EnvRtype: a software to interplay enviromics and quantitative genomics in agriculture

Germano Costa-Neto*, Giovanni Galli*, Humberto Fanelli*, José Crossa†, Roberto Fritsche-Neto*

* Department of Genetics, "Luiz de Queiroz" Agriculture College, University of São Paulo, São Paulo, Brazil.

† Biometrics and Statistics Unit, Genetic Resources Program, and Global Wheat Program, International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Mexico.

To whom correspondence should be addressed:

Germano Costa Neto: germano.cneto@usp.br

Running Title: Envirotyping Software for Genomics Studies

Keywords

G×E: genotype × environment interaction;

EnvRtype: Envirotyping and Enviromics in R;

BGGE: Bayesian Genomic Genotype × Environment Interaction;

ABSTRACT

Envirotyping is a core of techniques 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), but also for gathering environmental covariables as quantitative descriptors for predictive purposes (W_matrix function); (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 ecophysiology knowledge in genomic best unbiased predictions (GBLUP) and emulating reaction norm effects (get_kernel function). We highlight literature mining concepts in fine-tuning envirotyping parameters for each plant species and target growing environment. We show that envirotyping for predictive breeding not only collects raw data, but processes it in an ecophysiology-smart way. Examples of its use for creating global-scale envirotyping networks and the integration of reaction-norm modeling in GP are also outlined. We conclude that EnvRtype provides a cost-effective envirotyping pipeline capable of providing good quality enviromic data for a diverse set of genomic-based studies, especially for increasing accuracy in GP across multiple environments.

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 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 as a consequence of 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 due to the use of envirotyping-based data, i.e., a core of techniques to collect, process, type, and integrate environmental information in genetic-informed studies (Costa-Neto et al., 2020a).

Over the last ten years, envirotyping has been incorporated into whole-genome prediction (GP) 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 ecophysiological enriched envirotyping data has

led to outstanding results in model 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., 2020a). Combined with phenotyping and genotyping data, the use of envirotyping data may leverage molecular breeding strategies to understand historical trends and cope with future environmental change scenarios (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 (Bustos-Korts et al., 2019; Galli et al., 2020).

Despite the advance in the development of hypothesis 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 end 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 envirotype descriptors. Consequently, workflow to interplay enviromics (pool of environmental types, abbreviated as envirotypes) and genomic analysis is lacking, especially for GP conditions for multi-environment testing (MET) where G×E is the main concern hampering the model's accuracy.

In this study, we introduce EnvRtype, a novel R package used to integrate macro-environmental factors in many fields of plant, animal or ecological science. We approach basic eco-physiological concepts underlying the collection and processing of raw-environmental data in a biological and statistical manner. Then we present the functions for implementing remote data

collection and basic processing, as well as its applications for deriving quantitative and qualitative descriptors of relatedness. Finally, we present a comprehensive view of how envirome-based data can be incorporated into GP for predictive purposes across diverse environments. We highlight the use of different envirotyping levels to discover descriptors of environmental similarity, using crop species to exemplify the concepts.

METHODS

Envirotyping Pipeline

EnvRtype is a R package created for handling envirotyping by ecophysiology concepts in quantitative genetics and genomics for multiple environments. This means that envirotyping is not only a collection of raw environmental data that are used for exploratory or predictive processes, but a pipeline based on the collection of raw data and the processing of these data 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 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. 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, and 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 to find a viable alternative for expensive in-field environmental sensing equipment. More details about the theoretical basis of environmental sensing and the module itself are given in the section “Module 1: Remote Environmental Sensing”.

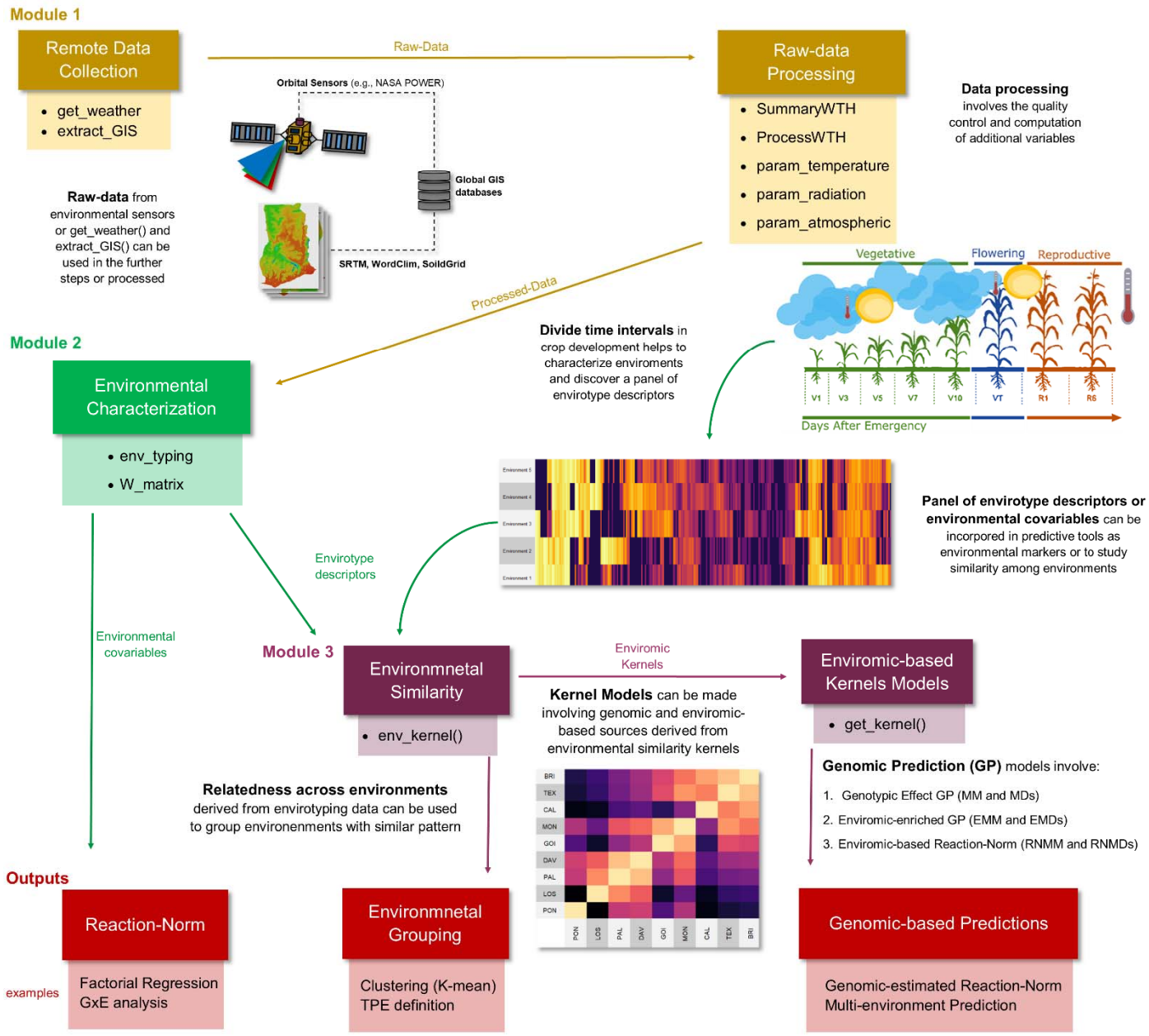


Figure 1. Workflow of the envirotyping pipeline implemented using EnvRtype in R. Yellow, green, wine, and red colored boxes denote the steps related to Modules 1, 2, 3 and the examples of outputs from EnvRtype. Straight arrows indicate the flux of the envirotyping pipeline passing by each Module. Curve arrow is a representation of a process between Modules 1 and 2 in which how field-growing conditions can be described as a panel of envirotype descriptors from each environmental factor processed and organized in the Module 2.

The processed environmental information can then be used for many purposes. In Module 2, we designed tools for environmental profiling (characterization of 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., aimed at discovering 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 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 (wine colors in 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 exploit it in terms of increasing the predictive ability of GP implemented in the BGGE package. About the last, more details are given in the section “Module 3: Environmental Similarity and Kernels for GP”. Below we give some theoretical details about each module and a description of the functions used to implement it.

Software

The R package EnvRtype is available at https://github.com/allogamous/EnvRtype [verified 19 October, 2020]). More details about graphical plots and additional codes can also be found in this Git Hub webpage. Typing the following command in R will automatically install the package (**BOX 1**):

BOX 1: Install EnvRtype
> install.packages(‘devtools’) > devtools::install_github(‘allogamous/EnvRtype’) > require(‘EnvRtype’)

MODULE 1: Remote Environmental Sensing

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 given from SRTM 90 m resolution and can be collected from any place between -60 and 60 latitude. 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 2**). Collection of environmental data for Nairobi, Kenya (latitude 1.367 N, longitude 36.834 E) from 01 march 2015 to 01 April 2015, is performed by:

