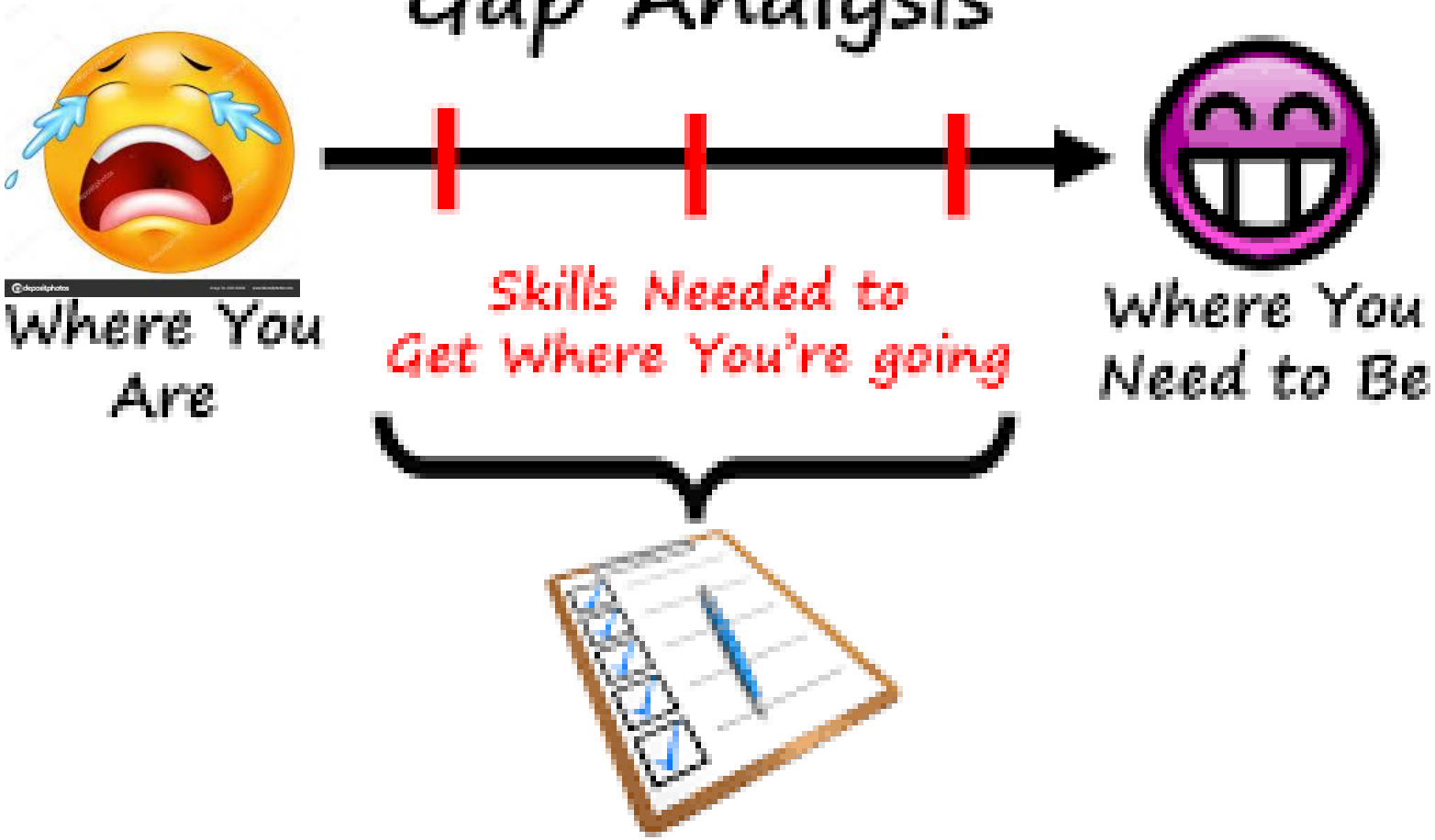
BIG data course: Dec 2018

Rabat, Morocco

Z.kehel@cigar.org


BIG
DATA
& ANALYTICS

Gap Analysis



Just to clarify!

Big data is a collection of data from ***traditional*** and ***digital*** sources ***inside*** and ***outside*** your company that represents a source for ***ongoing discovery and analysis***.

In defining big data, it's also important to understand the ***mix of unstructured and multi-structured data.***

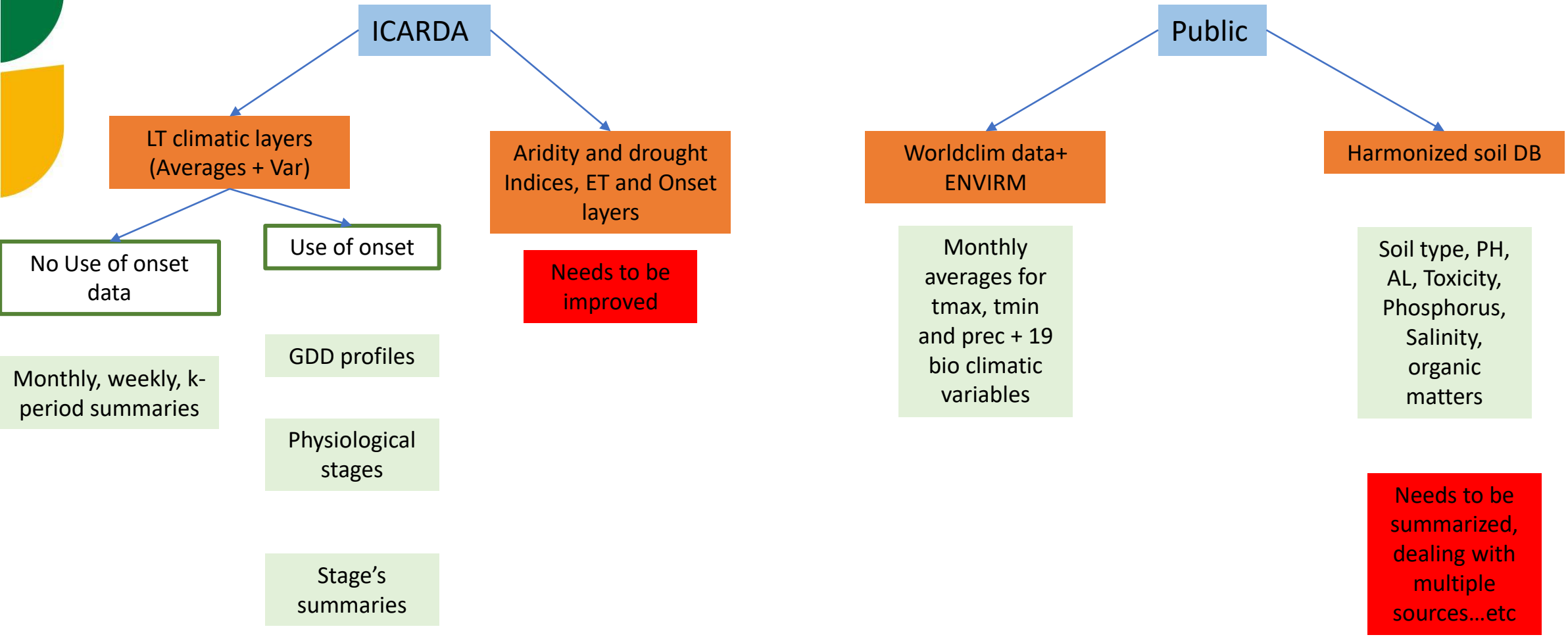
Every organization needs to ***fully understand big data*** (depending on the needs)– what it is to them, what is does for them, what it means to them –and the potential of ***data-driven decision making.*** Don't wait. Waiting will only delay the inevitable and make it even more difficult to unravel the confusion.

