GEMS-R
WEBINARS

Aug 28th, 2021

Germano Costa-Neto

Institute for Genomic Diversity – Cornell University

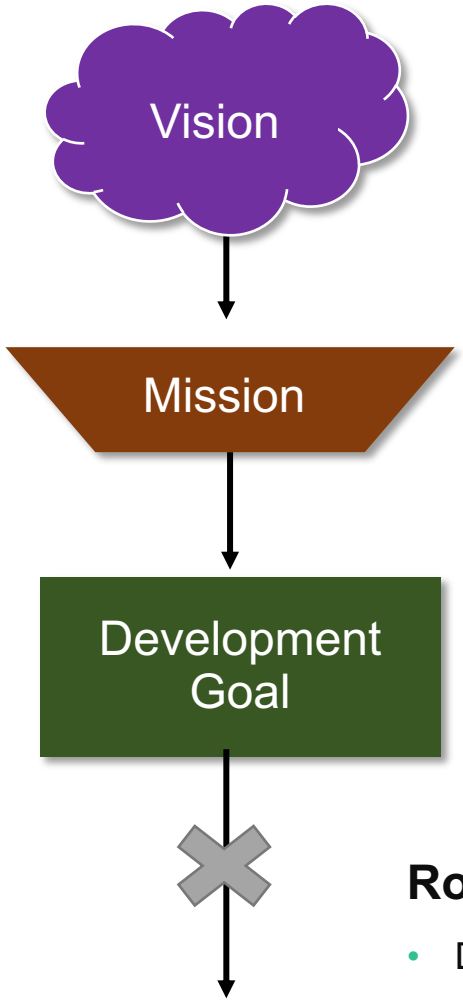
CIMMYT / BSU expert in envirotyping analytics

TRAINING COURSE

Basics for understanding envirotyping analytics and starting an enviromic pipeline in R

Overview

(GoHy, goal-directed hypothesis-driven)



If y = G+E, there is an urgent need to develop ways to better understand E in quantitative genetics

Merge genomics and enviromics: interplay quantitative genomics and ecophysiology matters in the plant breeding context

A pipeline for collect, process and integrate environmental information in the current quantitative genetics analysis and genomic prediction models across diverse growing conditions

Roadblock

- Does enviromics really improves the prediction ability of multi-environment GP?
- Does the use of enviromics might fill the lack of phenotypic records across drastically sparse multi-environment phenotyping networks?
- Which is the best kernel method to model the environmental relatedness ?
- Can we create an open-source software and a pipeline for it?

Part I

Theoretical Concepts & Data Set

Core ideas

- Accounting for ecophysiology knowledge using environmental information in predictive breeding
- Interplay environment, GxE, reaction-norm and phenotypic plasticity in the quantitative genetics and genomic prediction context

Environment

Check these references in the course repository
Cooper et al., (2014); Xu (2016); Resende et al., (2021); Costa-Neto et al. (2021a); Crossa et al (2021); Costa-Neto & Fritsche-Neto (2021)

Agronomic Standpoint

- Single-unit (planting date, location, management)
- Location + Planting date = Fixed + Random Factors
- Actual Management = Expected Management + biotic and abiotic interferences

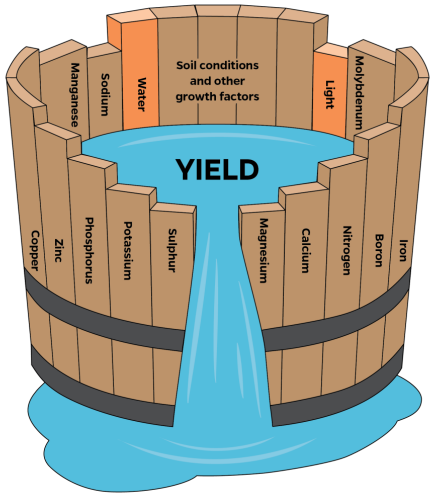
Geneticist Standpoint (Plant Breeders)

- Near-Iso Environments (managed conditions)
- Drought-Stress Trials (screening)
- Nitrogen level conditions (screening)
- A non-genetic source of variation (to be controled as a fixed factor)
- A sample of the theoretical **target population of environments (TPE)**

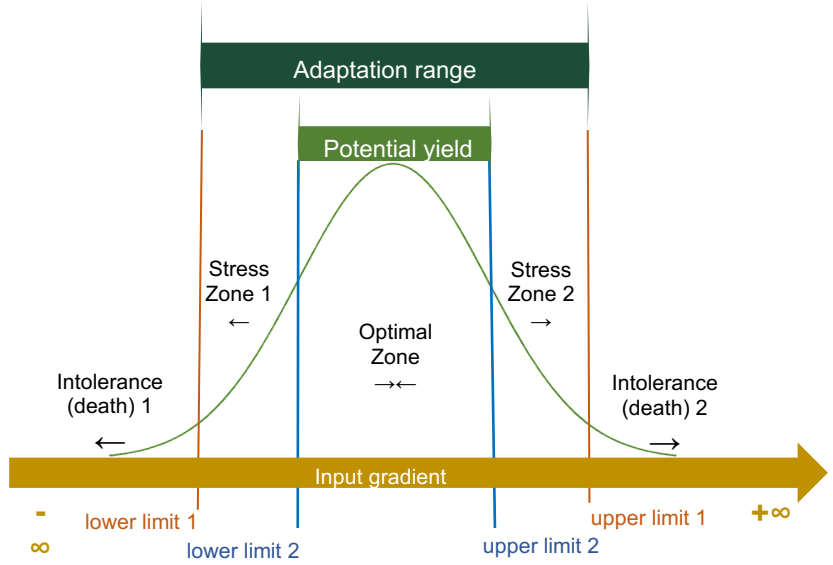
Ecophysiological Standpoint

- Relation between the **availability of inputs** and their resource absortion/alocation
- Availability as function of amount and frequency
- Core of events **linking** soil-plant-atmosphere dynamics (plant-env-pathogen)

Liebig's Law of Minimum (1840)
Whitson and Walster (1912)



Shelford (1931, 1932)
Tolerance Limits and adaptation



Then, what is envirotyping? enviromics? envirome?

- **Envirotyping** (environmental + typing) gathers the steps of collecting, processing, and associating environmental data with phenotypic data to understand the typology of environments for some MET, which is conducted relating it to the target population of environments (TPE) of the breeding program.