BOX 2: Practical use of get_weather

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

> head(env.data)
```

A second function is *extract_GIS*, which can be used to collect point values from large raster files from GIS databases. This function has 6 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 name of the environment, which is defined by the **env.id** argument (default is env.id = 'env'). **Latitude** and **Longitude** is given 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 (https://www.worldclim.org/), SoilGrids (https://soilgrids.org/) and EarthMaps (https://earthmap.org/).

A practical use of *extract_GIS* is given below (**BOX 3**). A collection of clay content (g/kg) for Nairobi using a raster file was downloaded from SoilGrids and the function *extract_GIS*. The file 'clay_5_15cm.tiff' is available in the Supplementary Data.

BOX 3: Practical use of extract_GIS
> clay_5_15cm = raster('clay_5_15cm.tiff') # from raster package > clay<-extract_GIS(covraster = clay_5_15cm,name.out = 'clay_5_15cm',env.data = env.data) > head(clay)

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 indispensable for identifying which variables will be summarized. The argument **days.id** indicates

the variable pointing to the time (days), and the default is the *daysFromStart* column from the *get_weather* function. A basic example of this use is given below (**BOX 4**):

BOX 4: Practical use of SummaryWTH
> summaryWTH(env.data = env.data, env.id = ‘env’, days.id = ‘daysFromStart’,statistic = ‘mean’) > summaryWTH(env.data = env.data) # by default

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 done 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 emergence date of the crop in the field, this can also be associated to 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.

Tools for basic data processing

Basic data processing is done by the *processWTH* function. As described for *summaryWTH*, this function can also be used to 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 the impact of atmospheric temperature on crop development. Below we describe these three functions and the ecophysiological temperature concepts underlying their application.

Radiation-related covariables

EnvRtype made available a function called *param_radiation* 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 additional, the global solar radiation incidence (SRAD, in MJ m^2 d^{-1}) is computed as described at the beginning of this section. This last is important in most computations of crop evapotranspiration (Allen et al., 1998) and biomass production (Muchow et al., 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).

Temperature-related covariables

The *param_temperature* function computes additional thermal-related parameters, such as *GDD* and *FRUE* and *T2M_RANGE*. This function has 8 arguments (**env.data**, **Tmax, Tmin**, **Tbase1**, **Tbase2**, **Topt1**, **Topt2** and **merge**). To run this function with data sources other than get_weather, it is indispensable to indicate which columns denote maximum air temperature (Tmax, default is Tmax = 'T2M_MAX') and minimum air temperature (Tmin, default is Tmin = 'T2M_MIN'). The cardinal temperatures must follow the previously described ecophysiology literature (Soltani and Sinclar, 2012, see Appendix 3). Consider the estimations for dry beans at the same location in Nairobi, Kenya (previous box examples). The cardinal temperatures for dry beans were collected from Table 2, which gathers different cardinal temperature for a diverse set of crops.

BOX 5: Practical use of param_temperature for Dry Beans in Nairobi, Kenya
> param_temperature(env.data = env.data,Tbase1 = 8,Tbase2 = 45,Topt1 = 30,Topt2 = 35)

Atmospheric demands

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). Detail about this inputs and equations are given in Appendix Section.

Below 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 = 1,795.

BOX 6: Practical use of param_atmospheric for Dry Bean Crop in Nairobi, Kenya
> param_atmospheric(env.data = env.data,Alt = 1795)

MODULE 2: Macro-Environmental Characterization

Envirotype profiling 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 environment 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). In addition, the relation 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 also can be referred to as environmental type (ET). Finally, the typing of environments can be done by discovering ETs, and the similarity among environments is a consequence of the number of ETs shared between environments.

Before the computation of ETs, a first step is the design based on ecophysiology concepts (e.g., plants' needs for some resource) or summarizing the raw data from the core environments being analyzed. Then, for each ET we computed the frequency of occurrence, which represents the frequency of certain 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 core of 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 in terms of their historical similarity. In this way, we are able not only to 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 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, 9 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 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 dry beans presented in Table 2, the cardinals are typed for Nairobi, Kenya as (**BOX 7**):

BOX 7: Basic use of env_typing for typing temperature in Nairobi, Kenya
> env_typing(env.data = env.data, var.id = ‘T2M’, cardinals = c(0,8,30,35,40))

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 that due to the researcher's expertise, precipitation values less than 10 mm.day^{-1} are insufficient to meet the demand of the crops. Values between 11 mm.day^{-1} and 40 mm.day^{-1} would be considered excellent water conditions, and values of 40 mm.day^{-1} would be considered excessive rainfall. In this scenario, such rainfall values could be associated negatively 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(0,10,40, Inf). For dew point, let's assume data-driven typing (cardinals = NULL) using the previously described quantiles. Taking the same example for Nairobi, Kenya (**BOX 8**):

BOX 8: Basic use of env_typing for more than one variable
> var = c('PRECTOT', 'T2DEW') # variables
> card = list(PRECTOT = c(0,10,40,Inf), T2DEW = NULL) # cardinals and data-driven limits
> Env_typing(env.data = env.data, var.id = var, cardinals = card)

Environmental Covariables with W_matrix

The quality of an environment is measured by the amount of resources available to fulfill the demands of the plants. In an experimental network composed of multi-environment trials (MET), the quality of an environment 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 the use of explicit environmental factors, given the difficulties in obtaining high quality envirotyping data. Here we make available the use of explicit environmental data arranged in a quantitative descriptor such as a covariate matrix (**W**), following the terminology used by Costa-Neto et al. (2020) 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 key 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. For creating environmental similarity kernels, Costa-Neto et al. (2020) suggested using quantile statistics to better describe the distribution of each variable across the experimental network. Thus, this allows a statistical approximation of the ecophysiological importance of the environmental variables 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). Conversely to *env_typing*, the *W_matrix* function was designed to sample quantitative values of each environmental factor 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) denotes mean-centered ($w - \bar{w}$) and scaled ($w - \bar{w}/\sigma$), in which w is the original variable, $\bar{w}$ 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's consider the same use for Nairobi, Kenya, involving only weather variables temperature, rainfall and precipitation, and assuming a quality control of *sd.tol* = 4. The time intervals were settled for each 10-day period (default), and statistic as 'mean' for each variable at each time interval (**BOX 9**).

BOX 9: Basic use of env_typing for more than one variable
> var = c('PRECTOT', 'T2DEW', 'T2M_MAX', 'T2M_MIN') # variables > W = W_matrix(env.data=env.data, var.id=var, statistic='mean', by.interval=TRUE, sd.tol=4) > dim(W)

MODULE 3: Enviromic Similarity and Kernels for Genomic Prediction (GP)

Whole-genome prediction (GP) 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 the use of 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., 2020).

In the third module of the EnvRtype package, we present the tools used to implement this type of modeling approach. Two main functions were designed for this purpose. First, for the construction of environmental relationship kernels, the *env_kernel* is used. The second tool is *get_kernel*, which aims to integrate these kernels into statistical models for GP-based selections across MET. In the next subsections, we describe the kernel methods for modeling envirotype relatedness. Then we present the statistical models that can be built with these kernels.

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:

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

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, diseases evaluations, managements).

The second method is a nonlinear kernel modeled by Gaussian processes, commonly called the Gaussian Kernel or GK. The use of GK for modeling $\boldsymbol{K}_E$ was proposed by Costa-Neto et al (2020a) and is described in a similar way to the approach already used for modeling genomic effects:

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

where *h* is the bandwidth factor (assume as *h* = 1 by default) factor multiplied by the Euclidean Distance $\boldsymbol{D}_{ii'}^2 = \sum_k (w_{ik} - w_{i'k})^2$ 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 pondered the environmental distance (assumed as *Q* = 0.5, equal to the median value of $\boldsymbol{D}_{ii'}^2$. The h can be computed using a marginal function as described by Pérez-Elizalde et al. (2015).

Both methods are implemented by the *env_kernel* function. This function 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 the kernel. 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 that will be integrated in the GP models (see **Genomic-enabled prediction models with envirotyping data**).

A basic use of env_kernel is presented below. Consider the **W** matrix created in Box 7 for the same environment in Nairobi, Kenya. The $\boldsymbol{K_E}$. value using linear covariance and the Gaussian kernel is given as (**BOX 10**):

BOX 10: Basic use of env_kernel
> env_kernel(env.data = W, gaussian = FALSE) > env_kernel(env.data = W, gaussian = TRUE)

Genomic-enabled prediction models with envirotyping data

After the construction of the relationship kernels for environmental relatedness, it is possible to fit a wide 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 naturally 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 4 times faster than BGLR and allows the incorporation of more kernel structures than BMTME. For this reason, EnvRtype has a function called *get_kernel* that aims to organize the genomic or envirotyping-based kernels in different statistical model structures to be run in the BGGE package.

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:

$$\boldsymbol{y} = \boldsymbol{1\mu} + \boldsymbol{X_f\beta} + \sum_{s=1}^{k} \boldsymbol{g_s} + \sum_{r=1}^{l} \boldsymbol{w_r} + \boldsymbol{\varepsilon} \tag{3}$$

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 the 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 ($\boldsymbol{\varepsilon}$). This 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 in several environmental kernels and a subsequent genomic by envirotyping (G×W) reaction-norm kernels. Based in the Equation 6, the theory underpinning the *get_kernel* function is summarized in three

- **Benchmark Genomic Models**. These are baseline models accounting only for genotype-based effects, mostly associated with pedigree-based or genomic realized kinships.

$\sum_{s=1}^{p} \boldsymbol{g_s} \neq 0$ and $\sum_{r=1}^{q} \boldsymbol{w_r} = 0$, 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 not consider any environmental effect.

- **Enviromic-enriched GP Models**. To the MM and MDs models, we added the acronym "E" to denote "enviromic-enriched" for EMM and EMDs models. These models consider $\sum_{s=1}^{p} \boldsymbol{g_s} \neq 0$ and $\sum_{r=1}^{q} \boldsymbol{w_r} \neq 0$, 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 fixed deviation (using $\boldsymbol{X_f \beta}$) plus a random envirotyping-based variation ($\sum_{r=1}^{q} \boldsymbol{w_r}$).
- **Enviromic-based Reaction-Norm GP.** To the MM and MDs models, we added the acronym "RN" for "reaction-norm", resulting in RNMM and RNMDs models, respectively. As described in (ii), the environmental effects can now be modeled as fixed deviations (using $\boldsymbol{X_f \beta}$) plus a random envirotyping-based variation ($\sum_{r=1}^{q} \boldsymbol{w_r}$). However, those RN models consider $\sum_{s=1}^{p} \boldsymbol{g_s} \neq 0$ and $\sum_{r=1}^{q} \boldsymbol{w_r} \neq 0$, 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 a 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.

Enviromic-based Kernel Models for GP 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 (**Y**),

composed of a vector of environment identification, a vector of genotype identification, and a vector of trait values; 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 $p \times p$ genotypes. In the same manner, **K_E** could have the dimension of $q \times q$ environments, but in some cases, the environmental kernels can be built at the phenotypic observation level. This means that for each genotype in each environment, there is a different EC, according to particular phenology stages or envirotyping at plant level. Thus, using the additional argument **dimension_KE** = ‘q’ (default is ‘environment’), the **K_E** may accomplish a kernel with $n \times n$, in which $n = pq$. The basic usage of *get_kernel* is given below (**BOX 11**) and detailed applications in Results section.

BOX 11: Basic usage of get_kernel function

