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

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

[1] Department of Genetics, "Luiz de Queiroz" Agriculture College, University of São Paulo, São Paulo, Brazil. E-mail: germano.cneto@usp.br (http://orcid.org/0000-0003-1137-6786);

[2] Biometrics and Statistics Unit, Genetic Resources Program, and Global Wheat Program, International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Mexico. E-mail: j.crossa@cgiar.org

Running Title: Envirotyping Software for Genomics

Key-words

G×E: genotype × environment interaction;

EnvRtype: Envirotyping and Enviromics in R;

BGGE: Bayesian Genomic Genotype × Environment Interaction;

Correponding Author: Germano Costa-Neto (germano.cneto@usp.br)

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 an 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 level; (2) environmental characterization by typing environments and profiling descriptors of environmental quality (EnvTyping 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 (EnvKernel function) in whole-genome prediction (GP), aiming to increase ecophysiology knowledge in genomic-best unbiased predictions (GBLUP) and emulate 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 is not only collect raw data, but process it into a ecophysiology-smart way. Examples of use for creating global-scale envirotyping networks and the integration of reaction-norm modeling in GP is also outlined. We conclude that EnvRtype provides a cost-effective envirotyping pipeline capable to provide good-quality enviromic data for a diverse set of genomic-based studies, especially for increasing accuracy in GP across multiple environments.

INTRODUCTION

Quantitative Genetics theory divides the phenotypic variation (P) into a genetic (G) and non-genetic source of variation (E). This last may involve micro-environmental effects that can be controlled by good experimental designs and phenotype correction strategies (e.g., Resende and Duarte, 2007; Galli et al., 2018). Conversely, most of the 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 in 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., core of techniques to collect, process, typing, and integrate the environmental information in genetic-informed studies (Costa-Neto et al., 2020a).

Over the last ten years, envirotyping has been incorporated in 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 ECs due to whole-genome factorial regressions (Heslot et al., 2014; Ly et al., 2018; Millet et al., 2019), but also using those ECs to create envirotyping-based kinships (Jarquín et al., 2014; Morais-Junior et al., 2018; Costa-Neto et al., 2020a). This last has the advantage to be faster in approaching putative environmental similarities that may drive a large amount of phenotypic variation, while the first has the advantage of allowing a deeper

61 understanding of what ECs may better explain the phenotypic plasticity of organisms. The
62 integration of ecophysiological enriched envirotyping data has led to outstanding results in
63 model crops such as maize, due to the use of Crop Growth Models (Cooper et al., 2016; Messina et
64 al., 2018) and Deep Kernel approaches (Costa-Neto et al., 2020a). Combined with phenotyping
65 and genotyping data, the use of envirotyping data may leverage the molecular breeding strategies
66 to understand historical trends and cope with future environmental change scenarios (Gillberg et
67 al., 2019; de los Campos et al., 2020).

68 Despite the advance in the development of theories supporting the inclusion of
69 envirotyping data in GP, there is difficult for most breeders to deal with the interplay between
70 envirotyping, ecophysiology, and genetics. For example, to use molecular data as a sign of the
71 allelic diversity of target germplasm or population, and then use it to build a genomic-
72 relationship matrix (GRM), many researches have been made to explored to associate the raw-
73 data into concepts and theories underlying quantitative genetics (e.g, Fisher's Infinitesimal
74 Model). Genotyping pipelines based on bioinformatics were successfully developed to translate
75 biochemical outputs collected from plant tissues onto biological significant markers of DNA
76 polymorphisms, e.g., genotyping-by-sequence (GBS, Elshire et al.,2011). To the best of our
77 knowledge, there is no publicly available user-friendly software to implement envirotyping
78 pipelines to translate raw-environmental data into a useful matrix of envirotype descriptors. As
79 consequence, is lacking a workflow to interplay enviromics (pool of environmental types,
80 abbreviated as envirotypes) and genomics analysis, especially for conditions of GP for multi-
81 environment testing (MET) where G×E is the main concern hampering the model's accuracy.

82 In this study, we introduce EnvRtype, a novel R package to integrate macro-environmental
83 factors in many fields of plant, animal or ecology science. We approached basic eco-physiological
84 concepts underlying the collection and processing of raw-environmental data into a biological

and statistical manner. Then, we present the functions for implementing remote data collection and basic processing, and also its applications in deriving quantitative and qualitative descriptors of relatedness. Finally, we present a comprehensive view of how envirome-based data can be incorporated in 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 the envirotyping is not only a collection of raw environmental data and their use for exploratory or predictive process, but a pipeline based from the collection of raw-data to the processing of this data in an ecophysiology-manner that make sense for describing the development of the organism in target environment. Here we consider *enviromic* as the large-scale envirotyping of a theoretical population of environments for a target specie or germplasm (the so-called envirome). It also may denote the core of possible growing conditions and technological inputs to create different productivity levels.

The envirotyping pipeline implemented by EnvRtype software are divided in three modules, in which will brief described above and detailed in the next sections (Fig1).

Module 1 (yellow toolboxes in Fig 1) starts for collecting raw-environmental data. Data collection may involve existing experimental trials (single trials sampling) 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 or air temperature on radiation use efficiency. Thus, englobes a toolbox with "Remote Data Collection" and "Data Processing" steps, both designed to assist researchers with lower budgets to fund in-field environmental sensing equipment. More detail about the theoretical basis of environmental sensing and the module itself is given in the section named "Module 1: Remote Environmental Sensing".

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

(1) discovering environmental types (envirotypes, hereafter abbreviated as ETs) and their frequency of occurrence at each growing environment (location, planting date, year). From the ET-discovering step, it is possible to create environmental profiles and group environments with the same ET pattern. It is also useful for running exploratory analysis, such as to discovery 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). This ECs can be used for many purposes, since basic interpretation of G×E to estimate gene-environment interactions. At the end of this process, a matrix of ECs (**W**) is created and integrated with tools from Module 3. More detail 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, i.e., the hereafter referred as

131 envirotype-informed GP The Module 3 (wine colors in Figure 1) aims to provide tools to compute
132 environmental similarity using correlations or Euclidean distances across different trials realized
133 from ECs. Thus, we develop a function to integrate this enviromic sources in GP as an additional
134 source of variation to bridge the gap between genomic and phenotypic variation. For that, we
135 provide at least four different structures, into a flexible platform to integrate multiple genomic
136 and enviromic kinships.

137 Figure 1 show some possible outputs of EnvRtype package (in red toolbox colors), in
138 which **W** can be used to interpret G×E (e.g., factorial regression, see) or exploit it in terms of
139 increasing the predictive ability of GP implemented in BGGE package. About this last, more detail
140 is given in the section "Module 3: Environmental Similarity and Kernels for GP". Below we gave
141 some theoretical detail about each module and description of the functions used to implement it.

142 **Software**

143 The R package EnvRtype is available at https://github.com/allogamous/EnvRtype
144 [verified 18 July. 2020]). More detail about graphical plots and additional codes can also be found
145 in this Git Hub webpage. Typing the following command in R will automatically install the
146 package:

BOX 1: Install EnvRtype

```
> install.packages('devtools')
> devtools::install_github('allogamous/EnvRtype')
> require('EnvRtype')
```

147

MODULE 1: Remote Environmental Sensing

What we call by "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 the collection of 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 near place. It is important to highlight that some equipment can be expensive or difficult to access for some research groups in certain regions, such as development 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 in the decision-making of breeders.

Firstly, in order to facilitate the steps of collection of the environmental data, we decided to insert a routine for collecting basic daily weather data through the Nasa Power database, which can access information on a daily scale anywhere on the globe. This database was integrated using the tools provided by the nasapower R package (Sparks, 2018). In addition, we integrate the raster R package to support the download of climatic data (from the WorldClim database, Fick and Hijmans, 2017) and SRTM (Shuttle Radar Topography Mission, providing information about elevation). The information from both data bases are freely available and can be downloaded using geographical coordinates (Latitude and Longitude, given in decimal degrees, both in WGS84 format) for a specific time window (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. It is because to be really effective in explaining the crops adaptation, the environmental data must be representative of some

171 envirotype-to-phenotype dynamic linked to a certain ecophysiological knowledge (e.g., air
172 temperature, relative air humidity and solar radiation driving the crops evapotranspiration and,
173 consequently, the soil-water balance).