- **Enviromics** is the large-scale envirotyping, based on the collections of environmental data across time and space to establish a global association between the *crops envirome* (the core of TPEs for a certain genotype or specie) and the phenotypic variation of key factors driving G×E.

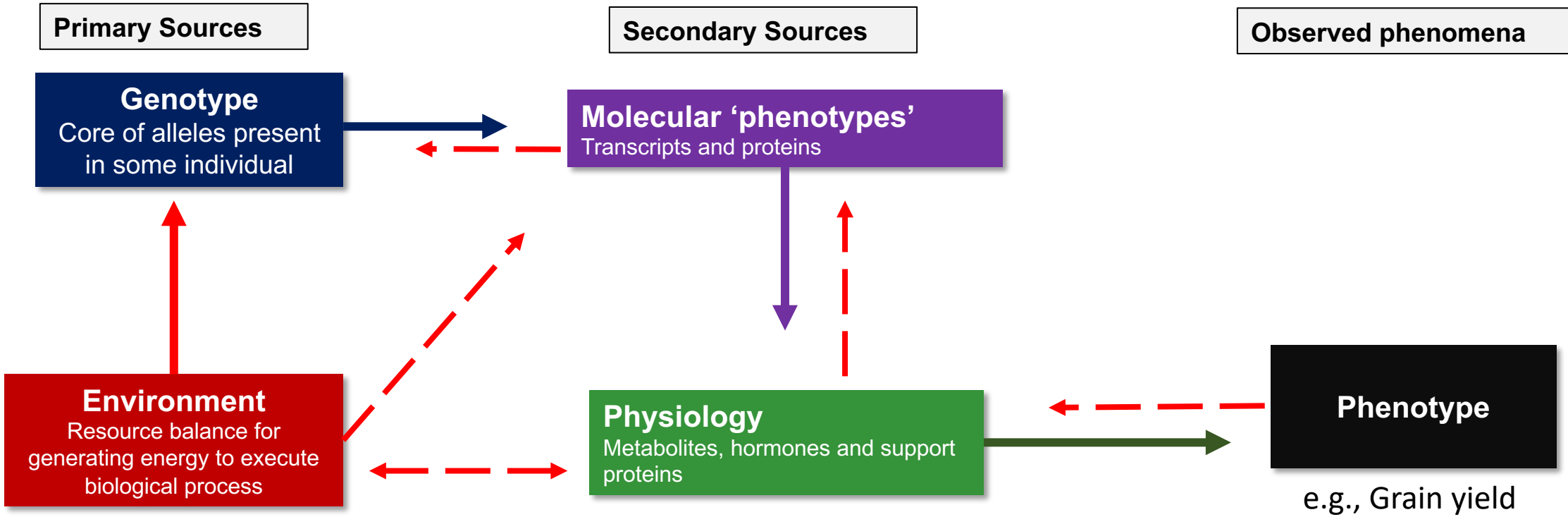
Why is envirOMICS ? Just a wave?

- **Because it can bridge different omics!**

Interplay Quantitative and Molecular Genetics in Predictive Breeding

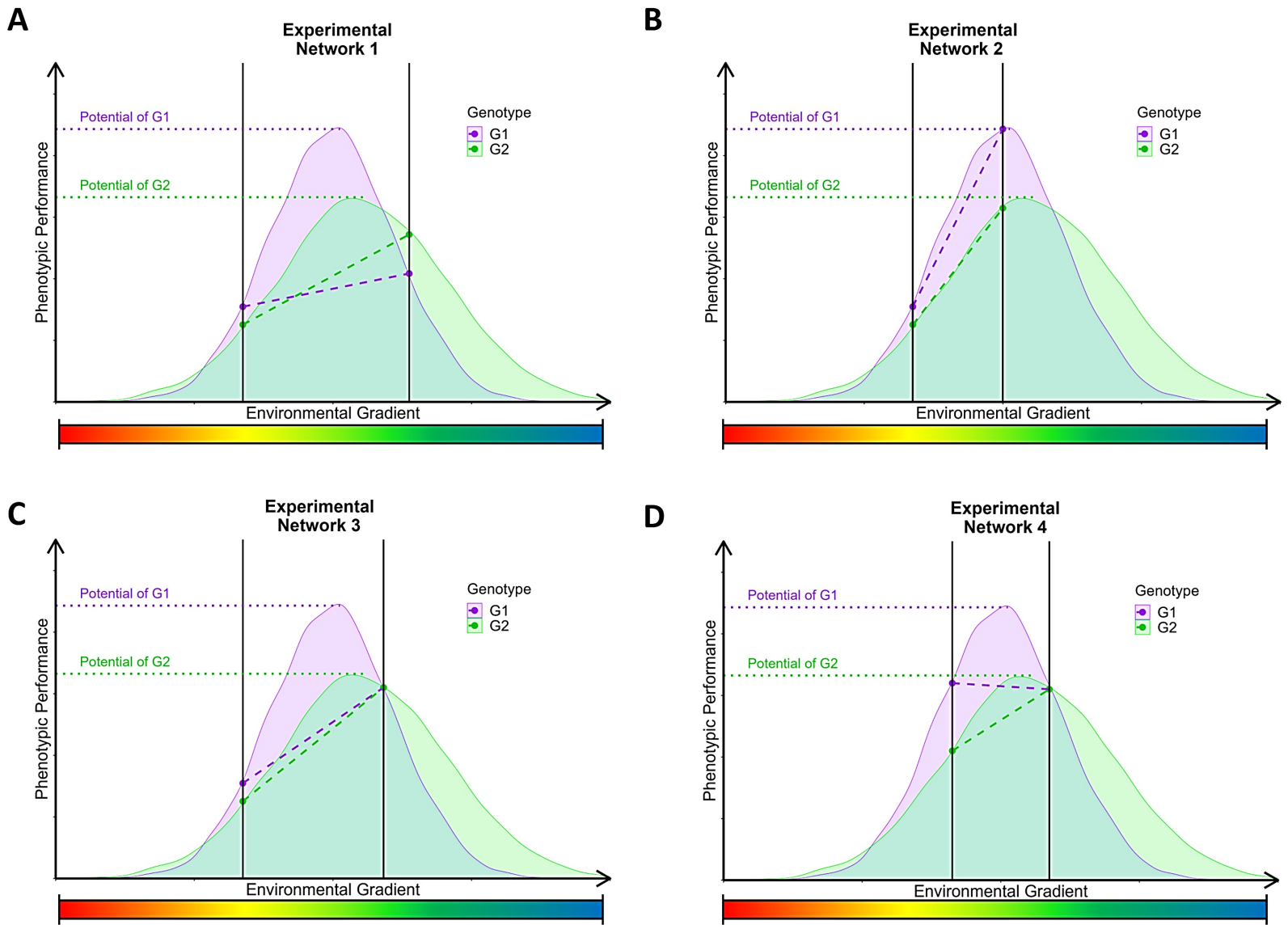
- **Phenotype:** end-result of multiple interactions among expressed genes, epigenome and environmental signals
- Environmental acts at nuclear, celular and physiologycal level (plant level)
- It is not a one-way road: compensatory effects, homeostasis, plasticity, pleiotropy, epigenetics...across time and space!