```
> data("maizeYield") # phenotype data (grain yield per environment)

> data("maizeG") # genomic relationship for additive effects

> data(“maizeWTH”) # environmental data

> ECs  = W_matrix(env.data = maizeWTH, var.id = c("FRUE",'PETP','SRAD','T2M_MAX'),statistic = 'mean')

> KE = list(W = env_kernel(env.data = ECs)[[2]])

> KG = list(G=maizeG);

> MM = get_kernel(K_G = KG, Y = maizeYield, model = "MM")

> MDs = get_kernel(K_G = KG, Y = maizeYield, model = "MDs")

> EMM = get_kernel(K_G = KG, K_E = KE, Y = maizeYield, model = "EMM")

> EMDs = get_kernel(K_G = KG, K_E = KE, Y = maizeYield, model = "EMDs")

> RMMM = get_kernel(K_G = KG, K_E = KE, Y = maizeYield, model = "RNMM")

> RNMDs = get_kernel(K_G = KG, K_E = KE, Y = maizeYield, model = "RNMDs")
```

RESULTS

Three sections of results were implemented to give 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. Second, we used a toy data set to demonstrate the creation of different environmental similarities based on different environmental factors. This type of application can be useful for researchers interested in predicting the particular genotypic responses shaped by genomic and enviromic-specific factors across existing experimental trials or for the assembly of virtual scenarios. Finally, we compare the kernel methods for modeling environmental similarities in adjusting GP models. In these last sections, we expect to give some insights to facilitate the use of enviromic data to boost GP for multiple environments.

Global-scale Envirotyping

To illustrate the use of EnvRtype for a global-scale envirotyping study, we consider different time periods (and years) within the summer season at 9 locations around the world (**BOX 12**): Goiânia (Brazil, 16.67 S, 49.25 W, from 15th March 2020 to 04th April 2020), Texcoco (Mexico, 19.25 N, 99.50 W, from 15th May 2019 to 15th June 2019), Brisbane (Australia, 27.47 S, 153.02 E, from 15th September 2018 to 04th October 2018), Montpellier (France, 43.61 N, 3.81 E, from 18th June 2017 to 18th July 2017), Los Baños (Philippines, 14.170 N, 121.431 E, from 18th May 2017 to 18th June 2017), Porto-Novo (Benin, 6.294 N, 2.361 E, from 18th July 2016 to 18th August 2016), Cali (Colombia, 3.261 N, 76.312 W, from 18th November 2017 to 18th December 2017), Palmas (Brazil, 10.168 S, 48.331 W, from 18th December 2017 to 18th January 2018) and Davis (United States, 38.321 N, 121.442 W, from 18th July 2018 to 18th August 2018).In this example, we use 'GOI', 'TEX', 'BRI', 'MON', 'LOS', 'PON', 'CAL', 'PAL' and 'DAV' to identify each location (Figure 2A).

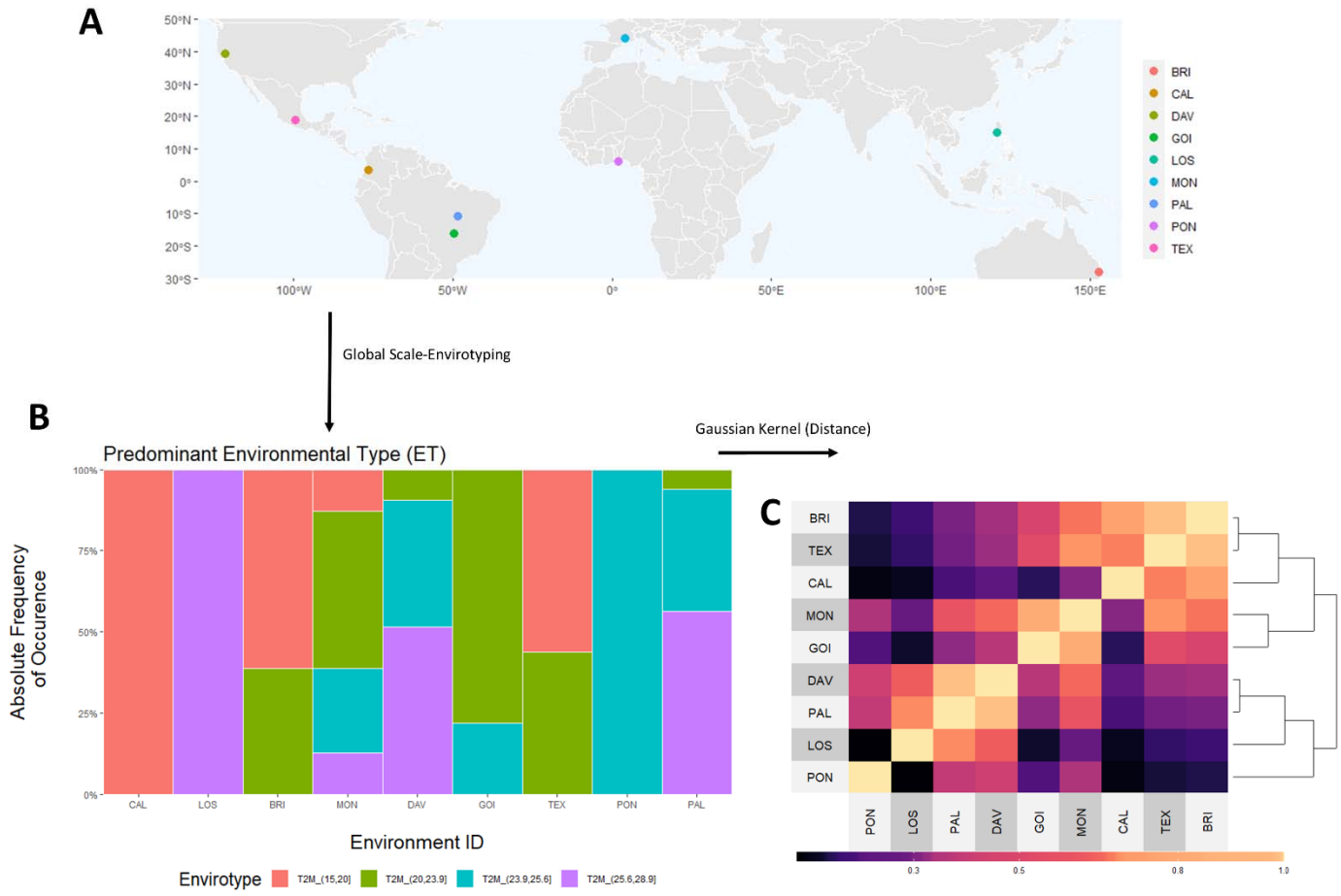


Figure 2. Global scale envirotyping highlights the possible environmental similarity of locations in different continents. **A.** Worldwide geographic positions of 9 locations used as toy examples. **B.** Panel of environmental types (ETs) for average air temperature during a certain month of a certain year in the summer season of each location. **C**. Environmental Similarity matrix based on the ETs and computed using Gaussian kernel.

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 13**), we used the variable 'T2M' (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., 2020a).

BOX 12: 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 13: Discovering ETs and similarity among locations

```
> ET = env_typing(env.data = env.data,env.id = 'env',var.id = "T2M"))