174 A direct example of the importance of processing raw-envirotyping data into
175 ecophysiogical enriched information is given for the variable "daily air temperature". This
176 variable can be processed in heat-units, heat-stress effect on radiation use efficiency and thermal
177 range, which is specie-specific for different crops such as maize, soybean, pinewood etc. For some
178 traits such as grain yield in maize, the impact of those temperature-derived factors differs from
179 the impact observed for traits such as plant height or flowering time. This dynamic has also a
180 variation across the crop development, which can be more or less suitable to become a stressful
181 factor in certain phenological stages (e.g., heat in flowering time in maize has a higher impact on
182 grain yield). Before this ecophysiological processing, some quality control of this data can also be
183 done in order to remove possible outliers. Below is detailed some of those subroutines.

184 **Remote data collection**

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

192 Table 1 shows the names of the outputs of *get_weather* and the *processWTH* (see Tools for
193 basic Processing). All weather information is given in daily scale. Altitude (*ALT*) information is

given from SRTM 90 m resolution and can be collected from any place between -60 and 60 latitude. This information are presented as a *data.frame* class output in R. For a same country, it is possible to create vectors of information to import the data for several environments at the same time.

A practical example of get_weather is given below. A 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',country = 'KEN')
```

. More examples are given in Results section.

A second function is *Extract_GIS*, that can be used to collect point-values from large raster files from GIS databases. This function has 6 the arguments. The argument **`env.data`** indicates the name of the environmental data set (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`** can be given in decimal format as WGS84, 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 data bases 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. A collection of clay content (g/kg) for Nairobi using a raster file 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
> Extract_GIS(covraster = clay_5_15cm,name.out = 'clay_5_15cm',env.data = env.data)
```

Summarizing raw-data

Basic data summary of the outputs from *get_weather* function done by *summaryWTH* function. This function has 10 arguments (**env.data, id.names, env.id, days.id, var.id, statistic, probs, by.interval, time.window,names.window**). The common arguments with *Extract_GIS* have also the same utility already described. Other identification columns (year, location, management, responsible researcher etc) may be indicated in **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 names of variables presented in Table 1. For other data sources, such as micro-stations outputs, this argument is indispensable to identify which variables will be summarized. The argument **days.id** indicates which day of the year (or days from the beginning of record), and the default is *daysFromStart* column from *get_weather* function. A basic example of this use is given below.

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 (in days after the beginning of the data). For example, `time.window = c(0,14,35,60,90,120)` denote the intervals of 0-14 days from the first day of record (0). If the first record denotes the emergence date of the crop in the field, this can also be associated a 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 will 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 quantiles: first (25%) second (50%, median) and third (75%).

Tools for basic data processing

Basic data processing is done by *processWTH* function. As described for *summaryWTH*, this function also can be used to process environmental data for get_weather outputs and other sources (micro-stations, in-field sensors) using the same arguments of identification (**`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 process, such as macro effects of the soil-plant-atmosphere dynamics and the impact of the atmospheric temperature on crop

development. Below we describe these three functions and the ecophysiological temperature concepts underlying their application.

Radiation-related covariables

The radiation balance in crop systems is regulated by the difference between the amount of incident radiation, absorbed energy by the plants and soil surface, and the converted thermal energy. From Nasa Power, the radiation outputs are given in terms of Top-of-atmosphere Insolation (ALLSKY_TOA_SW_DWN), Insolation Incident on a Horizontal Surface (Shortwave, ALLSKY_SFC_SW_DWN), and Downward Thermal Infrared Radiative Flux (Longwave, ALLSKY_SFC_LW_DW). Thus, the net solar radiation available for the physiological process of growth (biomass production) is given by the difference between longwave and shortwave, i.e., $SRAD = ALLSKY_S FC_L W_D W - ALLSKY_S FC_S W_D WN$, in MJ m^{-2} d^{-1}.

Most of the growth modeling approaches, the effect of radiation use efficiency (RUE) is the main target to describe the relations between the available energy in the environment and how the plants translate it in biomass (see subsection Processing Thermal Parameters). In this context, this source of environmental variation is important to understand the differences in potential yield observed for genotypes evaluated across diverse environments. Radiation is also important as a source to regulate the available energy for other biophysical process, such as evaporation, transpiration and temperature (see subsection Processing Atmospheric Parameters).

EnvRtype made available a function named *Param_Radiation* to compute additional radiation-based variables that can be useful for plant breeder and researchers from 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 the year (julian day, from 1 to 365) and cloudiness (for *n*). In additional, the lobal solar radiation incidence (SRAD, in MJ m^2 d^{-1}) is computed as described in 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 detail 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

Thermal variables are important for regulating the rates of important biochemical processes within the organisms. At cell level, the effect of temperature may regulate the rate of enzymatic reactions, in which critical values may led 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 in the absorption of nutrients from water flux in in the roots. At 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 is also strongly influenced by temperature (e.g., growing degree-day, *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 factor of temperature on radiation use efficiency (*FRUE*). The first is useful to predict the phenology development, while the second is an ecophysiology parameter 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 the temperature variations shapes the adaptation of some specie at the considered environment. *GDD* is also important for modeling plant-pathogen interactions, because some pests and diseases have their temperature-regulated growing.

In this context, dew point (*T2MDEW*) is another agrometeorology with great importance for crop sanity. In addition to being related to evaporation process in the stomata, this factor shapes the establishment of diseases (especially fungus) under the leaf pages. Finally, the daily temperature range (*T2M_RANGE*) is a factor impacting process such as floral abortion for crops were the main trait are related to grain productions. For more detail about the impact of temperature in diverse agricultural crops, please check Luo (2011).

The function *Param_Temperature* 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`**). For running this function with other data sources than get_weather, 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 ecophysiology literature previously described. Consider the dry bean crop at the same location in Nairobi, Kenya (previous box examples). The cardinals for dry bean were collected from Table 2.

BOX 5: Practical use of Param_Temperature for Dry Bean Crop in Nairobi, Kenya

```
> Param_Temperature(env.data = env.data,Tbase1 = 8,Tbase2 = 45,Topt1 = 30,Topt2 = 35)
```

Atmospheric demands

The dynamic of water precipitation (rainfall) and water demand (evaporation+plants transpiration) is regulated as a consequence of the balance of radiation and thermal-related process in the atmosphere (Soltani and Sinclair, 2012; Allen et al., 1998). The soil-plant-atmosphere continuum involves the water dynamics from the soil, passing by plant tissues and going back do the atmosphere by the stomata. The rate of this process is deeply related to the biomass production of plants, but also in the absorption of nutrients by mass-flux in roots. Because of that, the water demands are essential to measure the quality of some growing environment.

We created the function *Param_Atmosphere* to run basic computation 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°C, default is `Tmax='T2M_MAX'`), **`Tmin`** (minimum air temperature °C, default is `Tmin='T2M_MIN'`), **`RH`** (relative air humidity %, default is `RH='RH2M'`), **`Rad`** (net radiation, R_n, in MJ m^{-2} day^{-1}, default is `Rad ='Srad'`), **`alpha`**, (empirical constant accounting for vapour 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`). The usage of this function works in similar manner than the other two *Param* functions previously described.

From these inputs, we use 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*;

Below we present an example of usage for Nairobi, Kenya. Consider the same env.data collected in the previous box, and 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

Environmental characterization is a fundamental step to understand how the environment regulates the gene-expression and phenotypic variation of the genotypes under diverse growing conditions (Xu 2016; Costa-Neto *et al.* 2020a). In this step, the collected raw environmental data is translated in useful information for both predictive or exploratory analysis, such as factorial regression focused on G×E analysis, or for environmental grouping. The typing of the environmental variations provides a better profiling visualization of which factors are impacting in differing ways from one environment than others. If there is no difference among environments, it is reasonable to assume that all phenotypic variations in field are due the genetic-related differences. Thus, envirotyping has the power to detect these differences and explore it to enhance genomic analysis, such as genomic prediction (GP) for multiple environments.

Here we develop tools to facilitate the envirotyping of biophysical factors in two ways, a qualitative standpoint (discovering envirotype descriptors) and quantitative (creating quantitative covariables). Bellow we highlight some concepts underlaying the environmental characterization protocols. Next, we present the implementation of those concepts using the EnvRtype functions.