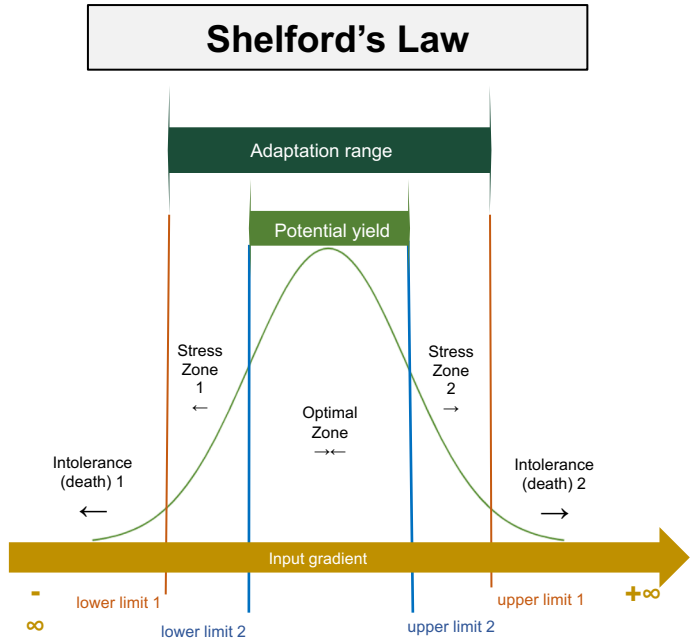
Solid lines: **Central Dogma**
Dashed lines: **Surrounding environment outside gene level**



Genotype x Environment Interaction (GxE) as an emergent property

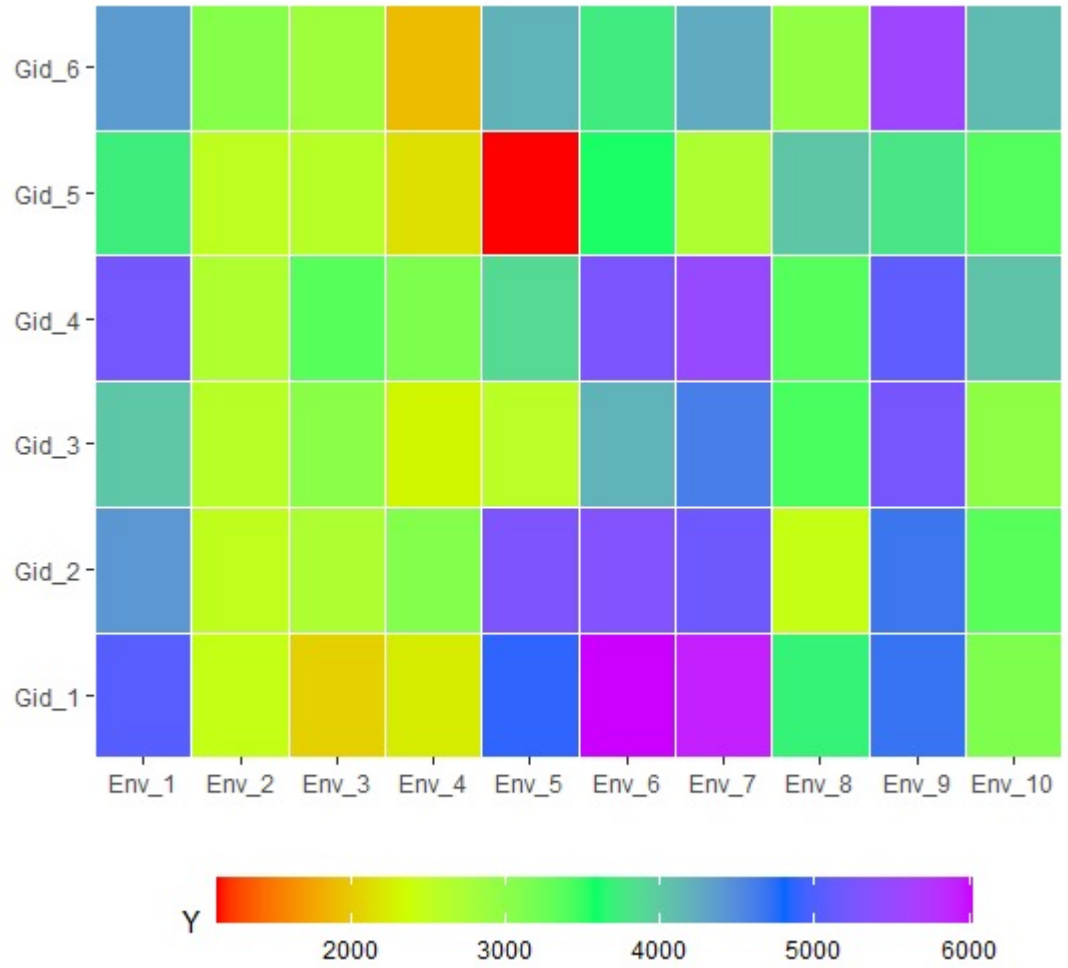
Product of reaction-norm, respecting the particular plasticity for each genotype, expressed across a given experimental gradient