> 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]]
```

It’s 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. 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., 2020b). It could also increase for a better definition of field trial positioning (Resende et al., 2020) and how breeding strategies have impacted crop adaptation in the past (Heinemann et al., 2019).

Panels of environmental similarity and reaction-norm

To illustrate the use of different ECs in GP, we ran a toy example involving a tropical maize data set available in EnvRtype (see **BOX 11**). This data set was included in Souza et al. (2017) and Cuevas et al (2019) and comes from the Helix Seed Company (HEL). However, to facilitate the demonstration of functions, we made available a subset of 150 hybrids per environment. Grain yield data are mean-centered and scaled (*MaizeYield* object). The genotyping relationship for additive effects is based on 52,811 SNPs that are also 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 5 environments (*maizeWTH* object).

We run an example of GP considering the reaction-norm and different levels of envirotyping per environment, which are: (1) envirotyping mean values per environment (entire croplife) and (2) envirotyping mean values for each time interval across crop life. We consider two types of ECs: factor of temperature effect over radiation use efficiency (FRUE) and the difference between rainfall precipitation and crop evapotranspiration (PETP). From equation (3), these ECs were arranged in four kernel structures using the RNMM model:

- Model 1: genomic plus an enviromic kernel built with a single EC (FRUE):

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \boldsymbol{g} + \boldsymbol{FRUE} + \boldsymbol{g} \times \boldsymbol{FRUE} + \boldsymbol{\varepsilon} \quad (4)$$

- Model 2: genomic plus an enviromic kernel built with a single EC (PETP):

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \boldsymbol{g} + \boldsymbol{PETP} + \boldsymbol{g} \times \boldsymbol{PETP} + \boldsymbol{\varepsilon} \quad (5)$$

- Model 3 [joint EC model]: genomic plus an enviromic kernel built with FRUE and PETP (**W**). This is the benchmark reaction norm model (Jarquín et al., 2014), but here it considers only two columns of covariates:

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \boldsymbol{g} + \boldsymbol{W} + \boldsymbol{g} \times \boldsymbol{W} + \boldsymbol{\varepsilon} \quad (6)$$

- Model 4 [multiple EC model]: genomic plus two enviromic kernels, from FRUE and PETP, respectively. Differently from model (6), here the effects of each environmental gradient is modeled separately as:

$$\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \boldsymbol{g} + \boldsymbol{FRUE} + \boldsymbol{PETP} + \boldsymbol{g} \times \boldsymbol{FRUE} + \boldsymbol{g} \times \boldsymbol{PETP} + \boldsymbol{\varepsilon} \quad (7)$$

Box 14 present the codes to implement these envirotyping levels and model structures can be implemented in EnvRtype: From a brief diagnosis of variance components and model fit (correlation between observed and predicted values without missing entries, *r*), it is possible to observe that for the same raw environmental data, each envirotyping level and modeling structure impacts on modeling the phenotypic variation. When envirotyping is done by environment, there is poor fitness (from 0.48 in Model 1 to 0.57 in Model 4) in relation to the models with envirotyping levels by time intervals (from 0.92 in Model 2 to 0.97 in Model 1). This also reflects how the different envirotyping levels impact on the understanding of which ECs explain the phenotypic variation across field trials.

BOX 14: Envirotyping levels and model structures for GP with ECs

```
### 1- Environmental Covariables (ECs)

> wFRUE1 = W_matrix(env.data = maizeWTH, var.id = 'FRUE',statistic = 'mean')

> wPETP1 = W_matrix(env.data = maizeWTH, var.id = 'PETP',statistic = 'mean')

> wJoint1 = W_matrix(env.data = maizeWTH, var.id = c("FRUE",'PETP'),statistic = 'mean')

### 2- Kernels

> K1 = env_kernel(env.data = wFRUE1)[[2]]

> K2 = env_kernel(env.data = wPETP1)[[2]]

> K3 = env_kernel(env.data = wJoint1)[[2]]

> K4 = list(FRUE = K1,PETP = K2);

> K1 = list(FRUE=K1); K2 = list(PETP=K2); K3 = list(Joint = K3); KG = list(G=maizeG);

### 3- Obtain Kernel Models

> M1 = get_kernel(K_G = KG, K_E = K1; Y = maizeYield, model = "RNMM")

> M2 = get_kernel(K_G = KG, K_E = K2, ,Y = maizeYield, model = "RNMM")

> M3 = get_kernel(K_G = KG, K_E = K3, Y = maizeYield, model = "RNMM")

> M4 = get_kernel(K_G = KG, K_E = K4, Y = maizeYield, model = "RNMM")

### 4- Genomic Prediction

> require(BGGE)

> fit1 = BGGE(y = maizeYield$value, K = K1, ne = table(maizeYield$env))

> fit2 = BGGE(y = maizeYield$value, K = K2, ne = table(maizeYield$env))

> fit3 = BGGE(y = maizeYield$value, K = K3, ne = table(maizeYield$env))

> fit4 = BGGE(y = maizeYield$value, K = K4, ne = table(maizeYield$env))
```

In addition, the envirotyping level per time interval leads to an increase in genomic variance components and drastically reduces residual variation. For Model 4 with the envirotyping level by time interval, it was possible to observe that the PETP effect is a determinant of genomic × enviromic interaction in this experimental network. In addition, those effects were better visualized in the model involving a second covariate (Model 4), because in this sense it is expected to better capture the single effect of PETP free of the inner effect of other covariates. This ECs represents the atmospheric demand over the soil-plant-atmospheric continuum. Thus, for predictive purposes, we suggest also using ECs from soil-water balance (e.g., soil water potential), which can be collected from field-based sensors or estimated using crop growth modeling approaches.

From the toy results demonstrated in this section, it is feasible to conclude that for this trait (grain yield), in this germplasm and experimental network, the increased knowledge about temporal variation of ECs also increased the ecophysiology knowledge of GP for explaining phenotypic variation. Consequently, it can lead to accuracy gains when predicting novel genotypes and novel environments. Obviously, further studies are needed in this sense, but here we introduce this concept as a potential application of EnvRtype for increasing ecophysiology knowledge in GP.

Benefits of the Gaussian kernel for GP with enviromic data

Finally, to illustrate the differences in kernel methods when reproducing environmental similarity, we compare the last results from models 4-7 now using a Gaussian kernel approach. For that, we use the same codes from Box 14, but now the Gaussian argument is gaussian = TRUE.

As expected, differences in modeling structures (Figure) and ECs importance (Table 4) were observed. However, there was an increase in all models in relation to the benchmark linear covariance matrix (Table 3). The models with envirotyping level per environment were the ones that benefitted most from using the Gaussian kernel. Model 1 using envirotyping level per

environment was the most successful model in reducing residual variation and fitness (r = 0.99) when FRUE are modeled from the Gaussian kernel. However, the models with envirotyping level per time-interval still outperformed the envirotyping per environment in adjusting models with more suitable to explain the phenotypic variance from genomic kinships. This is a sign that models with those levels of envirotyping data may also be useful to increase the accuracy of GP for multiple environments under conditions with low genomic heritability. Finally, in comparison with the linear-covariance matrix, the nonlinear Gaussian kernel was more efficient in capturing genomic × enviromic interactions (comparison between Tables 3 and 4). In addition, the importance of FRUE and PETP was better elucidated using the Gaussian kernel, although the internal differences between those two ECs are not different from the differences observed using the linear covariance matrix.

BOX 15: Use of the Gaussian Kernel for modeling enviromic kinships

```
### 2- Kernels

> K1 = env_kernel(env.data = wFRUE1,gaussian=TRUE)[[2]]

> K2 = env_kernel(env.data = wPETP1,gaussian=TRUE)[[2]]

> K3 = env_kernel(env.data = wJoint1,gaussian=TRUE)[[2]])

> K4 = list(FRUE = K1,PETP = K2);