Envirotype profiling with EnvTyping

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 resources availability, which can be described as a function of how much resources are available and the frequency of occurrence of those resources (e.g., transitory or constant effects). In addition, the relation of resource absorption and allocation depends on plant characteristics (e.g., phenology, current sanity status). Then, this particular environmental-plant influence is named after envirotype to differ to the concept of raw-environmental data (data collected directly from sensors). It also can be referred as environmental type (ET). Finally, the typing for environments can be done by discovering ETs, and the similarity among environments is a consequence of the number of ETs shared between each pariwise environment.

Before the discover of ETs, a first step is the design by ecophysiology concepts (e.g., plant necessity for some resource) or summarizing the raw data from the core of environments in analysis. Then, for each ET is computed the frequency of occurrence, which represent the frequency of certain quantities of resources for plant development. The typing by frequency of occurrence provides a deeper understanding about the distribution of the events, such as rainfall distribution across different growing cycles and occurrence of heat-stress conditions for a target

location (Heinemann *et al.* 2015). Thus, groups of environments can be better identified by analyzing the core of the events occurring on a target location, year or planting date. This step can be done not only using grade-point averages (e.g., accumulated sums or means for specific periods) but in terms of their historical similarity. In this way, we can not only group environments in the same year, but 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 interval) or specific phenological stages for a specific crop.

To implement envirotype profiling, we create the *EnvTyping* function. This function computes the frequency of occurrence of each envirotype across diverse environments. This function as 12 arguments in which the 9 of them (**`env.data, id.names, env.id, days.id var.id, statistic, by.interval, time.window,names.window`**) works in the same way already described in the previous functions. The novel argument **`cardinals`** are responsible to define the biological thresholds between envirotypes and adaptation zones. These cardinals must respect ecophysiological limits for each crop, germplasm or region. For that, we suggest reading the literature of ecophysiology and crop growth modeling, such as Soltani and Sinclar (2012). The argument **`cardinals`** can be fill as vector (for single-environmental factors) or as 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 as for Nairobi, Kenya:

BOX 7: Basic use of EnvTyping for typing temperature in Nairobi, Kenya

```
> EnvTyping(env.data = env.data, var.id  = 'T2M', cardinals = c(0,8,30,35,40))
```

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

For multiple environmental factors, a list of cardinals must be created. For example, considering the variables 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 demands 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 rainfalls could be associated negatively with flooding of the soil, drainage of fertilizers, among other factors related to crops lodging or diseases occurrence. Thus, for PRECTOT, the cardinals will be `cardinals = c(0,10,40,Inf)`. For dew point, let's assume a data-driven typing (`cardinals = NULL`) using the quantiles previously described. Taking the same example for Nairobi, Kenya:

BOX 8: Basic use of EnvTyping 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
> EnvTyping(env.data = env.data, var.id  = var, cardinals = card)
```

Environmental Covariables with W.matrix

The quality of an environment is measured by amount of resources availability to attempt the demands of the plants. Over 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 the use of phenotypic data as 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 as a covariate matrix (**W**), following the terminology used by Costa-Neto et al. (2020a) and de los Campos et al. (2020). From these **W** matrix, several analyzes can be used, such as: (1) dissecting the G×E interaction; (2) model the genotype-specific sensibility for key environmental factors; (3) dissecting the environmental factors of the QTL×E interaction; (4) integrate environmental data to model the gene × environment reaction-norm; (5) basic summary of the environmental gradient in some experimental network; (6) to produce environmental relationship matrices for genomic prediction.

To implement these applications, first the processed environmental data must to be translated into quantitative descriptors, by summarizing cumulative means, sums or quantile, such as in *summaryWTH*. However, this data must be mean-centered and scaled to assume a normal distribution and avoid variations due differences in scale dimension. For creating environmental similarity kernels, Costa-Neto et al (2020a) suggested the use of quantile statistics to better describe the distribution of each variable across the experimental network. Thus, this allows a statistic approximation of the ecophysiology importance of the environmental variables during crop growth and development. In this context, we developed the function *W.matrix* to create a double entry table (q environments/sites/years × k environmental factors). Conversely to *EnvTyping*, 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 *EnvTyping* are applicable (**`env.data, id.names, env.id, days.id var.id, statistic, by.interval,`**

462 **time.window,names.window**). However, in *W.matrix* the arguments **center** = TRUE (by
463 default) and **scale** = TRUE (by default) denotes mean-centered ($w - \bar{w}$) and scaled ($w - \bar{w}/\sigma$),
464 in which w is the original variable, $\bar{w}$ and σ are the mean and standard deviation of this
465 covariable across the environments (BOX line 9). A quality control (**QC** = TRUE argument) is
466 done by removing covariables with more than $\pm \sigma_{TOL}\pm \sigma$, where σ_{TOL} is the tolerance limit for
467 standard deviation, settled by default argument as *sd.tol* = 3.

468 To exemplify a basic use of *W.matrix*, lets consider the same use for Nairobi, Kenya,
469 involving only weather variables of temperature and rainfall precipitation, and assuming a
470 quality control of *sd.tol* = 4. The time intervals were settled for each 10-day (default), and statistic
471 as 'mean' for each variable at each time interval.

BOX 9: Basic use of EnvTyping 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)
```

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

473 Whole-genome prediction (GP) has revolutionized both plant and animal breeding
474 pipelines around the world. This technology enables an indirect selection of untested genotypes
475 using statistical and computational approaches to link the phenotypic records and high-dense
476 markers from related genotypes tested in the field trials. Since the first work proposing this
477 methodology (Meuwissen et al 2001), GP has evolved for multiple scenarios (multi-trait, multi-
478 environment), data sources (e.g., Westhues et al., 2017; Costa-Neto et al., 2020a) and
479 computational approaches (e.g., Morota and Gianola, 2014; Cuevas et al., 2019; Crossa et al.,
480 2019; de los Campos et al., 2020). Most of those approaches relies on increase the accuracy of

481 modeling genotype-phenotype patterns and explore it as predictive breeding tool. Among the
482 several enrichments in computational efficacy and breeding applications, the integration of
483 genomic by environment interaction (G×E) has boosted the ability of the genomic-assisted
484 selection for evaluating a wide number of genotypes under several growing conditions over
485 multiple environmental trials (MET).

486 Heslot et al (2014) and Jarquín et al (2014) introduced the use of environmental
487 covariables to model an environmental source of the phenotypic correlation across MET. These
488 approaches aim to model the reaction-norm of genotypes across MET, i.e., how different
489 genotypes react to the different environmental gradient variation. For most cases, the reaction-
490 norm modeling serve as additional source of variation for complementing the genomic
491 relatedness among individuals tested and untested under know environmental conditions. Thus,
492 in addition to the genomic kernels, now the envirotype-informed kernels can be used to capture
493 macro-environmental relatedness shaping the phenotypic variation of relatives, the so-called
494 *enviromic kernel* (Costa-Neto et al., 2020).

495 In the third Module of EnvRtype package, we present the tools implement this type of
496 modeling approach. Two main functions were designed for those purpose. First for the
497 construction of the environmental relationship kernels, the *EnvKernel*. The second is *get_kernel*,
498 aiming to integrate these kernels into statistical models for the GP-based selections across MET.
499 In the next subsections, we describe the kernel methods to model envirotype-relatedness. Then,
500 we present the statistical models that can be built with these kernels.

501 **Enviromic Kernels with EnvKernel**

502 In this package we use two types of kernel methods to compute enviromic-based
503 similarity. The first consists of the traditional method based on the linear variance-covariance

matrix (Jarquín et al., 2014). This kernel is similar to GBLUP for the purpose of markers and can be described mathematically as:

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

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 (2020) and is described in a similar way to the approach already used for modeling genomic effects:

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

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

Both methods are implemented by the EnvKernel 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 argument `gaussian` (default is `gaussian = FALSE`) denotes if the models (1) or (2) are used to compute $\boldsymbol{K_E}$. If `gaussian = TRUE`, so the gaussian kernel (equation 2) is used, and `h.gaussian` must be inserted to compute

kernel. In the argument **Y** (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).

EnvKernel function has two outputs, named *varCov* (for variable's covariance) and *envCov* (for environments covariance). The first is useful to deepen the understanding 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 'Statistical Models for Genomic Prediction").