www.forbes.com (accessed Dec2018)

Data Cycle at GRS (Passport, Eval, Charac)

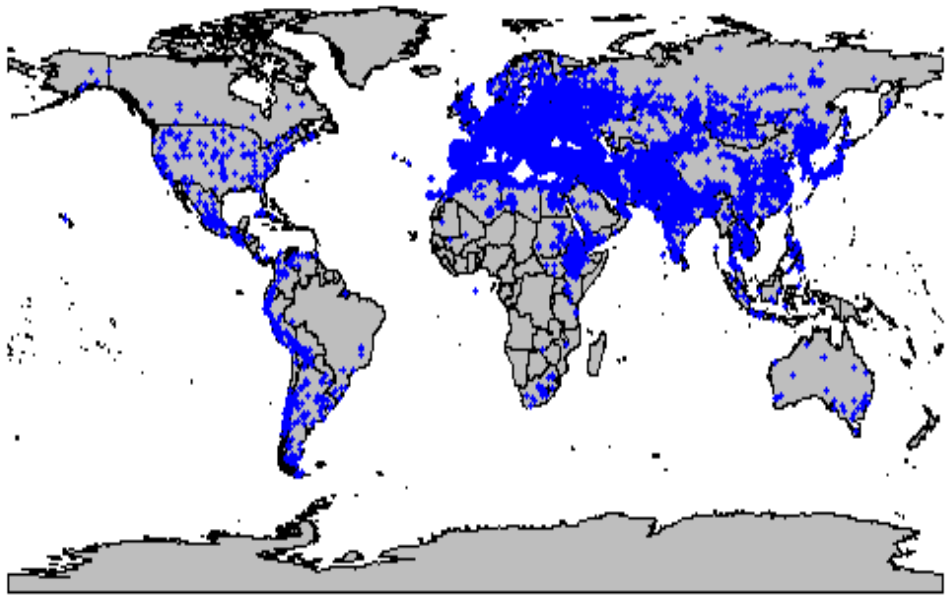
- Passport data and collecting information
- SMTA
- Characterization and evaluation
- Molecular data
- Reporting

Available data: Summary



Available data

- Passport and Daily long-term climatic surfaces; averages and variations over 30 years (GU)



a. temp---temperature at 2 meter (degC)
b. tmax---maximum temperature at 2 meter (degC)
c. tmin---minimum temperature at 2 meter (degC)
d. precip---precipitation (mm)
e. ABSH---absolute humidity (kg/m**3 scaled by 10**6)
f. RHY---relative humidity (%)
g. PAR---photosynthesis active radiation (mol PPFD=2.05*MJ)
h. uwind---wind at east-west direction(m/s)
i. vwind---wind at north-south direction(m/s)
j. VPD---vapor pressure deficit(Pa)

Available data

- Phenotypes + Markers

Morphology traits	Trait Type	No. of experiments	No. of accessions	No. of unique accessions
Growth habit	Categorical	79	73,912	29,699
Awness	Categorical	72	74,512	33,276
Waxiness of plant	Categorical	73	74,417	33,305
Spike length	Quantitative	69	63,157	21,865
Spike density	Categorical	76	86,274	33,172
Seed color	Categorical	12	21,626	10,729
Stem solidness	Categorical	54	48,447	21,432
Number of spikelets groups per spike	Quantitative	91	94,256	33,745

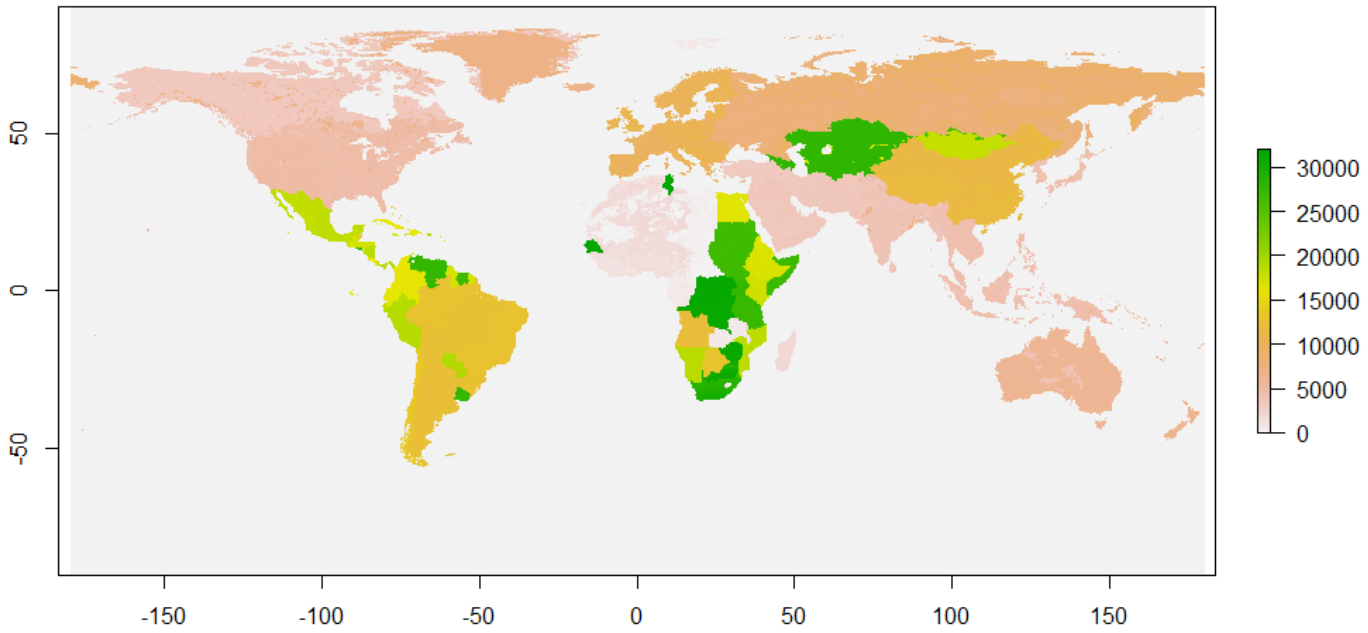
Agronomy traits	Trait Type	No. of experiments	No. of accessions	No. of unique accessions
Early growth vigour	Categorical	11	19,089	9,014
Productive tillering capacity	Categorical	63	72,625	33,295
Plant height	Quantitative	81	88,094	33,379
Number of kernels per spike	Quantitative	70	82,007	31,191
1000 kernel weight	Quantitative	75	83,041	31,756
Seed protein content	Quantitative	3	22,118	11,028
Vitreousness	Quantitative	1	9,325	9,325
Grain yield per plot	Quantitative	57	53,663	19,382
Lodging resistance	Categorical	68	82,049	31,755
Agronomic score	Categorical	12	19,679	8,720