> K1 = list(FRUE=K1); K2 = list(PETP=K2); K3 = list(Joint = K3); KG = list(G=maizeG);
```

CONCLUSION

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 ecophysiology knowledge in envirotyping not only increases the accuracy of statistical models in genomic prediction, but also provides a better explanation of the sources of variation and increases the efficiency of those 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, which depends on their sensibility 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 also showed that global envirotyping networks can be built using remote sensing tools and functions provided in EnvRtype. Other uses of the functions presented here may involve: (1) the creation of multiple environment scenarios for predictive breeding; (2) an enviromic scan to determine which ETs better explain environmental similarities; this information can be used to design better experimental networks and accelerate the screening of genotypes for target environments in which they perform better; (3) analysis of historical trends to quantify the impact of recent climate changes on long-term breeding strategies implemented in target regions; (4) provide climate data to be used as reference for field experiments (e.g. high-throughput phenotyping surveys); (5) integrate crop growth models, but as inputs in functions such as W_matrix and env_typing, as well as to predict crop performance across diverse water- or nitrogen-management levels, which may also involve the use of the collected data from get_weather and other sources.

APPENDIX

Appendix 1. What is 'environment' and how to remotely collect environmental data

What we call "environment" is a unit delimited for a combination of location, planting date, and management, which gathers the fluctuation for a core of environmental factors. Thus, the first step of any envirotyping study is collecting reliable environmental data. However, for most breeding programs around the world, this step is limited by the availability of sensing equipment (e.g., weather stations) installed in the field or in a site nearby. It is important to highlight that some equipment can be expensive or difficult to access for some research groups in certain regions, such as developing countries. For this reason, below we present two justifications for incorporating a remote environmental sensing routine (in-silico) into this package. Then, we present recommendations to enrich the envirotyping platforms in order to collect and organize environmental data that will be useful to breeders' decision making.

Firstly, in order to facilitate the step of collecting environmental data, we decided to include a routine for collecting basic daily weather data through the NASA Power database, which can access information on a daily basis anywhere on Earth. This database was integrated using the tools provided by the nasapower R package. In addition, we integrated the raster R package to support the download of climatic data (from the WorldClim database) and SRTM (Shuttle Radar Topography Mission, which provides information about elevation). The information from both databases are freely available and can be downloaded using geographical coordinates (latitude and longitude, given in decimal degrees, both in WGS84 format) for a specific window of time (e.g., from sowing to harvest).

Secondly, the processing of the collected environmental data requires some expertise in fields such as agrometeorology, soil physics and ecophysiology. This is because, to be effective in explaining crop adaptation, the environmental data must be representative of some envirotype-to-

phenotype dynamic linked to certain ecophysiological knowledge (e.g., air temperature, relative air humidity and solar radiation driving the crops' evapotranspiration and, consequently, the soil-water balance). A direct example of the importance of processing raw envirotyping data into ecophysiogical enriched information is given for the "daily air temperature" variable. This variable can be processed in heat units, heat stress effect on radiation use efficiency and thermal range, which is species-specific. For some traits such as grain yield in maize, the impact of temperature-derived factors differs from the observed in traits such as plant height or flowering time. This dynamic also varies across crop developmental stages, which can be more or less prone to become a stressful factor at certain stages (e.g., in maize, heat at flowering time has a greater impact on grain yield). Below is a detailed description of some of those subroutines.

