (2015) Basic steps in geostatistics: the variogram and kriging. Springer, Berlin
- Pérez-Rodríguez P, Crossa J, Bondalapati K et al (2015) A pedigree-based reaction norm model for prediction of cotton yield in multienvironment trials. *Crop Sci* 55:1143–1151
- Pérez-Rodríguez P, Crossa J, Rutkoski J et al (2017) Single-step genomic and pedigree genotype × environment interaction models for predicting wheat lines in international environments. *Plant Genome* 10:1–15
- Piepho H-P, Denis J-B, van Eeuwijk FA (1998) Predicting cultivar differences using covariates. *J Agric Biol Environ Stat* 151–162
- Poschenrieder C, Gunsé B, Corrales I, Barceló J (2008) A glance into aluminum toxicity and resistance in plants. *Sci Total Environ* 400:356–368
- Rauw WM, Gomez-Raya L (2015) Genotype by environment interaction and breeding for robustness in livestock. *Front Genet* 6:310
- Raza A, Razzaq A, Mehmood SS et al (2019) Impact of climate change on crops adaptation and strategies to tackle its outcome: A review. *Plants* 8:34
- Resende MDV, Rezende GDSP, Fernandes JSC (2001) Regressão aleatória e funções de covariância na análise de medidas repetidas. *Rev Mat Estat* 19:21–40
- Resende RT, Soares AAV, Forrester DI et al (2018) Environmental uniformity, site quality and tree competition interact to determine stand productivity of clonal *Eucalyptus*. *For Ecol Manage* 410:76–83
- Ribeiro S, Eler JP, Pedrosa VB et al (2015) Genotype × environment interaction for weaning weight in Nellore cattle using reaction norm analysis. *Livest Sci* 176:40–46

- Riggs DW, Yeager RA, Bhatnagar A (2018) Defining the human environment: an omics approach for assessing the environmental risk of cardiovascular disease. *Circ Res* 122:1259–1275
- Schmidt P, Hartung J, Rath J, Piepho H-P (2019) Estimating broad-sense heritability with unbalanced data from agricultural cultivar trials. *Crop Sci* 59:525–536
- Schmidt P, Möhring J, Koch RJ, Piepho H-P (2018) More, larger, simpler: How comparable are on-farm and on-station trials for cultivar evaluation? *Crop Sci* 58:1508–1518
- Shakoor N, Lee S, Mockler TC (2017) High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. *Curr Opin Plant Biol* 38:184–192
- Sorensen D, Gianola D (2002) Likelihood, Bayesian, and MCMC methods in quantitative genetics, 1st edn. Springer Science & Business Media, New York
- Sukumaran S, Crossa J, Jarquín D, Reynolds M (2017) Pedigree-based prediction models with genotype \times environment interaction in multienvironment trials of CIMMYT wheat. *Crop Sci* 57:1865–1880
- Teixeira AP, Dias JML, Carinhas N et al (2011) Cell functional environmentalomics: unravelling the function of environmental factors. *BMC Syst Biol* 5:92
- van Eeuwijk FA, Bustos-Korts D, Millet EJ et al (2018) Modelling strategies for assessing and increasing the effectiveness of new phenotyping techniques in plant breeding. *Plant Sci* 282:23–39
- van Eeuwijk FA, Bustos-Korts DV, Malosetti M (2016) What should students in plant breeding know about the statistical aspects of genotype \times environment interactions? *Crop Sci* 56:2119–2140
- van Ginkel M, Ortiz R (2017) Cross the best with the best, and select the best: HELP in breeding selfing crops. *Crop Sci* 58:17–30
- Varona L, Munilla S, Mouresan EF et al (2015) A Bayesian model for the analysis of transgenerational epigenetic variation. *G3 Genes, Genomes, Genet* 5:477–485
- Velazco JG, Rodríguez-Álvarez MX, Boer MP et al (2017) Modelling spatial trends in sorghum breeding field trials using a two-dimensional P-spline mixed model. *Theor Appl Genet* 130:1375–1392
- Vilhena DA, Antonelli A (2015) A network approach for identifying and delimiting biogeographical regions. *Nat Commun* 6:6848
- Voss-Fels KP, Cooper M, Hayes BJ (2019) Accelerating crop genetic gains with genomic selection. *Theor Appl Genet* 132:669–686
- Wolfert S, Ge L, Verdouw C, Bogaardt M-J (2017) Big data in smart farming—a review. *Agric Syst* 153:69–80
- Xu Y (2016) Envirotyping for deciphering environmental impacts on crop plants. *Theor Appl Genet* 129:653–673
- Xue J, Su B (2017) Significant remote sensing vegetation indices: A review of developments and applications. *J Sensors* 2017:1–17
- Yan W, Hunt LA, Sheng Q, Szlavnics Z (2000) Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Sci* 40:597–605

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.