A basic use of EnvKernel 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 gaussian kernel is given as:

BOX 10: Basic use of EnvKernel

```
> EnvKernel(env.data = W, gaussian = FALSE)
> EnvKernel(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 some packages available in R CRAN. However, it is important to consider that statistical models containing more complex structures (e.g., more than one genomic effect plus G×E and environmental information) are naturally models with a 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 in front to BMTME. For this reason, EnvRtype has a function named *get_kernel* aimed to organize the genomic or envirotyping-based kernels in different statistical model structures to be run in 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:

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

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 $\boldsymbol{y} = \left[\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots \boldsymbol{y_q}\right]^T$. The scalar $\mathbf{1}\boldsymbol{\mu}$ 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 types of modeling:

i. **Genotype-effect GP Models**. Involves the baseline models accounting only for genotype-based effects, mostly associated with pedigree-based or genomic realized kinships. Those models consider $\sum_{s=1}^{p} \boldsymbol{g_s} \neq 0$ 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, such as for additive (A) and dominance (D) deviations and other sources of information from "omics". All genomic kernels must to have the $p \times p$ dimension, in which p is the number of genotypes.

ii. **Enviromic-enriched GP Models**. From the MM and MDs models, we add the acronym "E" to denote "enviromic-enriched" for EMM and EMDs models. Those 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 effect (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 ($\sum_{r=1}^{q} \boldsymbol{w_r}$).

iii. **Enviromic-based Reaction-Norm GP.** From the MM and MDs models, we add the acronym "RN" from "reaction-norm", resulting in RNMM and RNMDs models, respectively. As described in (ii), the environmental effects can now be modeled as fixed deviation (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 effect (W) plus a envirotype × genomic interaction (G×W). In this context, RNMM accounts for the variation due G+W+GW, whereas RNMDS considers G+GE+W+GW.

589

590 The *get_kernel* function has four main arguments, which is a list of genomic relationship
591 kernels (**K_G**), a list of environmental relationship kernels (**K_E**), and phenotypic MET data set
592 (**Y**), organized as vector of environment identification, vector of genotype identification and
593 vector of trait values. Finally, the argument **model** sets the statistical model used ("MM", "MDs",
594 "EMM", "EMDs", "RNMM" and "RNMDs"). Each genomic kernel in **K_G** must have the dimension of
595 $p \times p$ genotypes. At the same manner, the **K_E** might have the dimension of $q \times q$ environments,
596 but in some cases the environmental kernels can be built at phenotypic observation level. This
597 means that for each genotype at each environment, there is a different ECs, according for
598 particular phenology stages or envirotyping at plant level. Thus, using the additional argument
599 **size_E** = 'full' (default is 'environment'), the **K_E** may accomplish a kernel with $n \times n$, in which n
600 = pq. The basic usage of *get_kernel* in given in Results section.

601 **RESULTS**

602 Three sections of results were implemented to give a comprehensive overview of the most
603 important functions of EnvRtype. First, we illustrate the use of EnvRtype in starting an
604 envirotyping pipeline over different locations in the world. Second, we used a toy data set to
605 demonstrate the creation of different environmental similarities based on different
606 environmental factors. This type of application can be useful for researchers interested in predict
607 the particular genotypic responses shaped by genomic and enviromic-specific factors across
608 existing experimental trials or for assembly virtual scenarios. Finally, we compare the kernel
609 methods for modeling environmental similarity in GP. For these last sections, we expect to give
610 some insights to facilitate the usage of enviromic data in boosting 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 of 9 locations around the world: 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 15 th May 2019 to 15th June 2019), Brisbane (Australia, 27.47 S, 153.02 E, from 15 th 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 the identification 'GOI', 'TEX', 'BRI' ,'MON', 'LOS', 'PON', 'CAL', 'PAL' and 'DAV' for each location, respectively (Figure 2A).

BOX 11: Remote Sensing for Serveral 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)
```

From the collected variables, its possible to type any environmental factor or a core of environmental factors (Figure 2B). As a toy exemplification, we use 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 sign of environmental distance, but it also can be used as kinship for predictive breeding (Costa-Neto et al., 2020a).

BOX 12: Discovering ETs and similarity among locations

```
> ET = EnvTyping(env.data = env.data,env.id = 'env',var.id = "T2M"))
> EC = W.matrix(env.data = env.data,var.id = 'T2M')
> distances =  EnvKernel(env.data = ET,gaussian = T)[[2]] # fig a
> kinship   = EnvKernel(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 similarity. The combination of the remote sensing + typing strategies is a powerful for turbocharging global patternships of field testing and germplasm exchange. It also contributes for increase 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 might also be leveraged for specific regions (Costa-Neto et al., 2020b). It also can increase for a better

641 definition of field trials positioning (Tassinari et al., 2020) and how breeding strategies have
642 impact on crops adaptation in the past (Heinemann et al., 2019);

643 **Panels of environmental similarity and reaction-norm**

644 To illustrate the use of different ECs in GP, we run a toy example involving a tropical maize
645 set available in EnvRtype. This data set was included in Souza et al. (2017) and Cuevas et al
646 (2019) and cames from the Helix Seeds Company (HEL). However, to facilitate the demonstration
647 of functions, we made available a subset of 150 hybrids per environment, thus counting 750
648 genotypes per environment observations. Grain yield data are mean-centered and scaled
649 (*MaizeYield* object). Genotyping relationship for additive effects is based on 52,811 SNPs are also
650 available to make the predictions (*maizeG* object). The phenotypic and genomic data of inbred
651 lines are credited to Helix Seeds Ltda. Company. Finally, weather data are presented for each one
652 of the 5 environments (*maizeWTH* object). This data sets becomes available in the R environment
653 by running the following R-code:

BOX 13: Toy data sets for illustrate GP examples

```
> data("maizeYield")   # phenotype data (grain yield per environment)
> data("maizeG")       #  genomic relationship for additive effects
> data("maizeWTH")      # environmental data
```

654

655 We run an example of GP considering reaction-norm and different levels of envirotyping
656 per environment, which is: (1) envirotyping mean values per environment (entire croplife) and
657 (2) envirotyping mean values for each time interval across crop life. We consider two types ECs:
658 factor of temperature effect over radiation use efficiency (FRUE) and the difference between

rainfall precipitation and crop evapotranspiration (PETP). From equation (3), this ECs were arranged in four kernel structures using the RNMM model:

- Model 1: genomic plus an enviromic kernel build 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 build with a single EC (PETP):
  - $\boldsymbol{y} = \mathbf{1}\boldsymbol{\mu} + \boldsymbol{g} + \boldsymbol{PETP} + \boldsymbol{g} \times \boldsymbol{PETP} + \boldsymbol{\varepsilon}$ (5)
- Model 3 [Joint EC model]: genomic plus an enviromic kernel build with FRUE and PETP ($\mathbf{W}$). This is the benchmark reaction norm model (Jarquín et al., 2014), but here considering 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:

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

These envirotyping levels and model structures can be implemented in EnvRtype as:

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 = EnvKernel(env.data = wFRUE1)[[2]]
> K2 = EnvKernel(env.data = wPETP1)[[2]]
> K3 = EnvKernel(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))
```

From a brief diagnosis of variance components and model fit (correlation between observed and predicted values without missing entries, *r*), its possible to observe that for the same raw-environmental data, each envirotyping level and modeling structure impacts on modeling the phenotypic variation. When envirotyping are done by environment, there is a bad fitness (from 0.48 in Model 1 to 0.57 in Model 4) in relation to the models with envirotyping level by time interval (from 0.92 in Model 2 to 0.97 in Model 1). It also reflected how the different levels of envirotyping impact on the understanding of which ECs explain the phenotypic variation across field trials.

In addition, the envirotyping level per time-interval lead to an increase of genomic variance components and drastically reduction of 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 effect were better visualized in the model involving a second covariate (Model 4), because in this

688 sense it is expected to better capture the single effect of PETP free of the inner effect of other
689 covariates. This ECs represents the atmospheric demand over the soil-plant-atmospheric
690 continuum. Thus, for predictive purposes, we suggest to use also ECs from soil-water balance
691 (e.g., soil water potential), in which can be collected from field-based sensors or estimated using
692 crop growth modeling approaches.