Appendix 2. Concepts underlying Radiation effects in Plants

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., $SRAD = ALLSKY_S FC_L W_D W - ALLSKY_S FC_S W_D WN$, in MJ m^{-2} d^{-1}.

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 source of environmental variation is important to understand the differences in potential yield observed in genotypes evaluated across diverse environments. Radiation is also important as a

source for regulating the available energy for other biophysical processes, such as evaporation, transpiration and temperature (see subsection Processing Atmospheric Parameters).

Appendix 3. Concepts underlying the Effect of Temperature in Plants

Thermal variables are important for regulating the rates of important biochemical processes within an organism. At 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 plant level, temperature related variables regulate the balance between photosynthesis (gross and net) and respiration in the canopy, impacting on radiation use efficiency (RUE). It also is 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 final yield of the crop, 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.

Table 2 summarizes the cardinal limits of temperature for several species. Those cardinal limits are used to compute growing degree-days (*GDD*) and the effect of temperature on radiation use efficiency (*FRUE*). 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 the adaptation of some species in the target environment. *GDD* is also important for modeling plant-pathogen interactions, because some pests and diseases have temperature-regulated growth.

In this context, dew point (*T2MDEW*) is another agrometeorological factor that is greatly important for crop health. In addition to being related to the evaporation process in the stomata,

this factor determines the establishment of diseases (especially fungus) under the leaf. Finally, the daily temperature range (*T2M_RANGE*) is a factor impacting processes such as floral abortion in crops where the main traits are related to grain production. For more details about the impact of temperature in diverse agricultural crops, please check Luo (2011).

Appendix 4. Concepts underlying the Effect of Atmospheric demands in Plants

The dynamics of water precipitation (rainfall) and water demand (evaporation+plant transpiration) 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. The rate of this process is deeply related to the biomass production of plants, but also to the absorption of nutrients by mass flux in roots. Because of that, water demands are essential for measuring the quality of some growing environments.

Here we used the Pristley-Taylor equation to compute the reference crop evapotranspiration. At this equation, the empirical constant (alpha = α) may range from 1 (at humidity conditions) to 2 (at arid conditions). First, we compute the vapour pressure, determined by: $e_a = RH \times e_s$ (Dingman, 2002), where e_s is the saturation vapour 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 vapour pressure (e_a), we compute the slope of the saturation vapour 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 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 agricultural crops, we encourage the use crop coefficient (K_c, dimensionless) to translate ET_0 in crop-specific evapotranspiration. This K_c is computed from empirical phenotypic records (crop height, albedo of the soil-crop surface, canopy resistance) combined with in-field sensors (evaporation from soil), or using K_c estimates for each crop specie. Allen et al. (1998) provides a wide number of general K_c values to be used in this sense. For a complete understanding of soil-water dynamics, we suggest the use of pedotransfer functions to derive some hydraulic properties of the soil, such as infiltration rate and water retention parameters. This can be done by soil samples or from remotely collected data from SoilGrids using *extract_GIS*.

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.*

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.

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

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.

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., 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.

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.