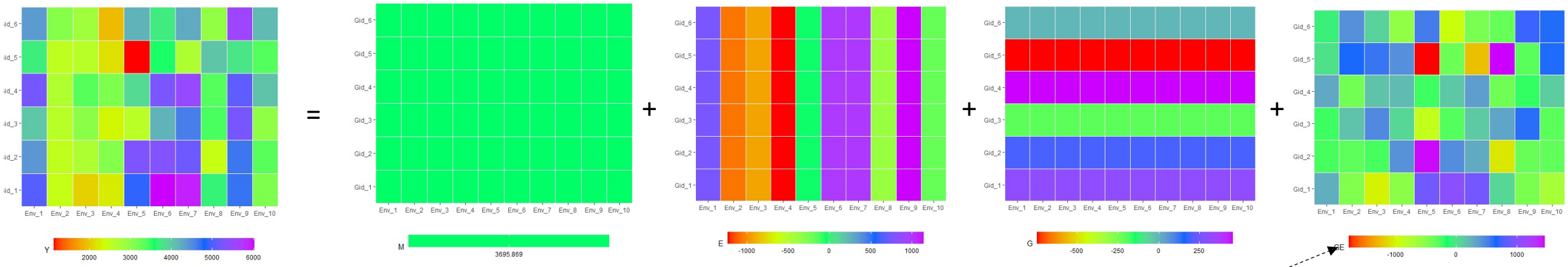
Two-way table (genotypes per environment)

$$\boldsymbol{Y_{pq}} = \begin{bmatrix} \hat{Y}_{11} & \cdots & \hat{Y}_{1q} \\ \vdots & \ddots & \vdots \\ \hat{Y}_{p1} & \cdots & \hat{Y}_{pq} \end{bmatrix} = (\mathbf{1}_{pq} \otimes \boldsymbol{\mu}) + (\mathbf{1}_{p} \otimes \boldsymbol{E}'_{q}) + (\mathbf{1}'_{q} \otimes \boldsymbol{G}_{p}) + \boldsymbol{GE}_{pq}$$



Two-way table (genotypes per environment)

$$\boldsymbol{Y_{pq}} = \begin{bmatrix} \hat{Y}_{11} & \cdots & \hat{Y}_{1q} \\ \vdots & \ddots & \vdots \\ \hat{Y}_{p1} & \cdots & \hat{Y}_{pq} \end{bmatrix} = (\mathbf{1}_{pq} \otimes \boldsymbol{\mu}) + (\mathbf{1}_{p} \otimes \boldsymbol{E}'_{q}) + (\mathbf{1}'_{q} \otimes \boldsymbol{G}_{p}) + \boldsymbol{GE}_{pq}$$



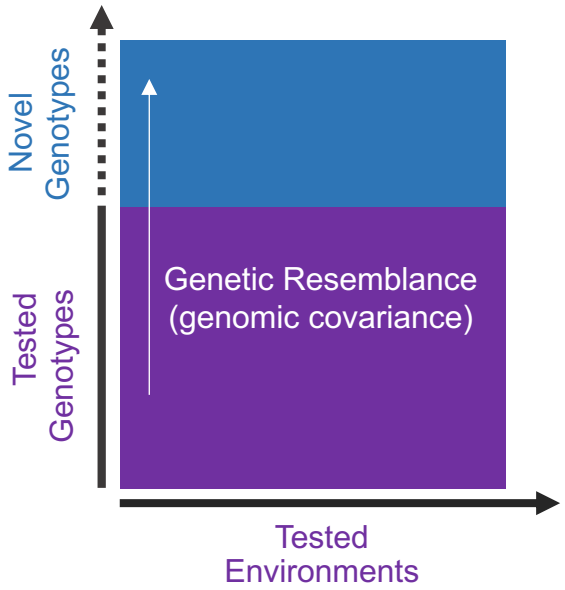
Enviromic (environmental covariables) realized relationship to 'estimate' reaction-norms or assembly environment-to-environment relatedness

Genomic (or Pedigree-based) realized relationship to explore genotype-genotype covariances, and a block-diagonal for mimic the genomic by environment effects

Genomic Prediction (GP)

Some key works:

- Introduced by Meeuwissen et al (2001) aimed for a whole-genome regression over some phenotype
- Burgueño et al (2012) proposed the marker x environment interaction model
- Jarquin et al (2014) and Lopez-Cruz et al (2015) presents the multi-environment GBLUP (using haddamard's product)
- Heslot et al (2014) and Ly et al (2018) presents a factorial regression with GP
- Bandeira e Sousa (2017) presents the use of gaussian kernel (GK) for modeling GxE in maize
- Morais-Júnior et al (2018) expands Jarquin's model to account for quantile covariables
- Crossa et al (2019) and Cuevas et al (2019) presents the use of Deep Kernel for modeling G effects



Vector of observed phenotypes for a given growing environment condition

Approximation of the "true" unknow genetic effects

$$\boldsymbol{y} = 1\boldsymbol{\mu} + \boldsymbol{X_f\beta} + \sum_{s=1}^{k} \boldsymbol{g_s} + \boldsymbol{\varepsilon}$$

"Know" and fixed effects and design structures over *y* variation

Catpures random nongenetic effects (imperfect LD between markers and QTL or model misspecifcation)

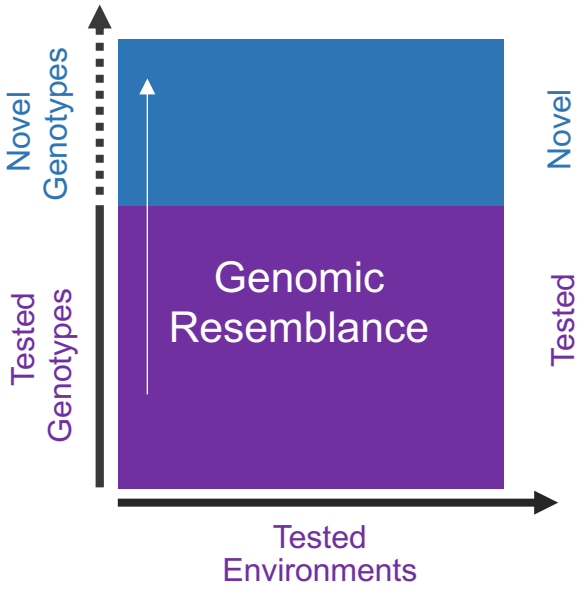
$$\boldsymbol{g_s} \sim N\left(0, \boldsymbol{K}\boldsymbol{\sigma}^2_{\mathbf{g}}\right)$$

Why is *Envirotyping* useful for multi-environment genomic prediction?