Phenology traits	Trait Type	No. of experiments	No. of accessions	No. of unique accession
Days to heading	Quantitative	87	93,275	33,395
Days to maturity	Quantitative	86	89,031	33,339
Grain filling period	Quantitative	61	85,886	33,297
Stresses traits				
Yellow rust severity of infection	Categorical	11	15,529	7,685
Common bunt resistance	Categorical	1	13,984	7,685
Septoria resistance	Categorical	1	8,176	8,176
Yellow rust severity and infection type		26	19,288	11286
Yellow rust coefficient of infection	Quantitative	14	9,580	5,319
Yellow rust infection type	Categorical	24	15,453	8,360
Cold tolerance	Categorical	12	15,440	7,522

Focused Identification of Germplasm Strategy (FIGS)

Available data

- Maps of soil characteristics from the harmonized world soil database HWSD



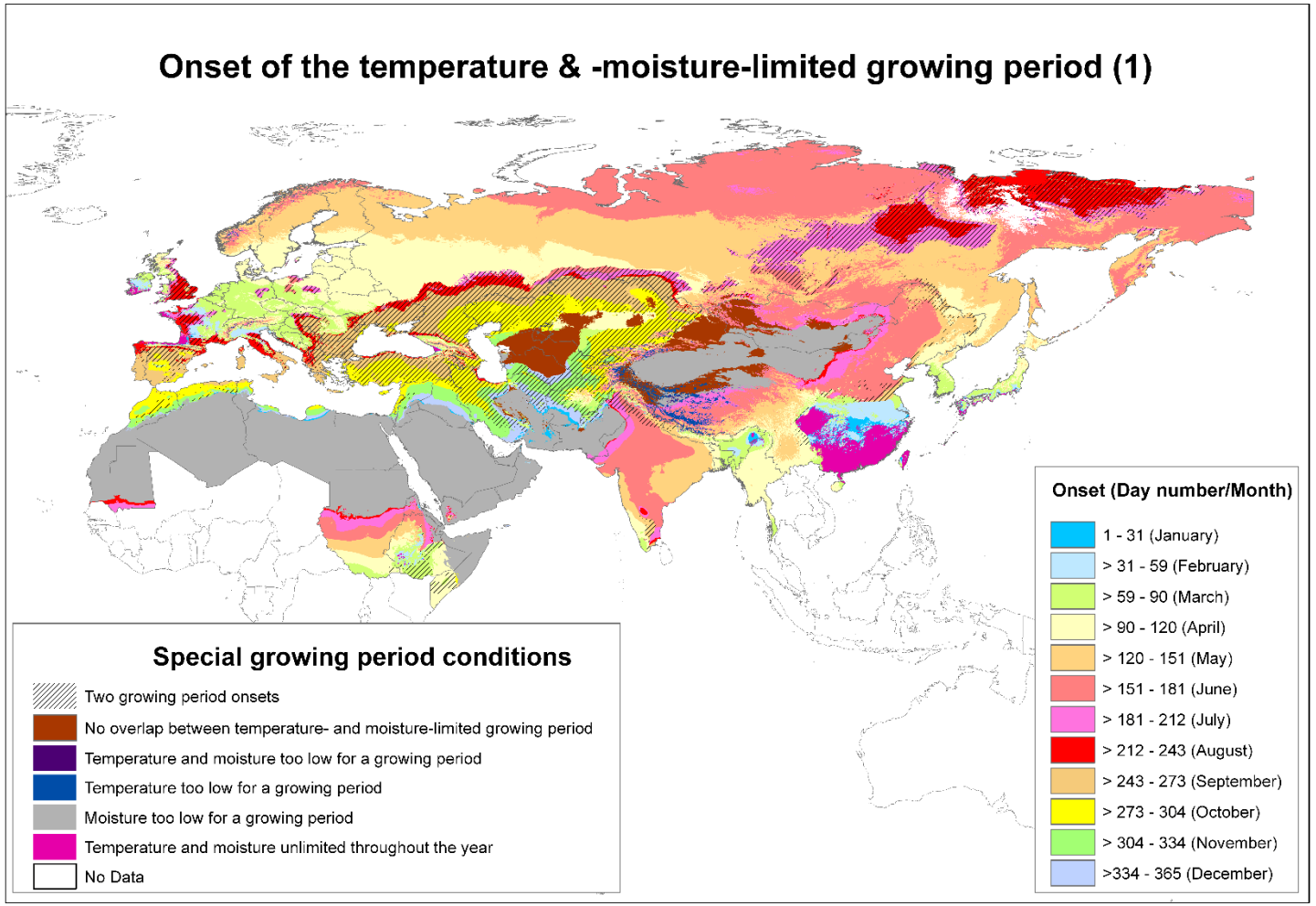
Variable	UNIT	DESCRIPTION
T_GRAVEL	%	Topsoil Gravel Content
T_SAND	%	Topsoil Sand Fraction
T_SILT	%	Topsoil Silt Fraction
T_CLAY	%	Topsoil Clay Fraction
T_USDA_TEX_CLASS		Topsoil USDA Texture Classification
T_REF_BULK_DENSITY	kg/dm3	Topsoil Reference Bulk Density
T_BULK_DENSITY	kg/dm3	Topsoil Bulk Density
T_OC	% weight	Topsoil Organic Carbon
T_PH_H2O		Topsoil pH (H2O)
T_CEC_CLAY	cmol/kg	Topsoil CEC (clay)
T_CEC_SOIL	cmol/kg	Topsoil CEC (soil)
T_BS	%	Topsoil Base Saturation
T_TEB	cmol/kg	Topsoil TEB
T_CACO3	% weight	Topsoil Calcium Carbonate
T_CASO4	% weight	Topsoil Gypsum
T_ESP	%	Topsoil Sodicity (ESP)
T_ECE	dS/m	Topsoil Salinity (ECe)
S_GRAVEL	%	Subsoil Gravel Content
S_SAND	%	Subsoil Sand Fraction
S_SILT	%	Subsoil Silt Fraction
S_CLAY	%	Subsoil Clay Fraction
S_USDA_TEX_CLASS		Subsoil USDA Texture Classification
S_REF_BULK_DENSITY	kg/dm3	Subsoil Reference Bulk Density
S_BULK_DENSITY	kg/dm3	Subsoil Bulk Density
S_OC	% weight	Subsoil Organic Carbon
S_PH_H2O		Subsoil pH (H2O)
S_CEC_CLAY	cmol/kg	Subsoil CEC (clay)
S_CEC_SOIL	cmol/kg	Subsoil CEC (soil)
S_BS	%	Subsoil Base Saturation
S_TEB	cmol/kg	Subsoil TEB
S_CACO3	% weight	Subsoil Calcium Carbonate
S_CASO4	% weight	Subsoil Gypsum
S_ESP	%	Subsoil Sodicity (ESP)
S_ECE	dS/m	Subsoil Salinity (ECe)