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.

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.

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

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., M. Malosetti, B. Ababaei, G. Touzy, A. Mini *et al.*, 2019 Using crop growth model stress covariates and AMMI decomposition to better predict genotype-by-environment interactions. Theor. Appl. Genet. 132: 3399–3411.

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.

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

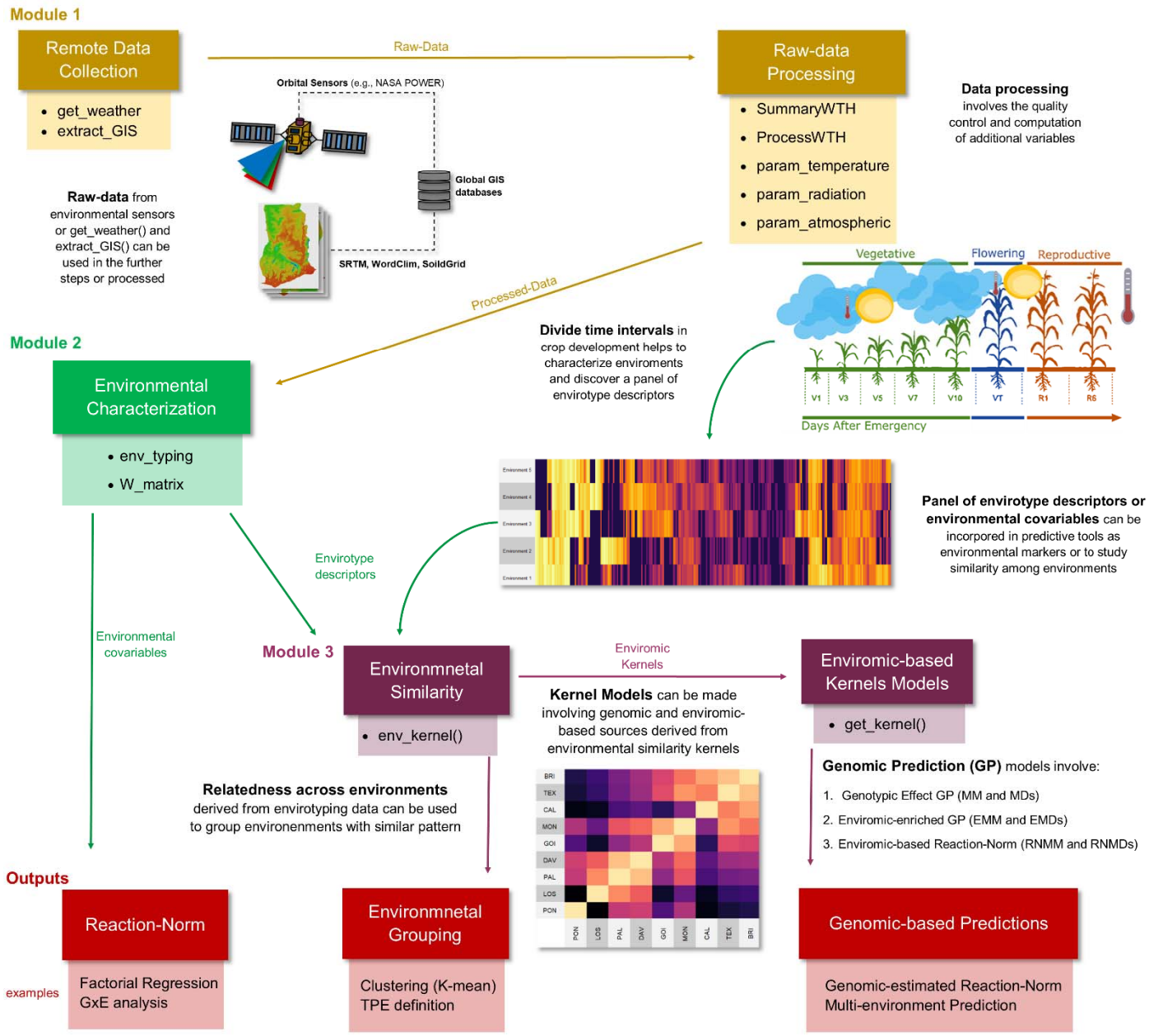


Figure 1. Workflow of the envirotyping pipeline implemented using EnvRtype in R. Yellow, green, wine, and red colored boxes denote the steps related to Modules 1, 2, 3 and the examples of outputs from EnvRtype. Straight arrows indicate the flux of the envirotyping pipeline passing by each Module. Curve arrow is a representation of a process between Modules 1 and 2 in which how field-growing conditions can be described as a panel of envirotype descriptors from each environmental factor processed and organized in the Module 2.

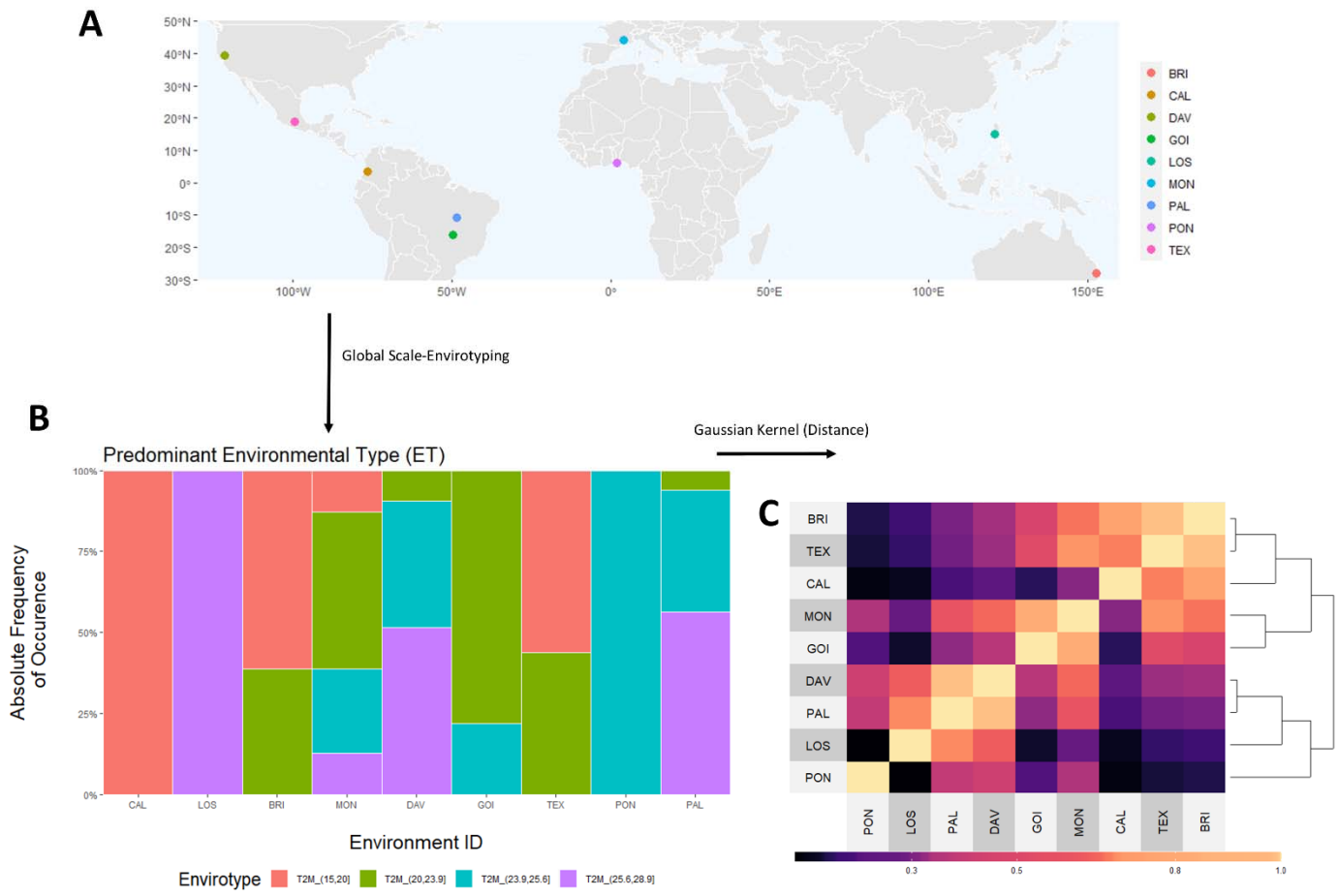


Figure 2. Global scale envirotyping highlights the possible environmental similarity of locations in different continents. **A.** Worldwide geographic positions of 9 locations used as toy examples. **B.** Panel of environmental types (ETs) for average air temperature during a certain month of a certain year in the summer season of each location. **C**. Environmental Similarity matrix based on the ETs and computed using Gaussian kernel.

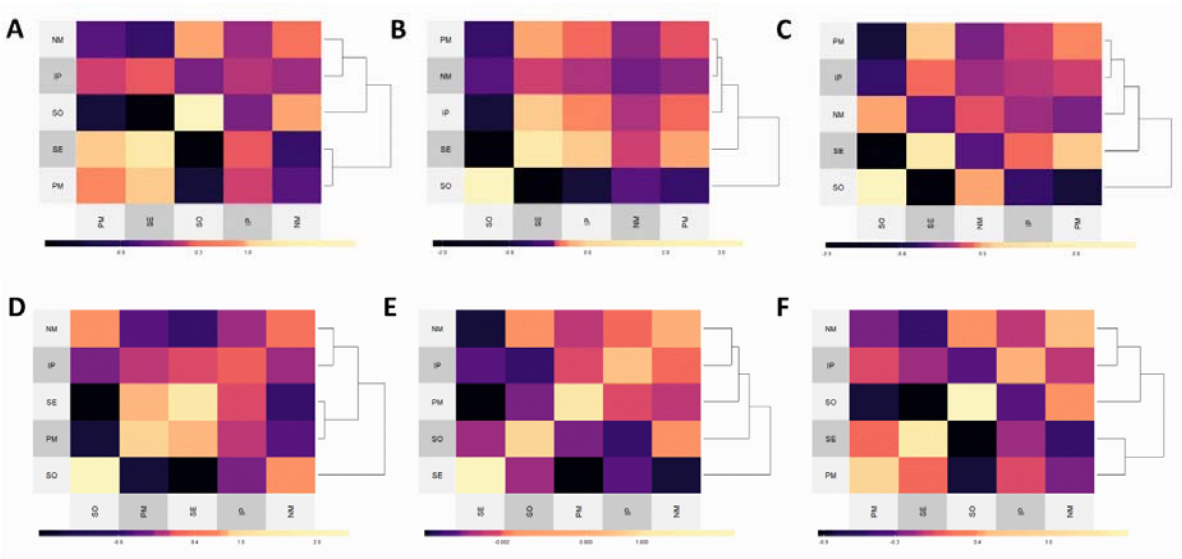


Figure 3. Linear enviromic kernels based on the combination of two environmental covariates (ECs) and two envirotyping levels for 5 locations across an experimental network of tropical maize. Kernel were based on a FRUE variable (impact of temperature on radiation use efficiency) for envirotyping during the entire crop life (**A**) or divided by time intervals (**D**); **B.** PETP variable (deficit of evapotranspiration, mm.day^{-1}) for the entire crop life (**B**) or divided by time intervals (**E**) and a combination of those two variables in a single kernel for the entire crop life (**C**) or by time intervals of crop life (**F**).

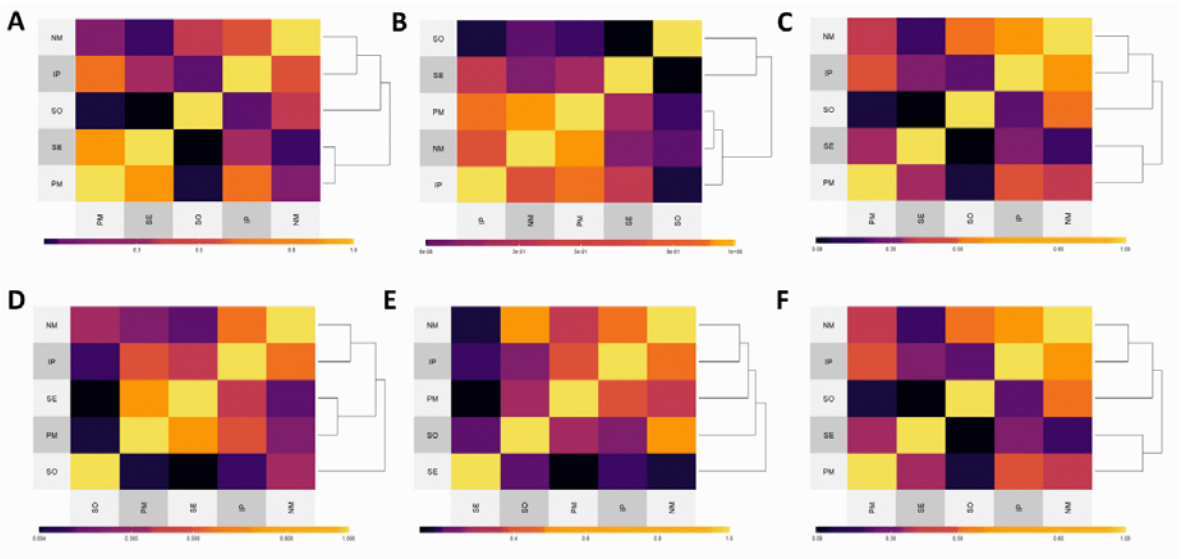


Figure 4. Nonlinear enviromic kernels (Gaussian) based on the combination of two environmental covariates (ECs) and two envirotyping levels for 5 locations across an experimental network of tropical maize. Kernels were based on a FRUE variable (impact of temperature on radiation use efficiency) for envirotyping during the entire crop life (**A**) or divided by time intervals (**D**); **B.** PETP variable (deficit of evapotranspiration, mm.day^{-1}) for the entire crop life (**B**) or divided by time intervals (**E**) and a combination of those two variables into a single kernel for the entire crop life (**C**) or by time intervals of crop life (**F**).

Table 1. Core of environmental factor available using the 'Environmental Sensing Module' of *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}
	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	-
	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] collected from NASA orbital sensors (Stackhouse Jr., 2014); [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).

Table 2. Synthesis of some cardinal limits for the effect of temperature on the phenology development in the main agricultural crops. These estimates were adapted from Soltani and Sinclar (2012), Lago et al (2009), Steinmetz (2004), Buriol et al. (1991), Venkataraman et al. (2007)

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

Table 3. Summary of a variance components and correlation between observed and predicted values (r) from a preliminary reaction-norm study involving two envirotyping levels (per environment, by.interval = FALSE; and per time intervals by environment, by.interval=TRUE), for a RNMM model involving additive genomic effects (G) and different structures for environmental covariates (ECs). The r values were computed using all phenotypic data. Enviromic kinships were build using a linear-covariance matrix (gaussian = FALSE).

Effect per Model	Envirotyping level	
	Environment	**Time Interval × Environment**
Model 1: FRUE		
E = [FRUE]	2.025	15.841
Genomic (G)	0.428	0.482
GxE	0.037	0.074
Residual	0.855	0.152
Total	3.345	16.548
r =	0.48	0.97
Model 2: PETP		
E = [PETP]	3.058	4.093