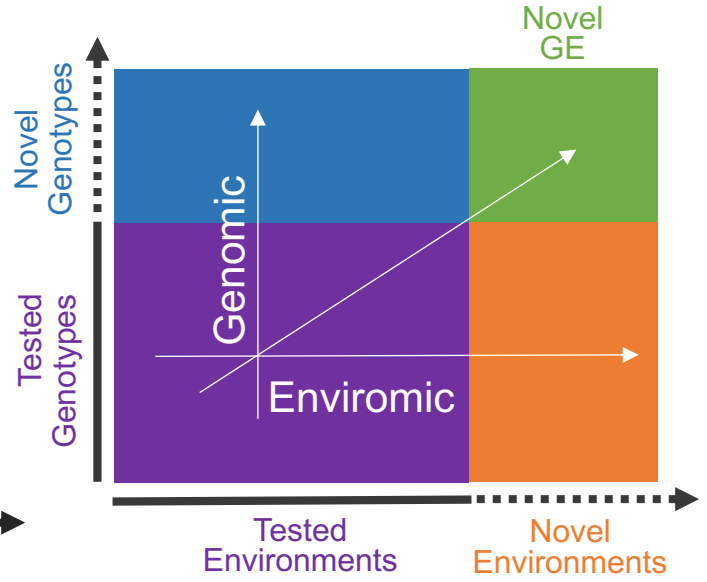
Try to deal with the environmental-phenotype (E2P) covariances in the phenotypic records in the context of MET

- Benchmark GP models were conceived for modeling genotype-phenotype covariances (G2P)
- Different field trials, different E2P; consequently, different G2P pattern (and GP accuracy)
- Multi-environment trials = more source of noise due to E2P?
- Implicit and Explicit covariates can shape E2P and translate some noise in exploitable pattern

MET Prediction (Pure Genomic)



MET Prediction (Enviromic + Genomic)



Adding value to Predictive Breeding

- Model genomic reaction-norms
- Realization of yet-to-be-seem GxE
- Optimize MET for training GP
- Orient cultivar targeting

Adapted from Malosseti et al. (2013)

Models for Multi-Environment Genomic Prediction (MET-GP)

$$y = 1\mu + Z_E\beta + \sum_{s=1}^{k} g_s + \sum_{s=1}^{k} gE_s + \sum_{r=1}^{l} W_r + \sum_{s=1}^{k}\sum_{r=1}^{l} gW_{sr} + \varepsilon$$

$$y = [y_1, y_2, \ldots y_q]^T$$

Model	Effect			
	Genetic	Environment	GxE	
Main Genomic Effects (MM)	$\sum_{s=1}^{p} g_s \neq 0$	$\sum_{r=1}^{l} W_r = 0$	$\sum gE_s = 0$	$\sum\sum gW_r = 0$
MM + single GxE deviation (MDs)	$\sum_{s=1}^{p} g_s \neq 0$	$\sum_{r=1}^{l} W_r = 0$	$\sum gE_s \neq 0$	$\sum\sum gW_r = 0$
MM + main Enviromic effecs (EMM)	$\sum_{s=1}^{p} g_s \neq 0$	$\sum_{r=1}^{l} W_r \neq 0$	$\sum gE_s = 0$	$\sum\sum gW_r = 0$
EMM + single GxE deviation (EMDs)	$\sum_{s=1}^{p} g_s \neq 0$	$\sum_{r=1}^{l} W_r \neq 0$	$\sum gE_s \neq 0$	$\sum\sum gW_r = 0$
EMM + reaction norm GxW (RNMM)	$\sum_{s=1}^{p} g_s \neq 0$	$\sum_{r=1}^{l} W_r \neq 0$	$\sum gE_s = 0$	$\sum\sum gW_r \neq 0$
EMDs + reaction norm GxW (RNMDs)	$\sum_{s=1}^{p} g_s \neq 0$	$\sum_{r=1}^{l} W_r \neq 0$	$\sum gE_s \neq 0$	$\sum\sum gW_r \neq 0$

Implementation

- Independent Residuals = I
- Hierarchical Bayesian Modeling
- BGGE (now within EnvRtype)

Phenology (example for maize)

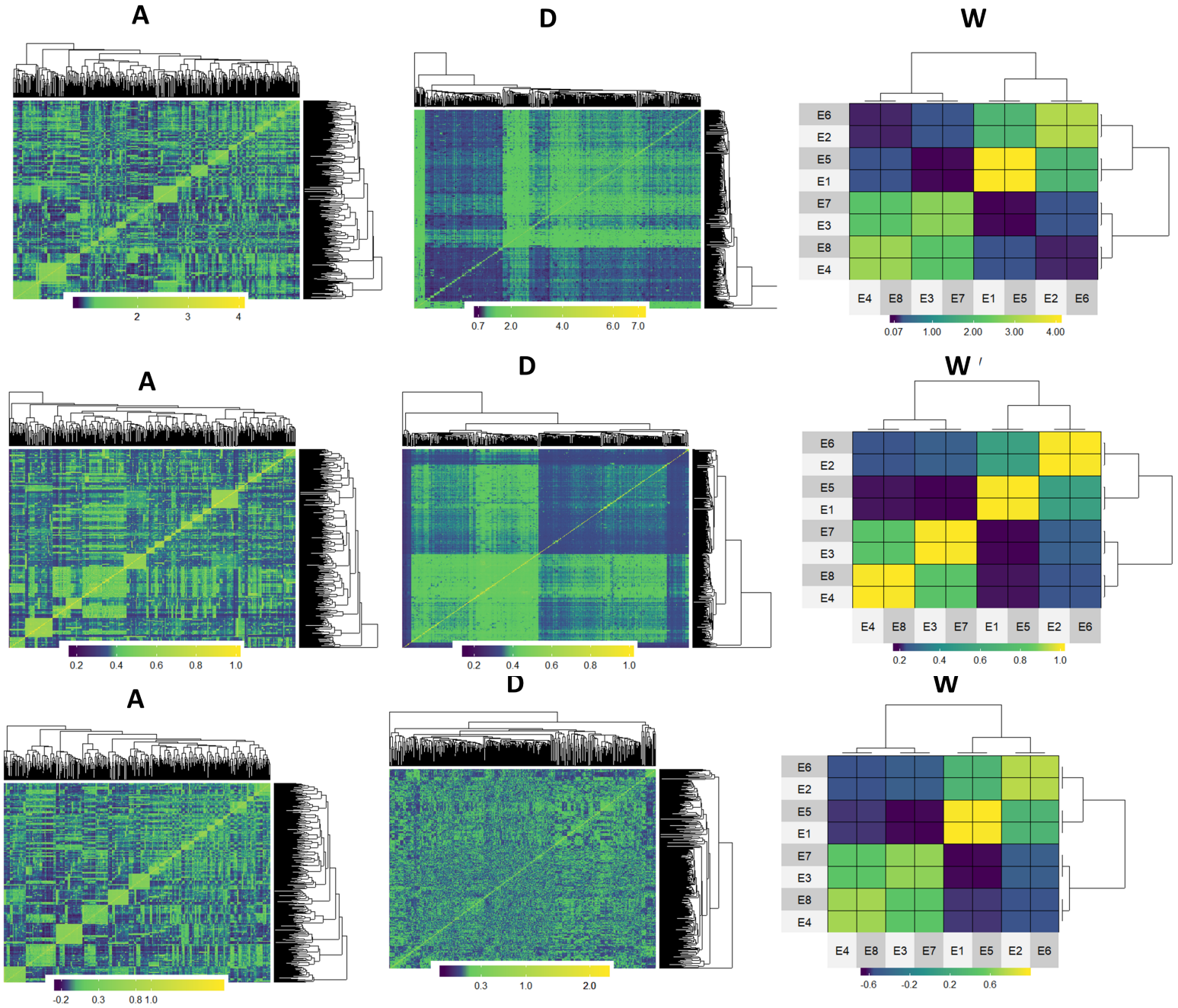
- emergence to the first leaf (V1, 14 DAE)
- V1 to the fourth leaf (V4, 35 DAE)
- V4 to the tasseling stage (VT, 65 DAE)
- VT to the kernel milk stage (R3, 90 DAE)
- R3 to physiological maturity (R6, 120 DAE)