source: http://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/harmonized-world-soil-database-v12/en/

Focused Identification of Germplasm Strategy (FIGS)

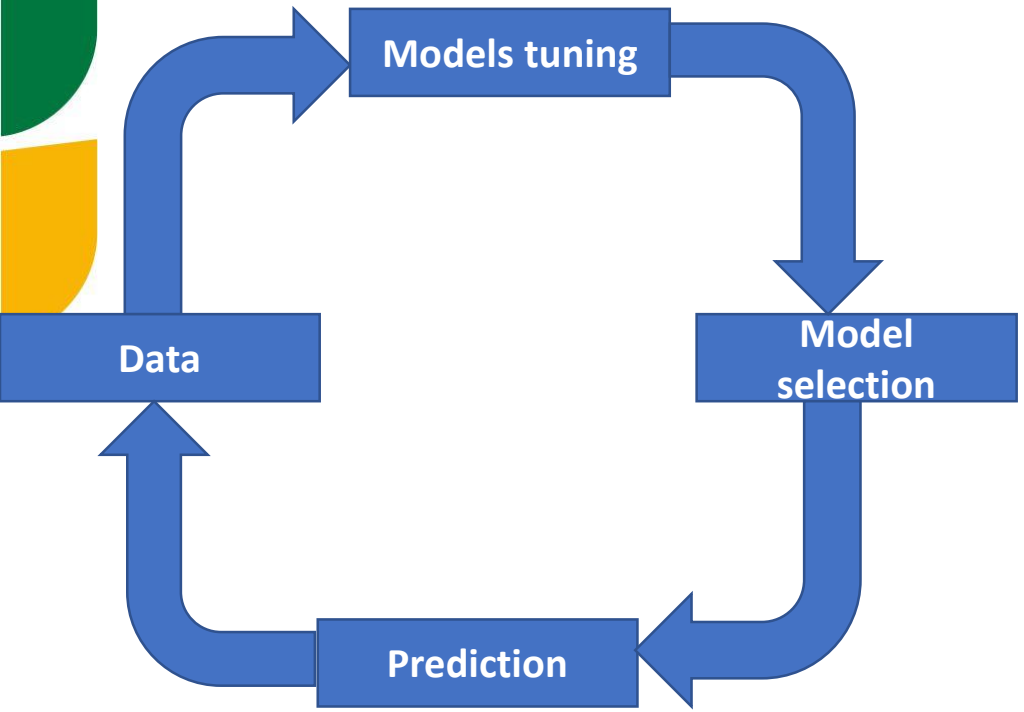
Available data

- Onset surfaces for ICARDA mandate crops, aridity and drought indices



Predictive characterization at ICARDA using FIGS

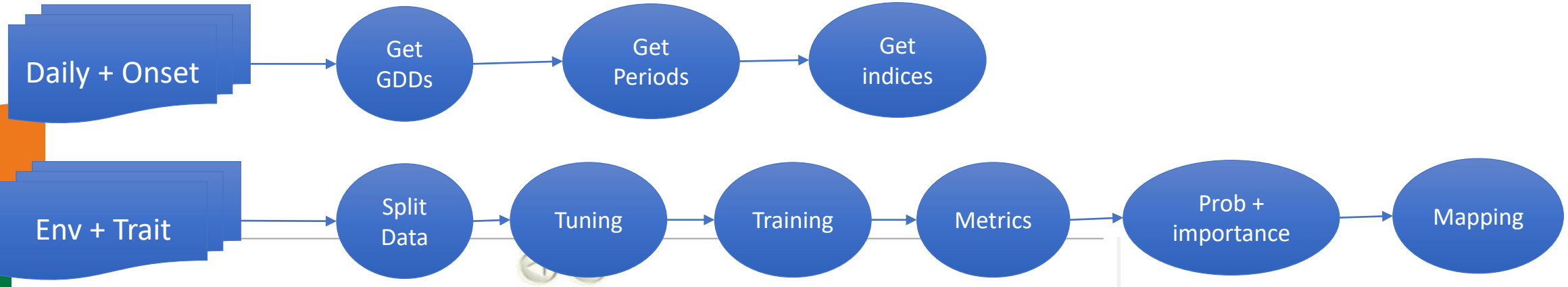
Durum wheat Grain filling period example



FIGS modeling pathway

Performance Measures	k-Nearest Neighbours	Random Forest	Support Vector Machine
Accuracy	0.834	0.838	0.817
95% CI	(0.799, 0.865)	(0.804, 0.868)	(0.781, 0.849)
No Information Rate	0.762	0.762	0.762
P-Value [Acc > NIR]	3.58E-05	1.37E-05	0.001423371
Kappa	0.563	0.557	0.467
Sensitivity	0.722	0.675	0.54
Specificity	0.869	0.889	0.903
True Positive	91	85	68
True Negative	351	359	365
False Positive	53	45	39
False Negative	35	41	58

High accuracy showing that there is a strong relationship between GFP and longterm climatic conditions



Documentation for package 'icardaFIGS' version 0.1.0

- DESCRIPTION file.