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

700 **Benefits of Gaussian Kernel for GP with enviromic data**

701 Finally, to illustrate the differences in kernel methods in reproducing environmental
702 similarity, we compare the last results from models 4-7 now using a gaussian kernel approach.
703 For that, we use the same codes from the box 14, but now the argument gaussian is `gaussian =`
704 `TRUE`.
705

BOX 15: Use of Gaussian Kernel for modeling enviromic kinships

```
### 2- Kernels
> K1 = EnvKernel(env.data = wFRUE1,gaussian=TRUE)[[2]]
> K2 = EnvKernel(env.data = wPETP1,gaussian=TRUE)[[2]]
```

```
> K3 = EnvKernel(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);
```

As expected, it was observed differences in modeling structures (Figure) and ECs importance (Table 4). However, it was observed an increase in all models in relation to the benchmark linear covariance matrix (Table 3). The models with envirotyping level per environment were the most benefit from using gaussian kernel Model 1 using envirotyping level per environment were the most successful model in reducing residual variation and fitness (r = 0.99) when FRUE are modeled from gaussian kernel. However, the models with envirotyping level per time-interval still outperforms the envirotyping per environment in adjusting models with more suitable to explain the phenotypic variance from genomic kinships. This is a signal that models with those level of envirotyping data may also be useful to increase the accuracy of GP for multiple environments for conditions with low genomic-heritability. Finally, in comparison with the linear-covariance matrix, the nonlinear gaussian kernel were more efficient in capturing genomic × enviromic interactions (comparison between Tables 3 and 4). In addition, the importance of FRUE and PETP were better elucidated using gaussian kernel, despite the internal differences between those two ECs are not changed from the observed using linear covariance matrix.

CONCLUSION

The collection, processing and use of envirotyping data in genomic-based studies depends not only of the quality of the data sources. Here we demonstrate that the increased ecophysiology knowledge in envirotyping is benefit not only to increase accuracy of statistical models in

726 genomic prediction, but also to provide a better explanation of the sources of variation and
727 increase efficiency in those models. The correct use of envirotyping data depends on the quality
728 of data processing and it is specific for each crop specie (or living organism). A same
729 'environment' (considering a time interval for a target location) may result in different
730 environmental types (ETs) for each organism, which depends on their sensibility in respond to
731 constant and transitory variations of the environment. Thus, in this study we presented some of
732 those concepts and created functions (and gather others from different R packages) to facilitate
733 the use of envirotyping data in quantitative genomics. We also show that global envirotyping
734 networks can be build using remote sensing tools and functions provided in EnvRtype. Other
735 uses of the functions presented here may involve: (1) the creation of multiple environment
736 scenarios for predictive breeding; (2) an enviromic scan of which ETs better explain
737 environmental similarities. Then, this information can be used for design better experimental
738 networks and accelerate the screening of genotypes for target environments in which perform
739 better; (3) analysis of historical trends to quantify the impact of recent climate changings in long-
740 term breeding strategies conducted in target regions; (4) integrate crop growth models, but as
741 inputs in functions such as W.matrix and EnvTyping, but also to predict the crops performance
742 across diverse water or nitrogen management levels, which also may involve the use of the
743 collected data from get_weather and other sources.

744

745 **REFERENCES**

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

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.

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

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

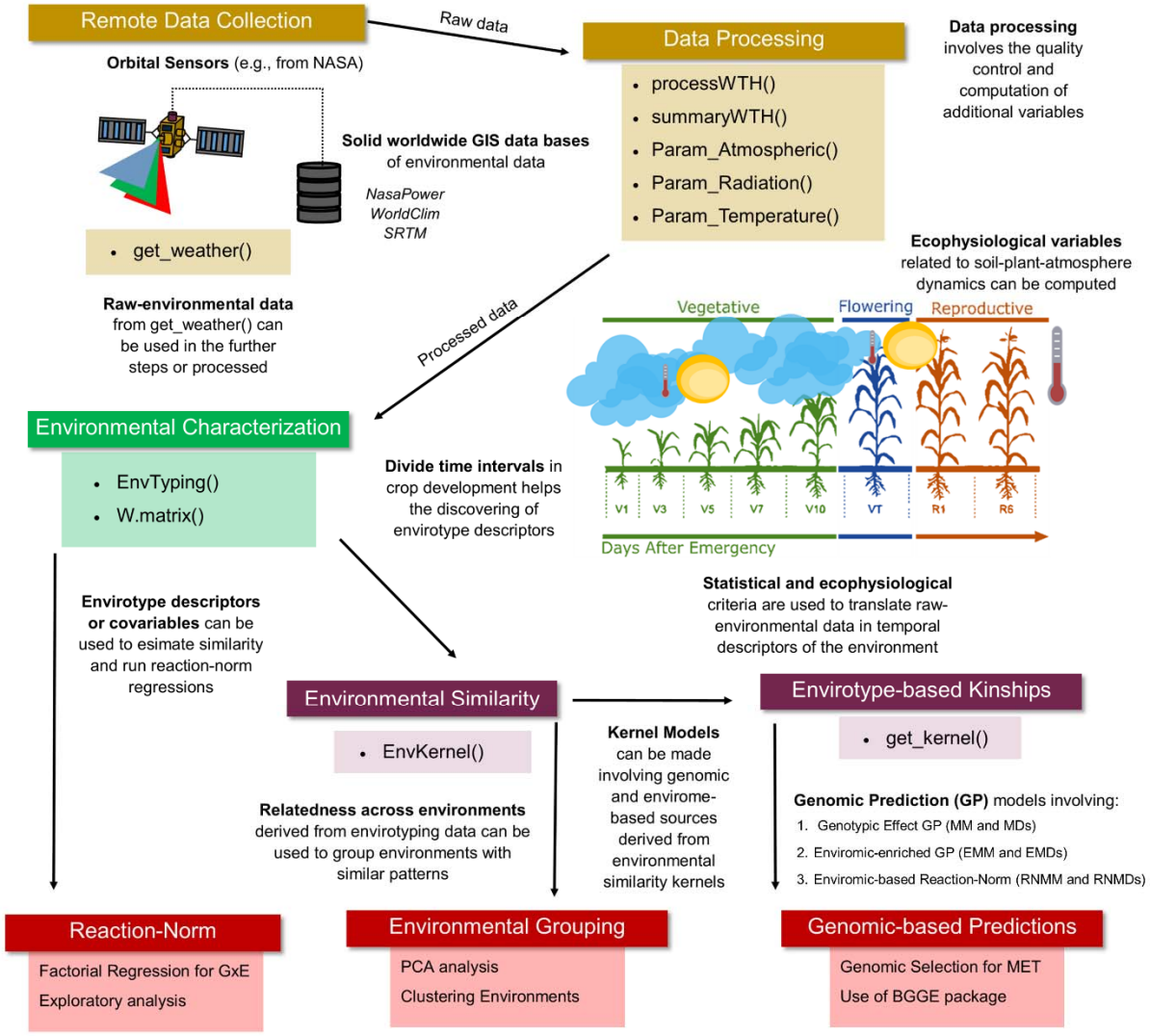


Figure 1. Workflow of the envirotyping pipeline implemented using EnvRtype in R. Yellow, Green, Wine and Red colors box denote the steps related to the Modules 1, 2, 3 and the outputs from EnvRtype. Black arrows indicate the flux direction of the envirotyping pipeline