Kernel Methods (GB, GK and DK)

Benchmark GBLUP
(linear covariances)

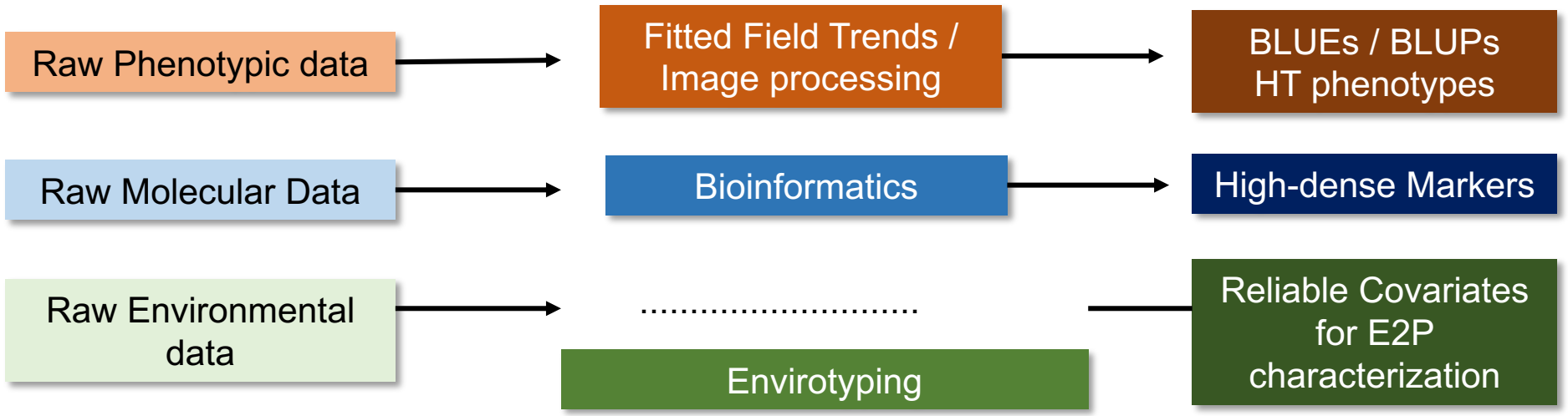
Gaussian Kernel
(distances)

Deep Kernel based
on Arc-cosine
estimation (nonlinear)



Why EnvRtype?

- Availability vs Usefulness of Environmental Data
- Increase efficiency without increase costs of field trials
- Use any environmental information ?
- A clearly definition of how envirotyping can be exploited in predictive breeding
- From envirotyping we acess the ‘envirome’ of the crop and better ‘explain’ some of the nongenetic effects

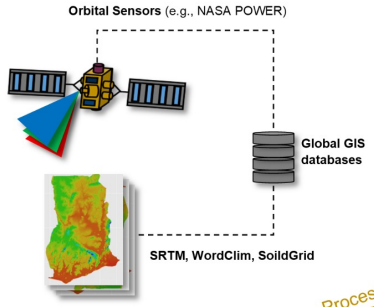


Module 1

Remote Data Collection

- get_weather()
- extract_GIS()

Raw-Data



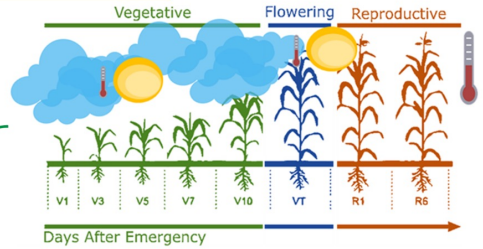
Raw-data from environmental sensors or get_weather() and extract_GIS() can be used in the further steps or processed

Raw-data Processing

- SummaryWTH()
- ProcessWTH()
- param_temperature()
- param_radiation()
- param_atmospheric()

Data processing involves the quality control and computation of additional variables

Processed-Data



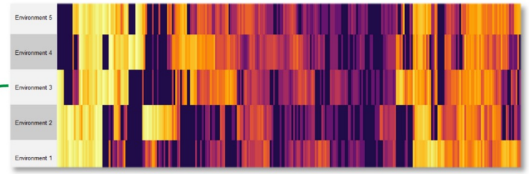
Divide time intervals in crop development helps to characterize enviroments and discover a panel of envirotype descriptors

Module 2

Environmental Characterization

- env_typing()
- W_matrix()