Help Pages

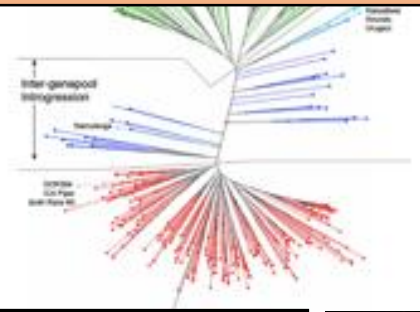
durumDaily	DATASET_TITLE
durumWC	DATASET_TITLE
extractWCdata	Extracting WorldClim Data
getAccessions	Getting Accession Data from ICARDA's Genebank Documentation System
getCrops	Getting List of Crops Available in ICARDA's Genebank Documentation System
getDaily	Extracting Daily Values of Climatic Variables from ICARDA Data
getGrowthPeriod	Calculating Growing Degree Days and Lengths of Growth Stages for Various Crops Using Onset Data from ICARDA's Database
getMetrics	Obtaining Performance Measures from Confusion Matrix
getMetricsPCA	Obtaining Performance Measures from Confusion Matrix for algorithms with PCA pre-processing
getOnset	Extracting Daily Values of Climatic Variables from ICARDA Data Based on Onset of Planting
getTraits	Getting Traits Associated with Crops from ICARDA's Genebank Documentation System
getTraitsData	Getting Values of Accessions for Associated Traits
hello	Hello, World!
mapAccessions	Plotting Accessions on a World Map
septoriaDurumWC	DATASET_TITLE
splitData	Splitting Data into Train and Test Sets
tuneTrain	Splitting the Data, Tuning and Training the Data, and Making Predictions
varimpPred	Calculating Variable Importance and Making Predictions

ICARDA-FIGS-R package

Landrace gap spatial analysis methodology

1 Review literature on landrace genetic structure and its relationship with ecogeography

Knowledge on crop genetic and ecogeographic variation



2 Gather passport data and other data relevant to genetic structure (e.g. characterization, names, molecular)

Database (passport, characterization, genetic)

Genesys

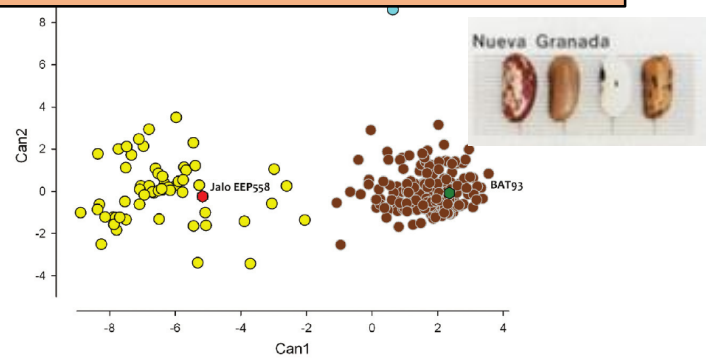
GRIN-Global

GBIF Global Biodiversity Information Facility

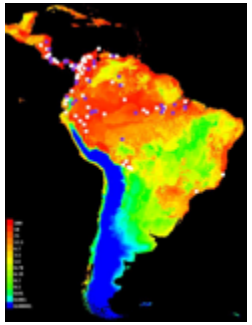
3 Classify dataset into its structural groups:

3.1. Define levels & sub-groups
3.2 Assess predictability
3.3 Classify at chosen level

Classified occurrence dataset



4 Model geographic distributions of each subgroup



Probability distribution of landraces in given sub-group

5 Calculate gap score for each subgroup:

5.1 Geographic score
5.2 Environmental score
5.3 Combined gap score

Maps of geographic, environmental and gaps

6 Expert evaluation of gap analysis results

Validated / updated results

CIAT

LGA *S. phureja*: Species distribution model

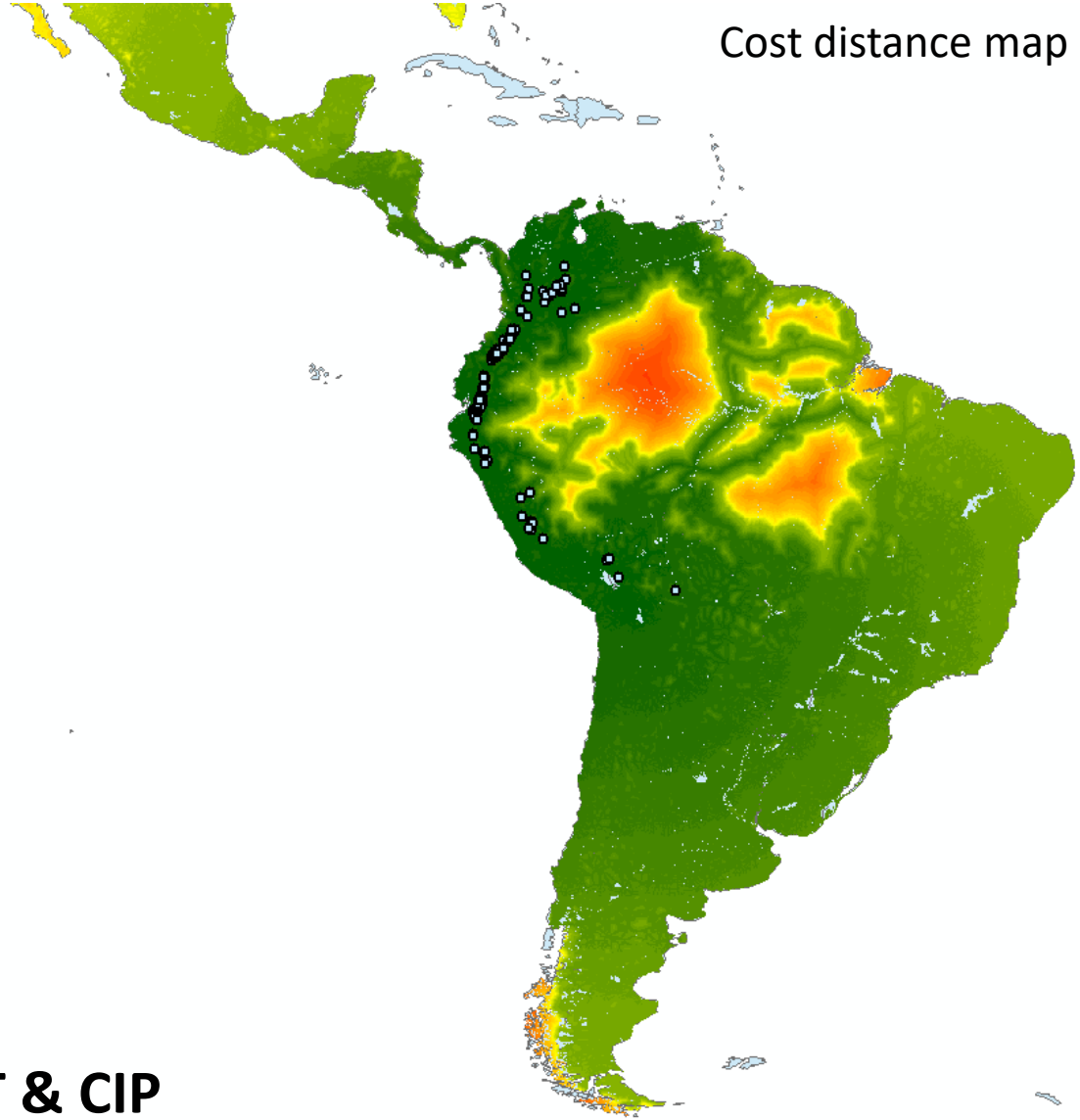
Occ map

SDM map

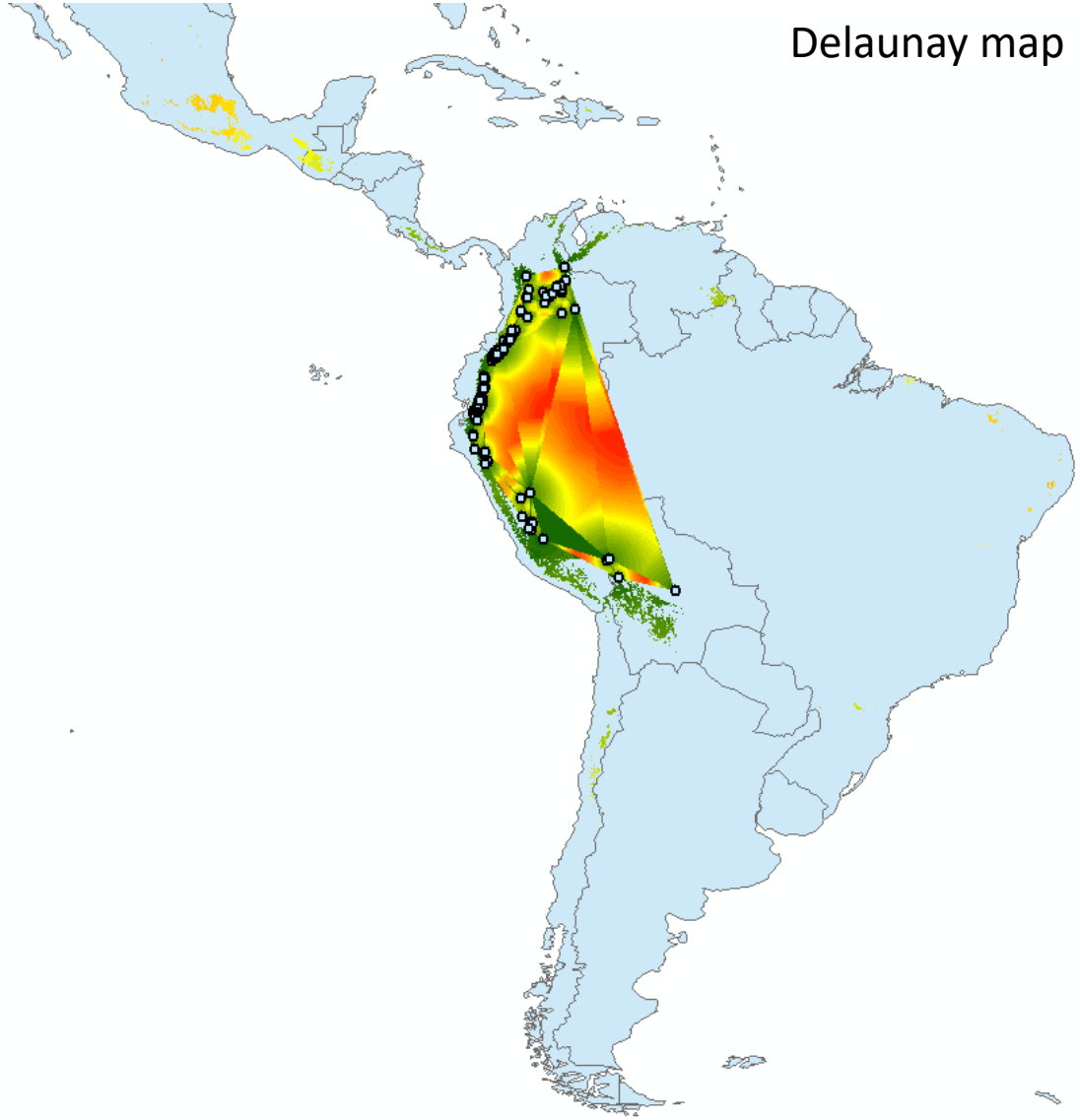
CIAT & CIP

LGA *S. phureja*: Geographic scores

Cost distance map



Delaunay map



CIAT & CIP

LGA *S. phureja*: Environmental score

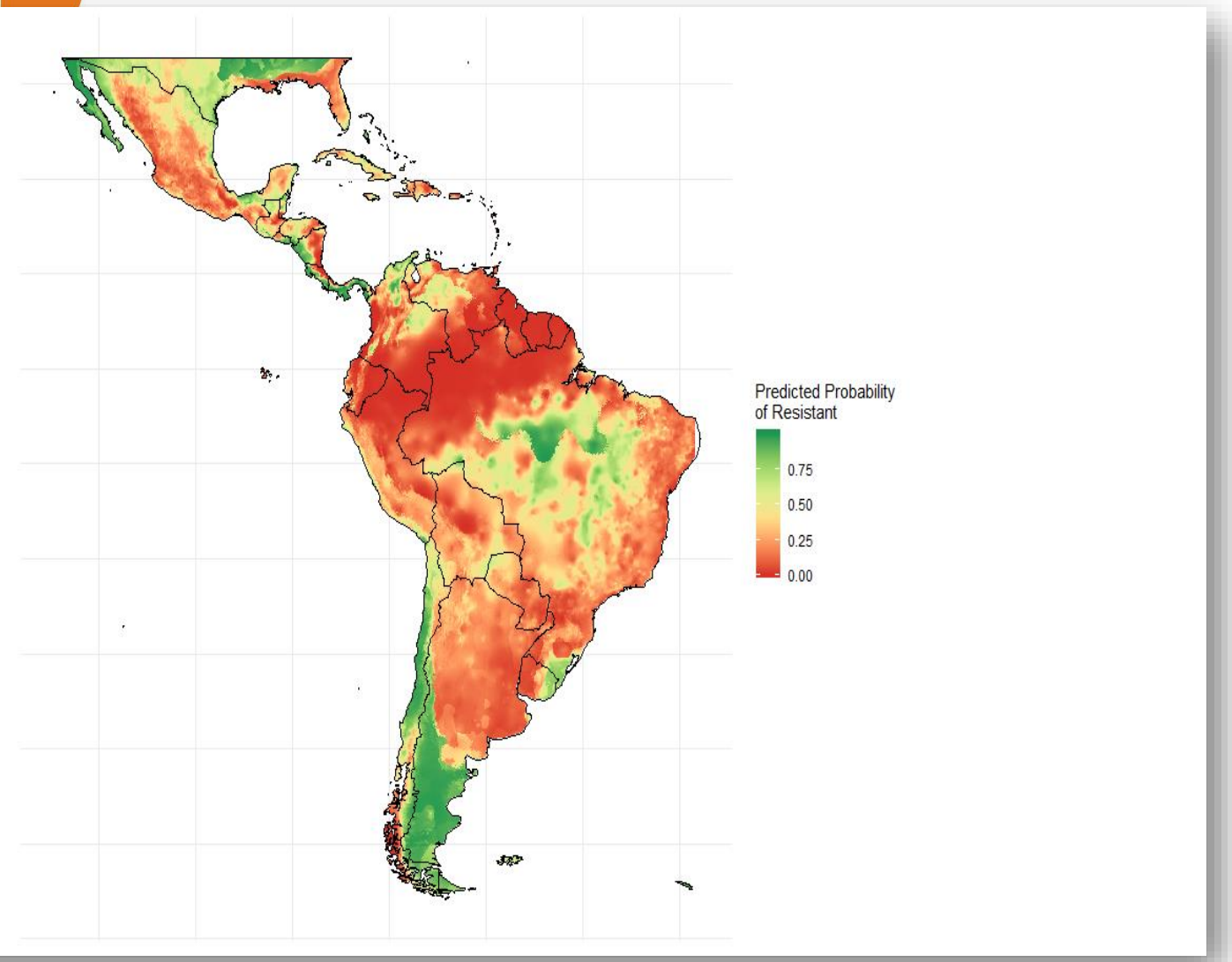