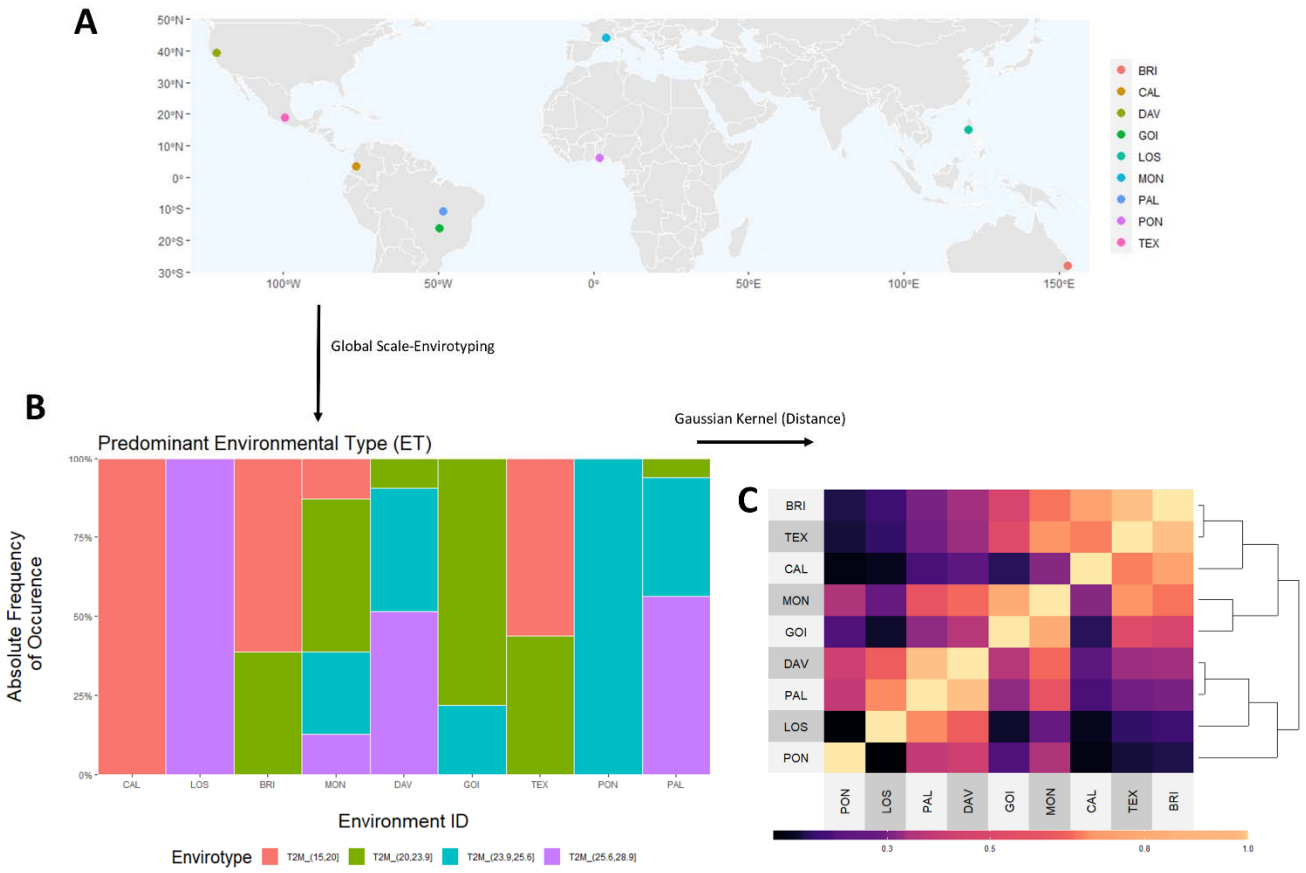


Figure 2. Global scale envirotyping highlight possible environmental similarity for locations in different continents. **A.** Worldspread geographic positions of 9 locations used as toy-example. **B.** Panel of environmental types (ET) for average air temperature during a certain month of a certain year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

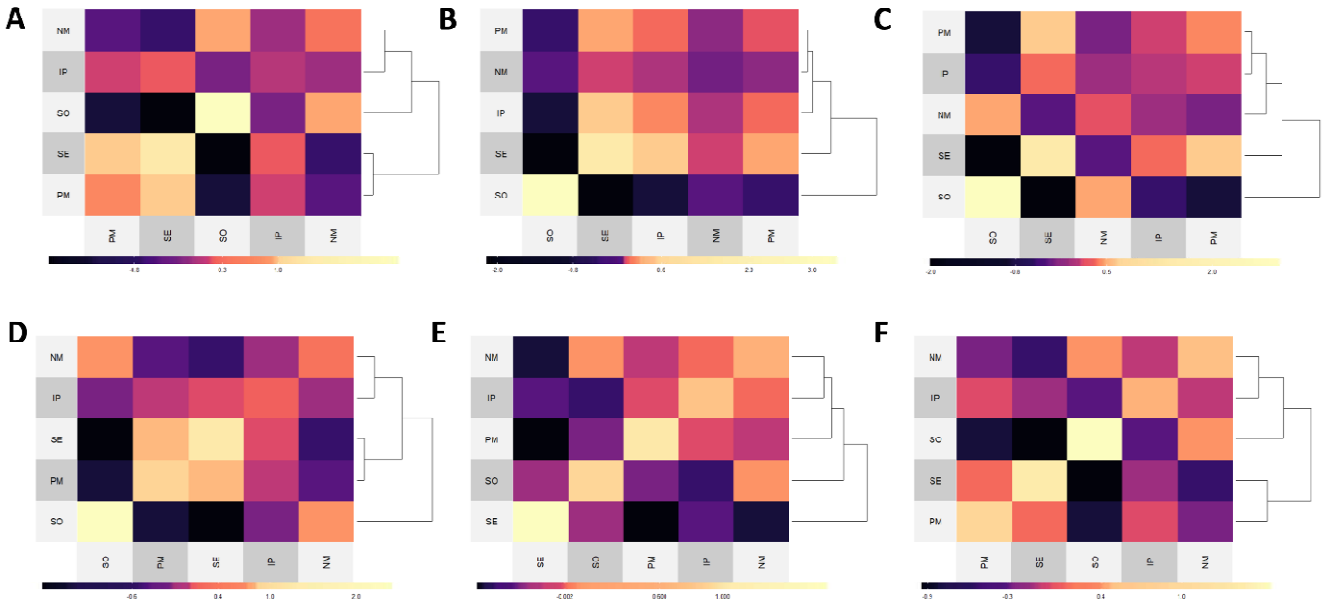


Figure 3. Linear enviromic kernels based the combination of two environmental covariates (ECs) and two envirotyping levels for 5 locations over an experimental network of tropical maize. Kernel were based on FRUE variable (impact of temperature on radiation use efficiency) for envirotyping at entire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

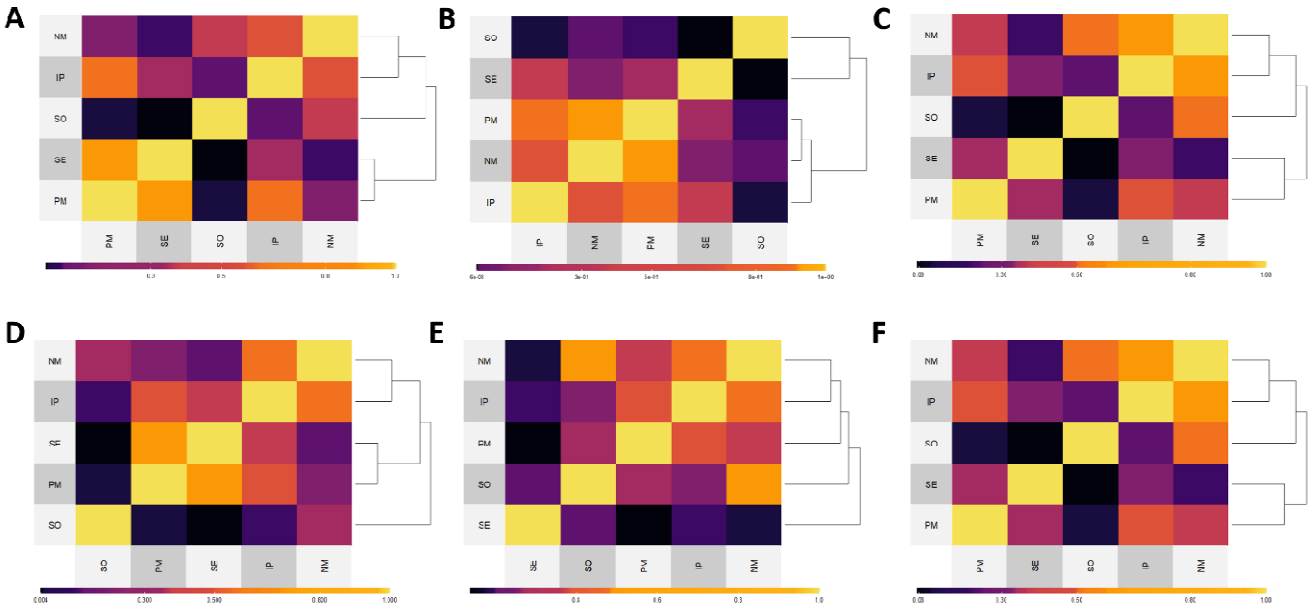


Figure 4. Nonlinear enviromic kernels (gaussian) based the combination of two environmental covariates (ECs) and two envirotyping levels for 5 locations over an experimental network of tropical maize. Kernel were based on FRUE variable (impact of temperature on radiation use efficiency) for envirotyping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