Envirotype descriptors



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

Environmental covariables

Module 3

Environmental Similarity

- env_kernel()

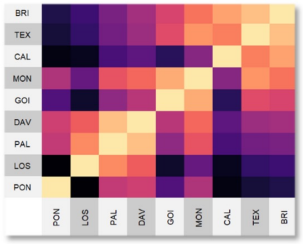
Enviromic Kernels

Enviromic-based Kernels Models

- get_kernel()
- kernel_model()

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

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



Models:

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

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

Remote Sensing (Digital Envirotyping)

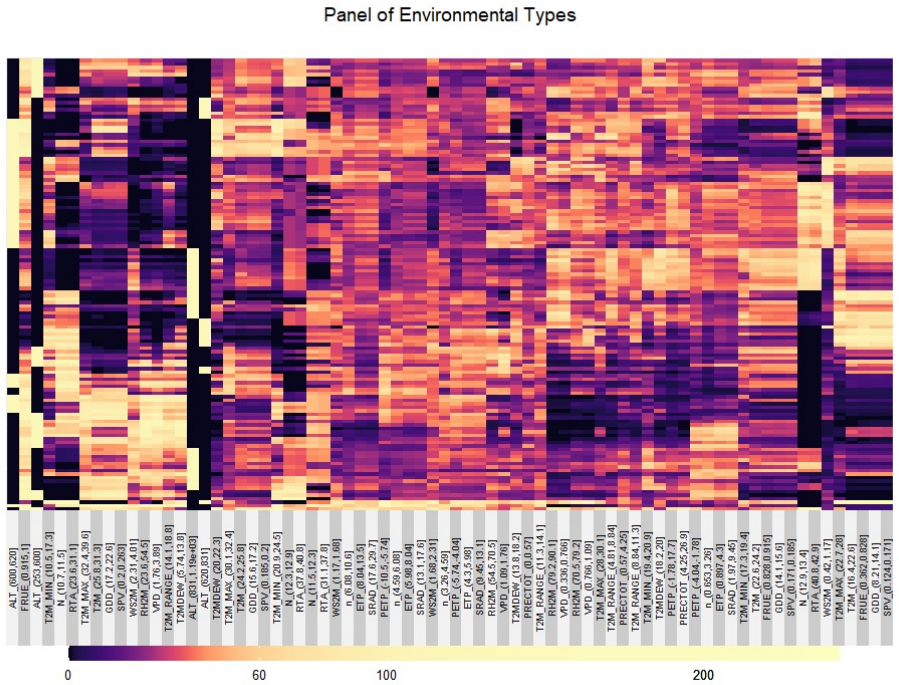
Source	ID	Environmental Factor	Unit
Nasa Power[1]	ALLSKY_SFC_SW_DWN	All sky insolation incident on a horizontal surface	MJ m^{-2} d^{-1}
	ALLSKY_SFC_LW_DWN	Thermal infrared longwave radiative flux	MJ m^{-2} d^{-1}
	WS2M	Wind speed at 10 m above the surface of the earth	m s^{-1}
	T2M_MIN	Minimum air temperature at 2 m above the surface of the earth	°C d^{-1}
	T2M_MAX	Maximum air temperature at 2 m above the surface of the earth	°C d^{-1}
	T2MDEW	Dew-point temperature at 2 m above the surface of the earth	°C d^{-1}
	RH2M	Relative air humidity at 2 m above the surface of the earth	%
	PRECTOT	Rainfall precipitation (P)	mm d^{-1}
SRTM[2]	ALT	Elevation (above sea level)	m
Computed[3]	FRUE	Effect of Temperature on Radiation use Efficiency	-
	GDD	Growing Degree-days	C° d^{-1}
	ETP	Evapotranspiration (ETP)	mm d^{-1}
	PETP	Atmospheric water deficit P-ETP	mm d^{-1}
	DVP	Deficit of vapor pressure	kPa d^{-1}
	SVP	Slope of saturation vapor pressure curve	kPa C° d^{-1}
	T2M_RANGE	Temperature Range	°C d^{-1}
	RTA	Global Solar Radiation based on Latitude and Julian Day	MJ m^{-2} d^{-1}

Worldwide soil data : https://soilgrids.org/

Cardinals for temperature (from literature)

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

Environmental Matrix (W_matrix or env_typing)



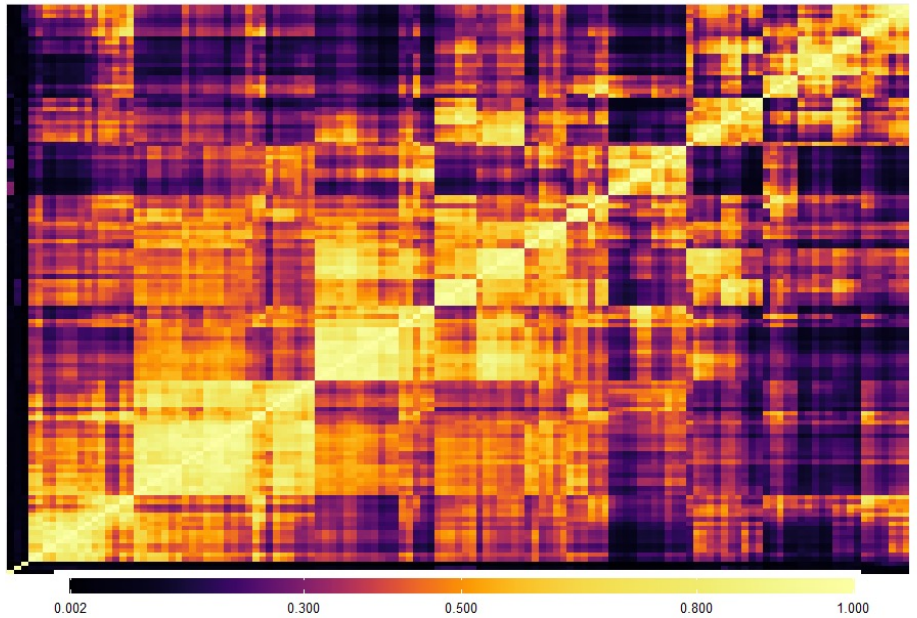
Used as Explicit Covariates

Kernelization

(env_kernel)