Genomic (G)	0.454	0.481
GxE	0.039	0.093
Residual	0.826	0.221
Total	4.378	4.888
r =	0.52	0.92
Model 3: Joint		
E=[FRUE + PETP]	3.331	5.185
Genomic (G)	0.430	0.511
GxE	0.046	0.093
Residual	0.819	0.213
Total	4.626	6.002
r =	0.54	0.93
Model 4: FRUE+PETP		
RUE	1.838	1.247
PETP	2.661	5.148
Genomic (G)	0.403	0.502
GxRUE	0.048	0.045
GxPETP	0.050	0.131
Residual	0.811	0.186
Total	5.811	7.260
r =	0.57	0.95

Table 4. Summary of a variance components and correlation between observed and predicted values (r) from a preliminary reaction-norm study involving two envirotyping levels (per environment, by.interval = FALSE; and per time intervals by environment, by.interval=TRUE), for a RNMM model involving additive genomic effects (G) and different structures for environmental covariates (ECs). The r values were computed using all phenotypic data. Enviromic kinships were build using Gaussian Kernel (gaussian = TRUE).

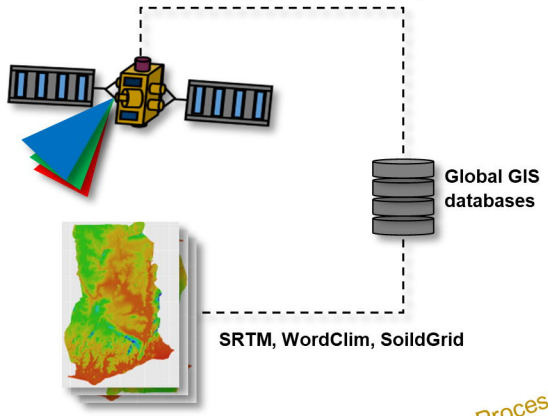
Effect per Model	Envirotyping level	
	Environment	**Time Interval × Environment**
Model 1: FRUE		
E = [FRUE]	31.974	10.131
Genomic (G)	0.470	0.530
GxE	0.104	0.182
Residual	0.076	0.161
Total	32.624	11.004
r =	0.99	0.96
Model 2: PETP		
E = [PETP]	9.104	6.059
Genomic (G)	0.446	0.517
GxE	0.124	0.155
Residual	0.230	0.195
Total	9.904	6.927
r =	0.93	0.94
Model 3: Joint		
E=[FRUE + PETP]	12.549	7.877
Genomic (G)	0.484	0.518
GxE	0.149	0.144
Residual	0.144	0.185
Total	13.326	8.724
r =	0.97	0.95
Model 4: FRUE+PETP		
RUE	3.172	4.860
PETP	5.500	4.203
Genomic (G)	0.538	0.553
GxRUE	0.134	0.127
GxPETP	0.133	0.139
Residual	0.158	0.151
Total	9.635	10.033
r =	0.96	0.97

Module 1

Remote Data Collection

- get_weather
- extract_GIS

Raw-Data



Raw-data from environmental sensors or get_weather() and extract_GIS() can be used in the further steps or processed

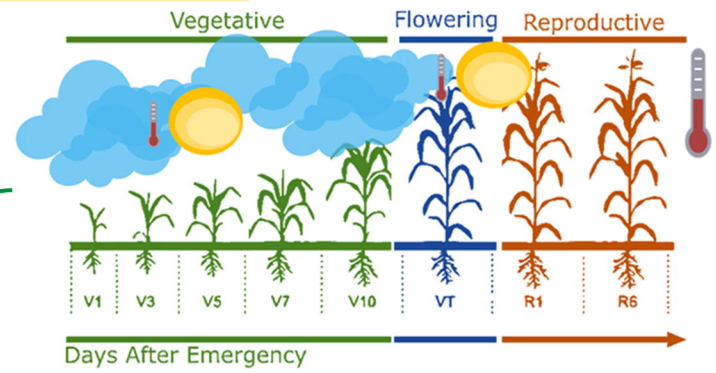
Raw-data Processing

- SummaryWTH
- ProcessWTH
- param_temperature
- param_radiation
- param_atmospheric

Data processing involves the quality control and computation of additional variables

Processed-Data

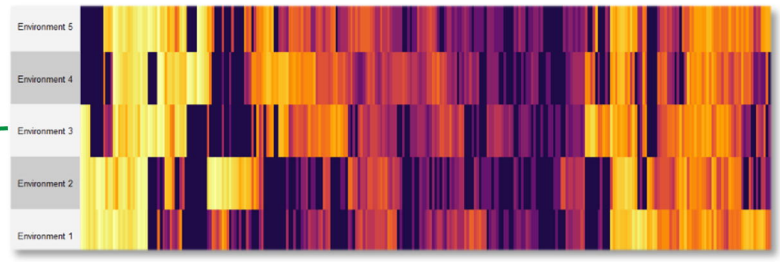
Divide time intervals in crop development helps to characterize enviroments and discover a panel of envirotype descriptors



Module 2

Environmental Characterization

- env_typing
- W_matrix



Panel of envirotype descriptors or environmental covariables can be incorpored in predictive tools as environmental markers or to study similarity among environments

Envirotype descriptors

Environmental covariables

Module 3

Environmnetal Similarity

- env_kernel()

Enviromic Kernels

Kernel Models can be made involving genomic and enviromic-based sources derived from environmental similarity kernels

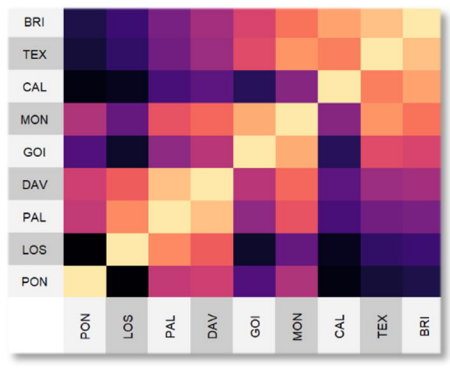
Enviromic-based Kernels Models

- get_kernel()

Genomic Prediction (GP) models involve:

1. Genotypic Effect GP (MM and MDs)
2. Enviromic-enriched GP (EMM and EMDs)
3. Enviromic-based Reaction-Norm (RNMM and RNMDs)

Relatedness across environments derived from envirotyping data can be used to group environenments with similar pattern



Outputs

Reaction-Norm

examples

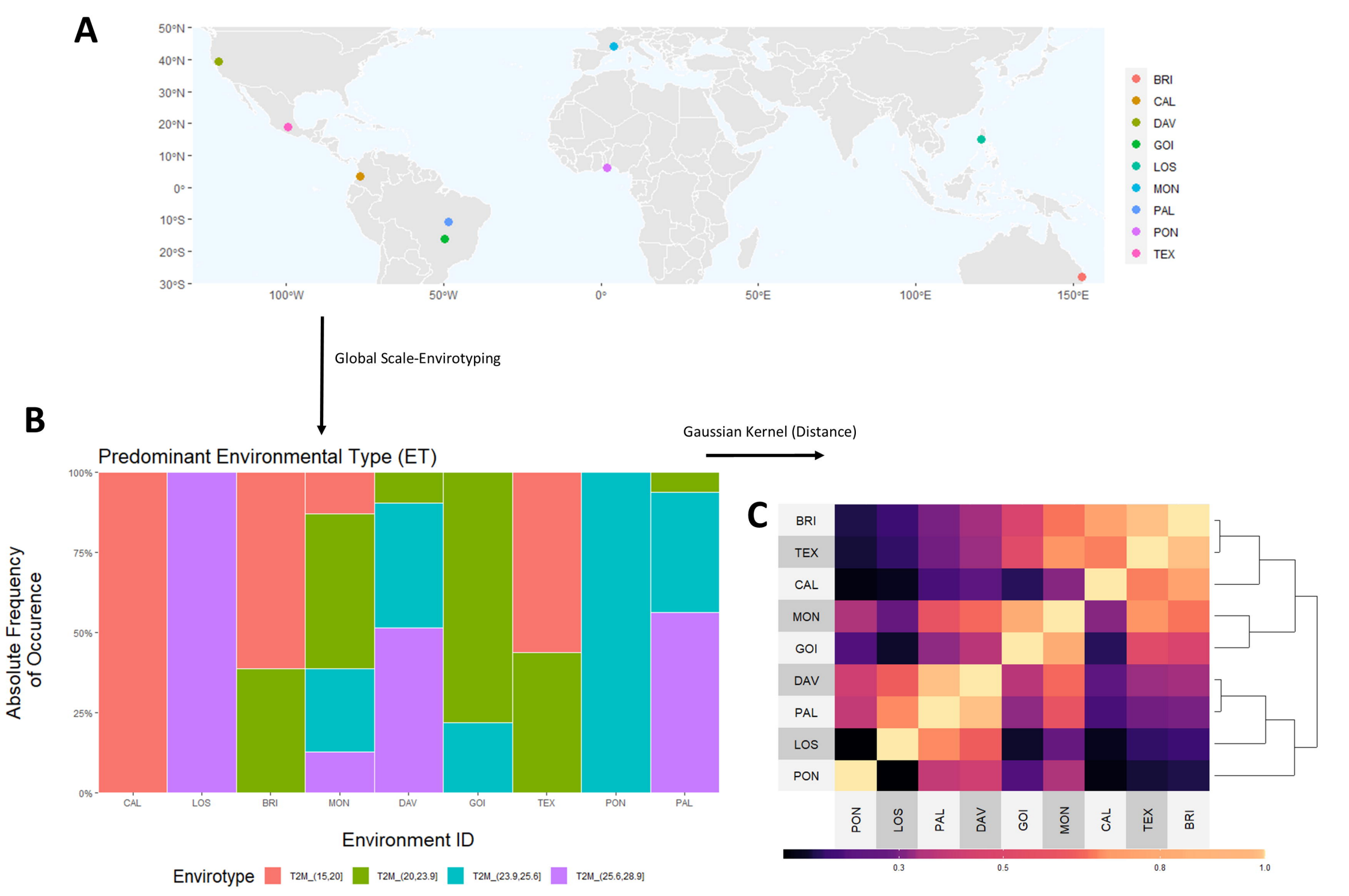
Factorial Regression
GxE analysis

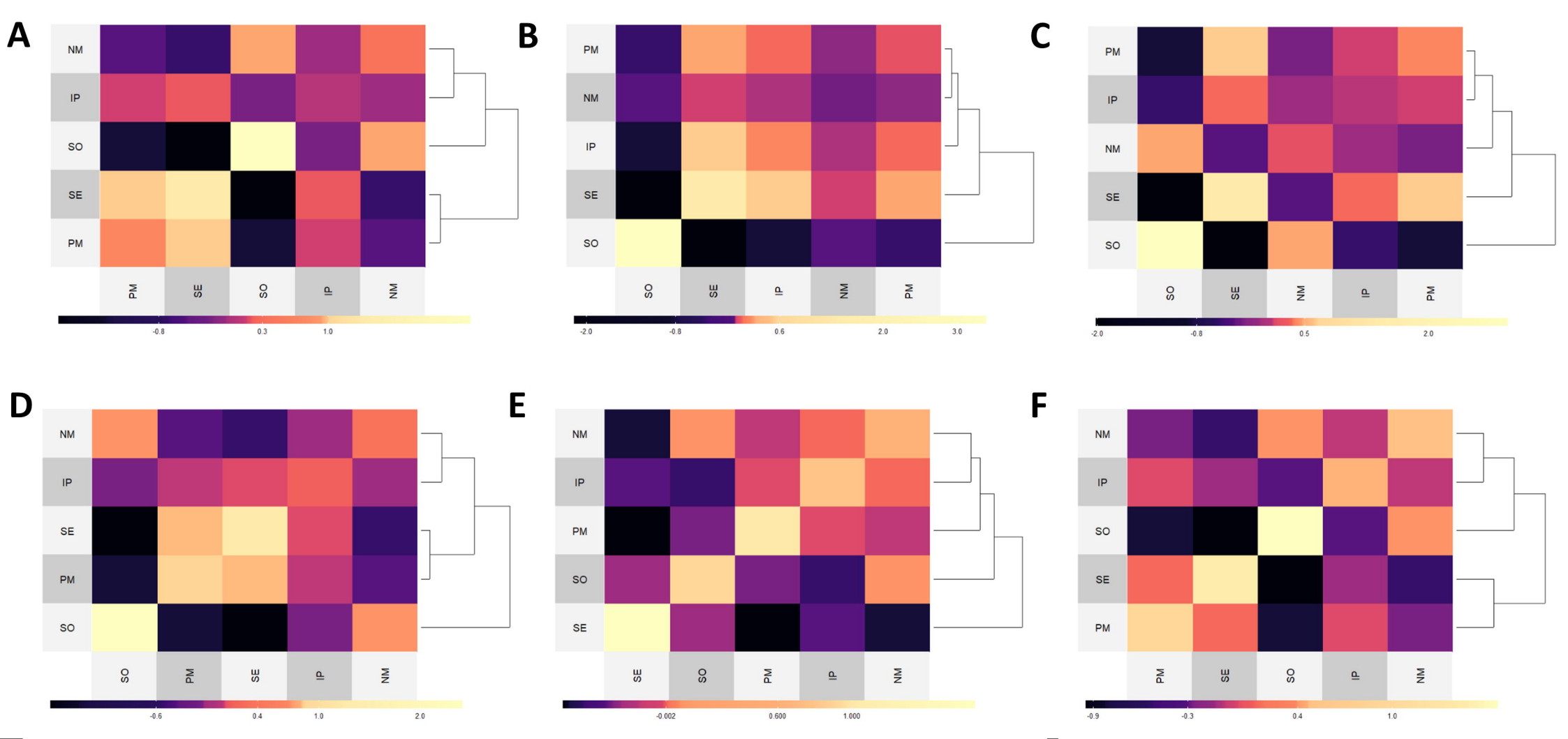
Environmnetal Grouping

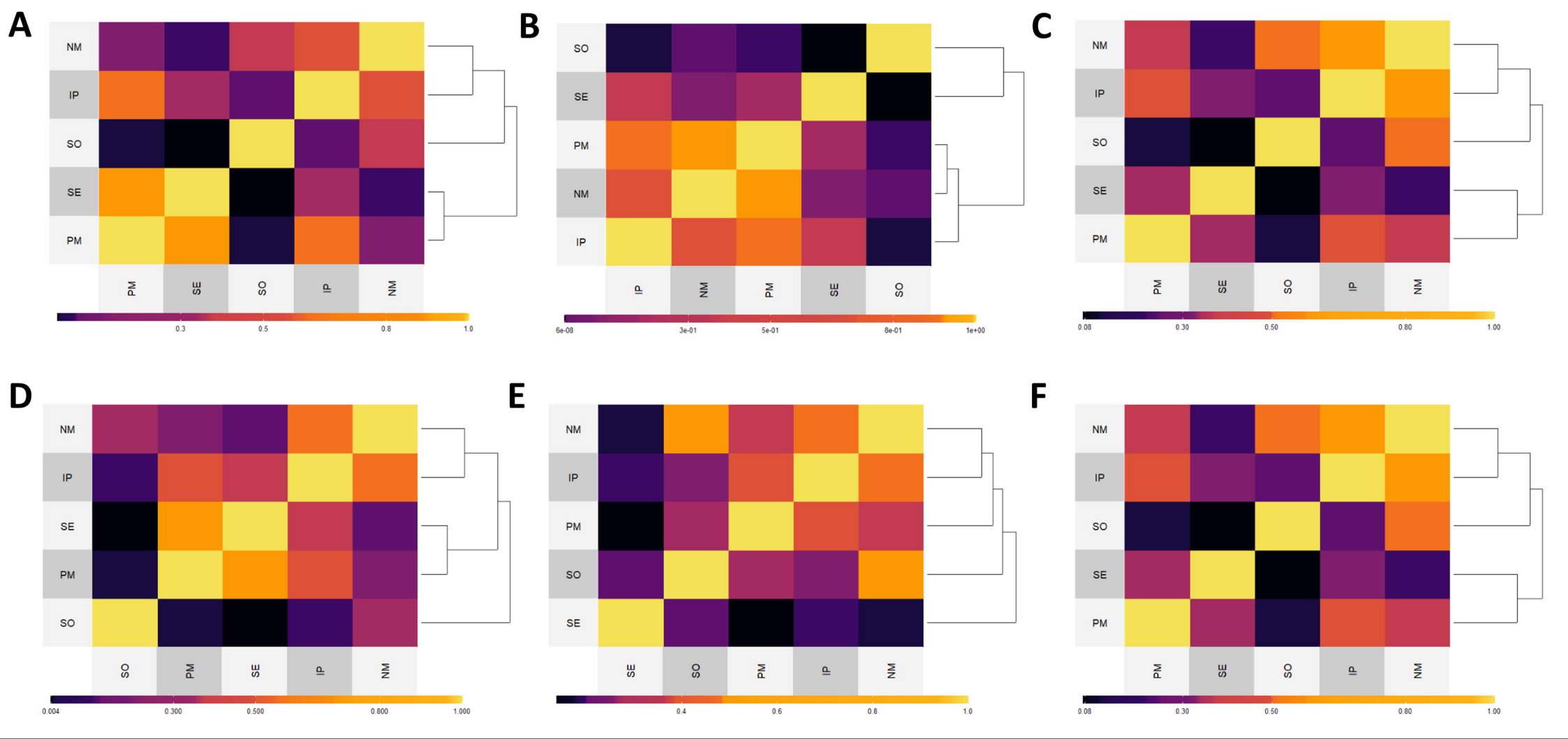
Clustering (K-mean)
TPE definition

Genomic-based Predictions

Genomic-estimated Reaction-Norm
Multi-environment Prediction


A
50°N
40°N
30°N
20°N
10°N
0°
10°S
20°S
30°S
100°W
50°W
0°
50°E
100°E
150°E
BRI
CAL
DAV
GOI
LOS
MON
PAL
PON
TEX
Global Scale-Envirotyping
B
Predominant Environmental Type (ET)
Absolute Frequency of Occurence
100%
75%
50%
25%
0%
CAL
LOS
BRI
MON
DAV
GOI
TEX
PON
PAL
Environment ID
Envirotype
T2M_(15,20]
T2M_(20,23.9]
T2M_(23.9,25.6]
T2M_(25.6,28.9]
Gaussian Kernel (Distance)
C
BRI
TEX
CAL
MON
GOI
DAV
PAL
LOS
PON
0.3
0.5
0.8
1.0


A
NM
IP
SO
SE
PM
PM SE SO IP NM
-0.8 0.3 1.0
B
PM
NM
IP
SE
SO
SO SE IP NM PM
-2.0 -0.8 0.6 2.0 3.0
C
PM
IP
NM
SE
SO
SO SE NM IP PM
-2.0 -0.8 0.5 2.0
D
NM
IP
SE
PM
SO
SO PM SE IP NM
-0.6 0.4 1.0 2.0
E
NM
IP
PM
SO
SE
SE SO PM IP NM
-0.002 0.600 1.000
F
NM
IP
SO
SE
PM
PM SE SO IP NM
-0.9 -0.3 0.4 1.0


A
NM
IP
SO
SE
PM
PM SE SO IP NM
0.3 0.5 0.8 1.0
B
SO
SE
PM
NM
IP
IP NM PM SE SO
6e-08 3e-01 5e-01 8e-01 1e+00
C
NM
IP
SO
SE
PM
PM SE SO IP NM
0.08 0.30 0.50 0.80 1.00
D
NM
IP
SE
PM
SO
SO PM SE IP NM
0.004 0.300 0.500 0.800 1.000
E
NM
IP
PM
SO
SE
SE SO PM IP NM
0.4 0.6 0.8 1.0
F
NM
IP
SO
SE
PM
PM SE SO IP NM
0.08 0.30 0.50 0.80 1.00