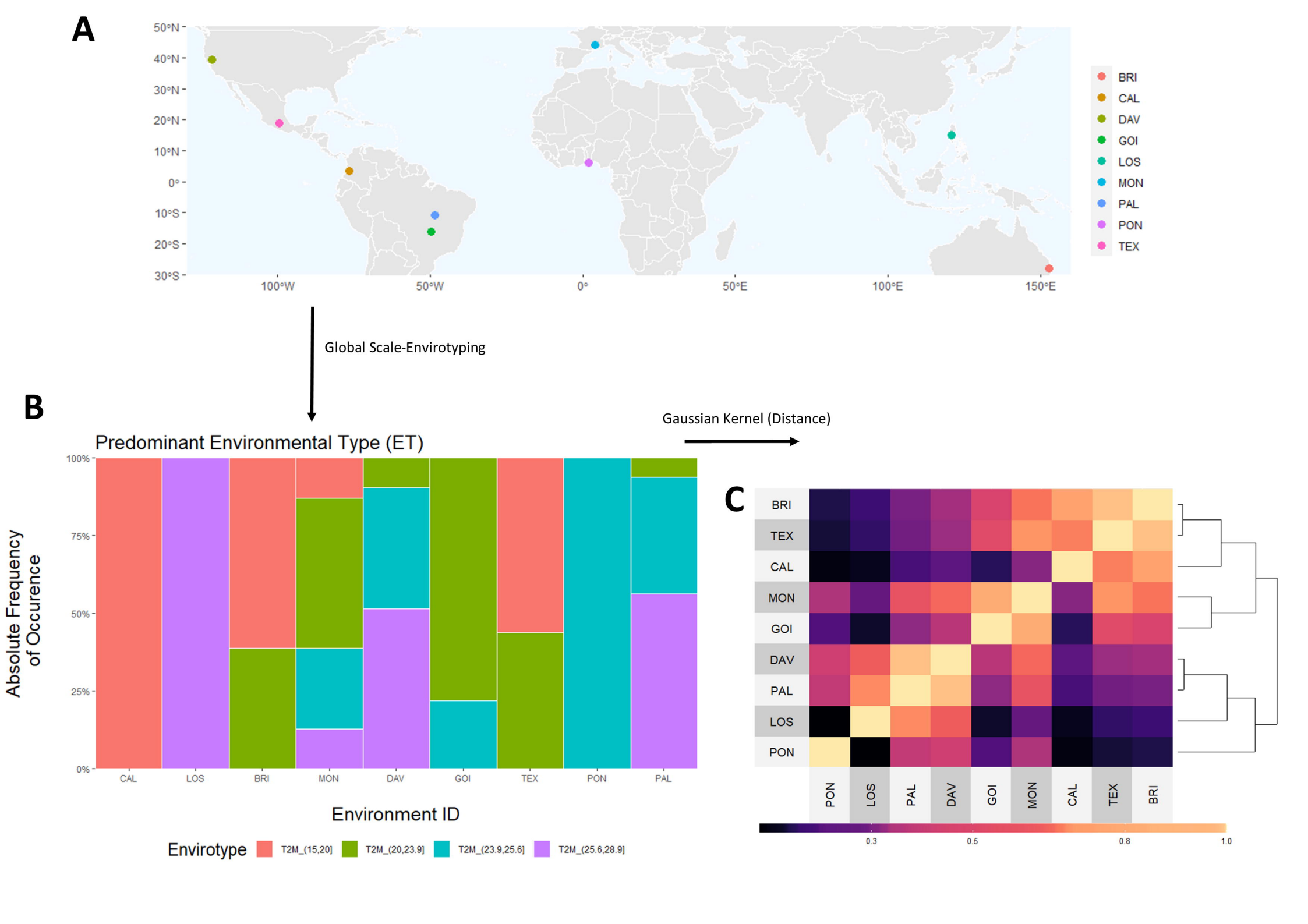
Specie	Suggested Cardinal Limit			
	Tbase1	Topt1	Topt2	Tbase2
Maize	8.0	30.0	37.0	45.0
Wheat	0.0	25.0	28.0	40.0
Rainfed Rice	8.0	30.0	37.0	45.0
Irrigated Rice (only vegetative stage)	8.0	28.0	40.0	45.0
Irrigated Rice (only reproductive stage)	15.0	25.0	35.0	45.0
Sorghum	8.0	30.0	37.0	45.0
Soybean	8.0	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

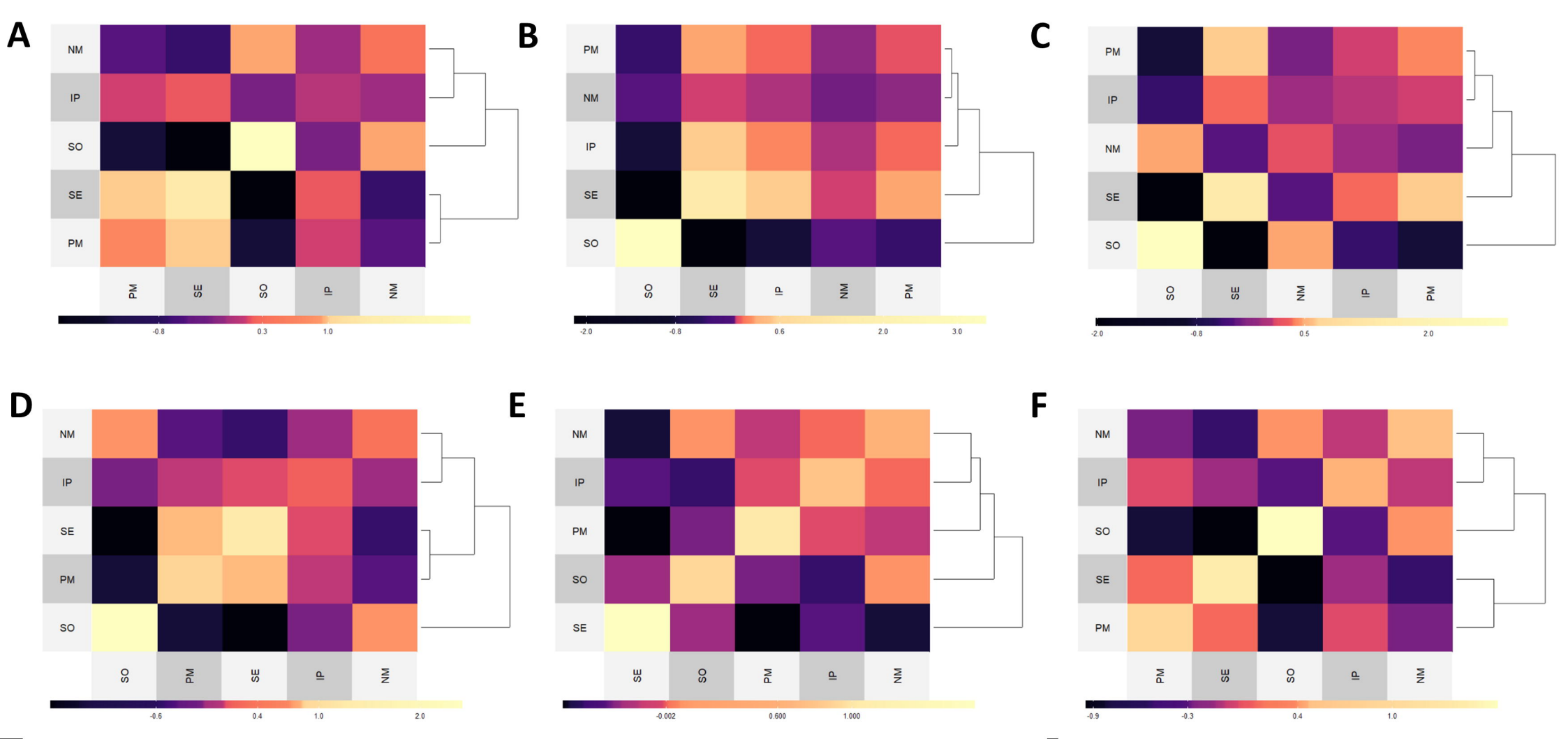
Table 3. Summary of a variance components and correlation between observed and predicted values (r) from a preliminary reaction-norm study involving two envirotyping levels (per environment, by.interval = FALSE; and per time intervals by environment, by.interval=TRUE), for a RNMM model involving additive 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).