1. Linear Kernel (same as GBLUP)
2. Nonlinear Gaussian Kernel (GK)
3. Nonlinear Deep Kernel (Arc-cosine approach)

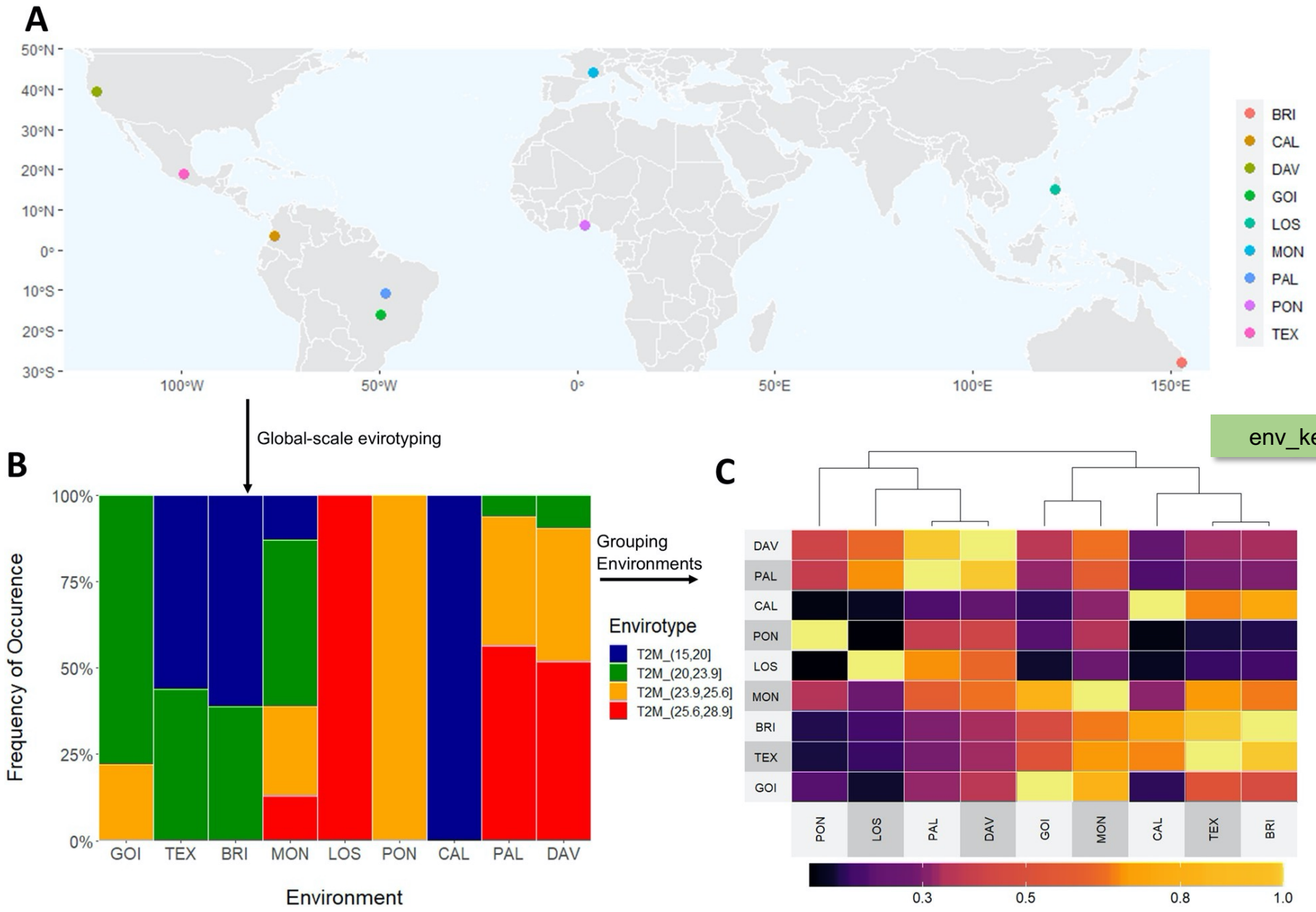
Environmental Relatedness



Approaching Similarity

(in silico)

Global Envirotyping Network



env_typing()

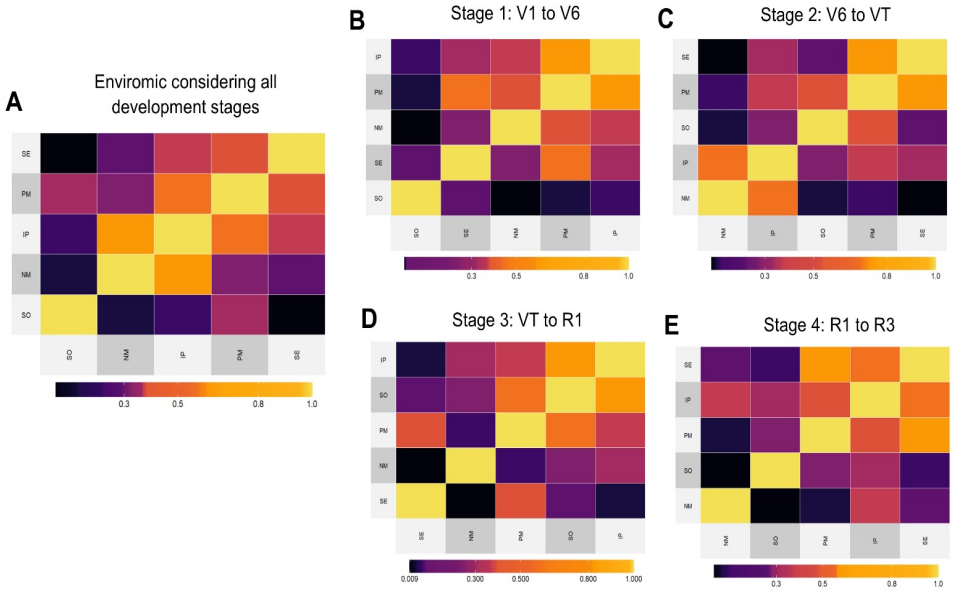
env_kernel(gaussian=TRUE)

“New” reaction-norm approaches

GK+different model structures

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

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



“New” reaction-norm approaches

GK+specific or joint covariates effects

$$y = 1\mu + X_f\beta + g + gE + EC + gEC + \varepsilon$$

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

Final remarks

“[…] Enviromics is like a chocolate box: you never know what you gonna get”