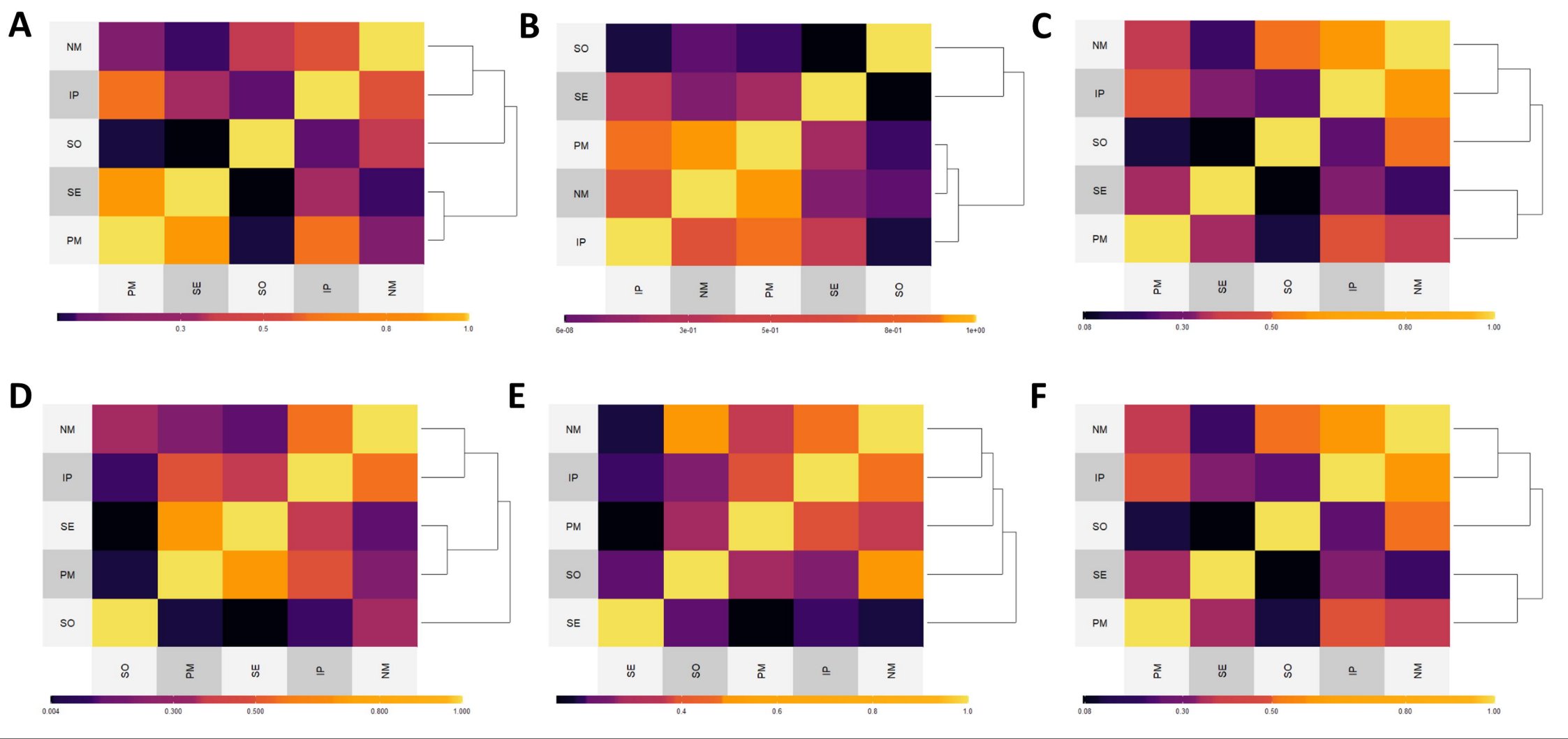
Effect per Model	Envirotyping level	
	Environment	Time Interval × Environment
Model 1: FRUE		
E = [FRUE]	2.025	15.841
Genomic (G)	0.428	0.482
GxE	0.037	0.074
Residual	0.855	0.152
Total	3.345	16.548
r =	0.48	0.97
Model 2: PETP		
E = [PETP]	3.058	4.093
Genomic (G)	0.454	0.481
GxE	0.039	0.093
Residual	0.826	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


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

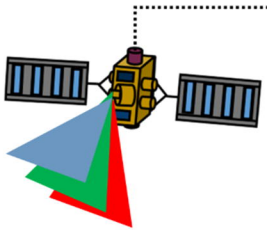

A
NM
IP
SO
SE
PM
PM
SE
SO
IP
NM
-0.8
0.3
1.0
B
PM
NM
IP
SE
SO
SO
SE
IP
NM
PM
-2.0
-0.8
0.6
2.0
3.0
C
PM
IP
NM
SE
SO
SO
SE
NM
IP
PM
-2.0
-0.8
0.5
2.0
D
NM
IP
SE
PM
SO
SO
PM
SE
IP
NM
-0.6
0.4
1.0
2.0
E
NM
IP
PM
SO
SE
SE
SO
PM
IP
NM
-0.002
0.600
1.000
F
NM
IP
SO
SE
PM
PM
SE
SO
IP
NM
-0.9
-0.3
0.4
1.0


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

Remote Data Collection

Raw data

Orbital Sensors (e.g., from NASA)



Solid worldwide GIS data bases of environmental data

NasaPower
WorldClim
SRTM

- get_weather()

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

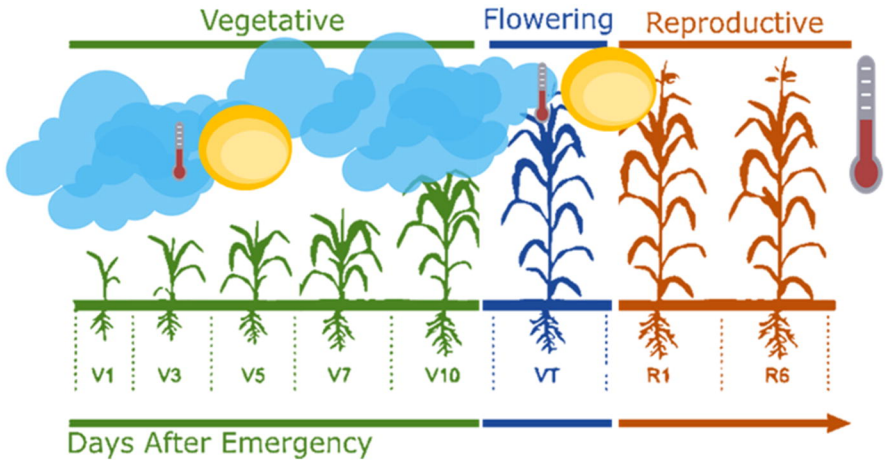
Data Processing

- processWTH()
- summaryWTH()
- Param_Atmospheric()
- Param_Radiation()
- Param_Temperature()

Data processing involves the quality control and computation of additional variables

Ecophysiological variables related to soil-plant-atmosphere dynamics can be computed

Processed data



Environmental Characterization

- EnvTyping()
- W.matrix()

Divide time intervals in crop development helps the discovering of envirotype descriptors

Envirotype descriptors or covariables can be used to esimate similarity and run reaction-norm regressions

Statistical and ecophysiological criteria are used to translate raw-environmental data in temporal descriptors of the environment

Environmental Similarity

- EnvKernel()

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

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

Envirotype-based Kinships

- get_kernel()

Genomic Prediction (GP) models involving:

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

Reaction-Norm

Factorial Regression for GxE

Exploratory analysis

Environmental Grouping

PCA analysis

Clustering Environments

Genomic-based Predictions

Genomic Selection for MET

Use of BGGE package