CIAT & CIP

LGA *S. phureja*: Gap score

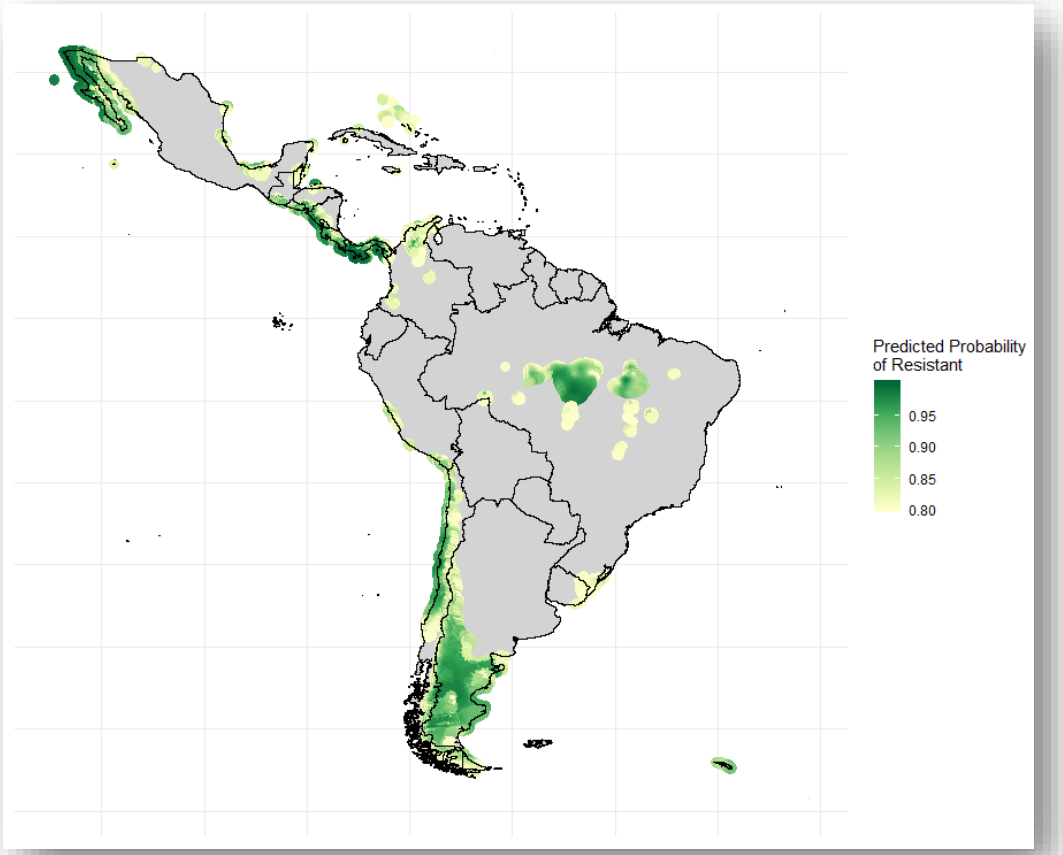
Cost distance approach

Delaunay approach

CIAT & CIP


Predicted Probability
of Resistant
0.75
0.50
0.25
0.00

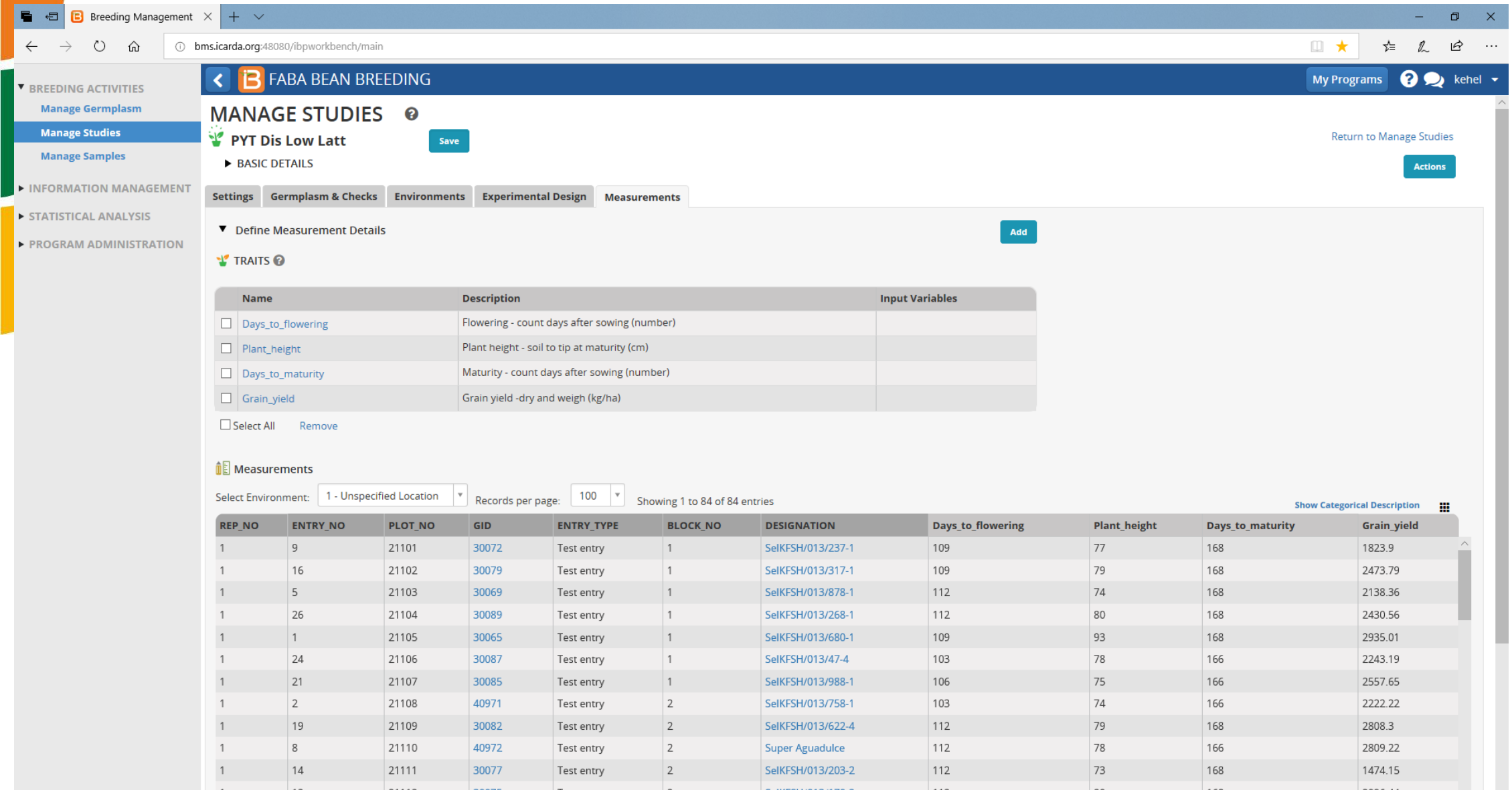
- Prob >= 0.8


Predicted Probability
of Resistant
0.95
0.90
0.85
0.80

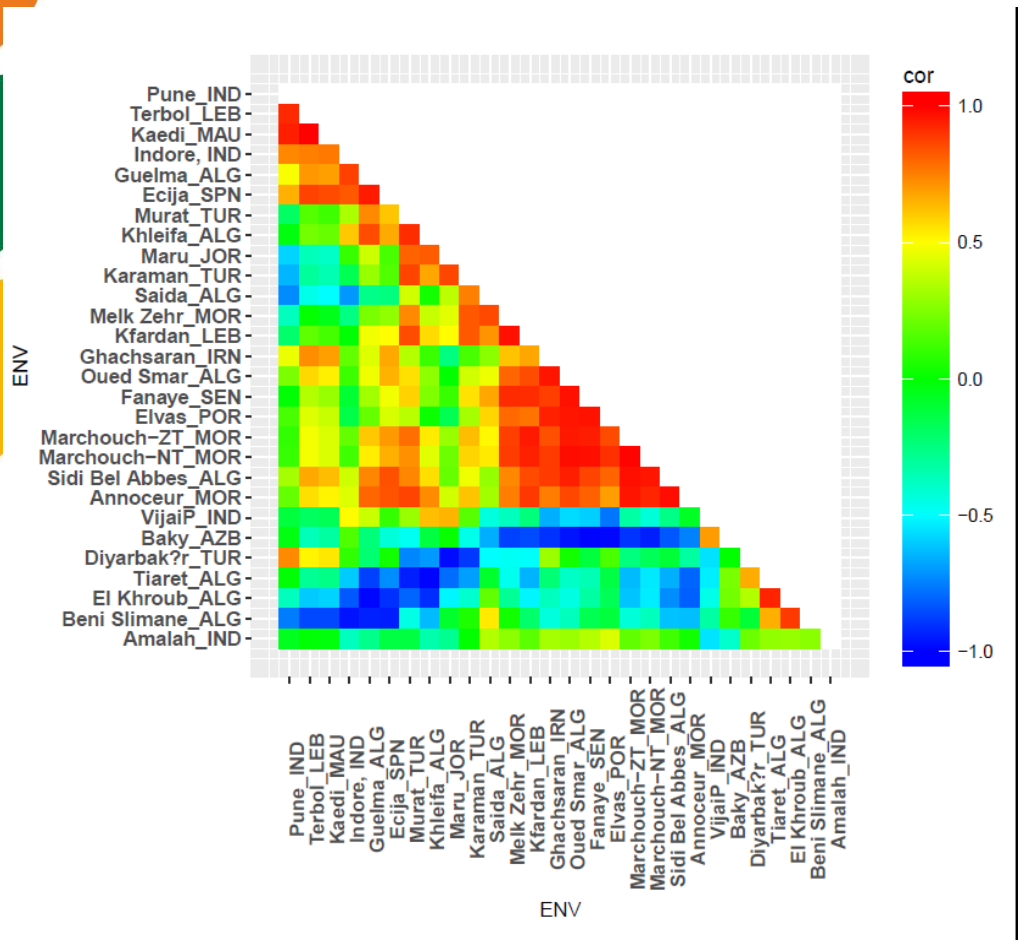
Data Cycle at BCI (Pheno)

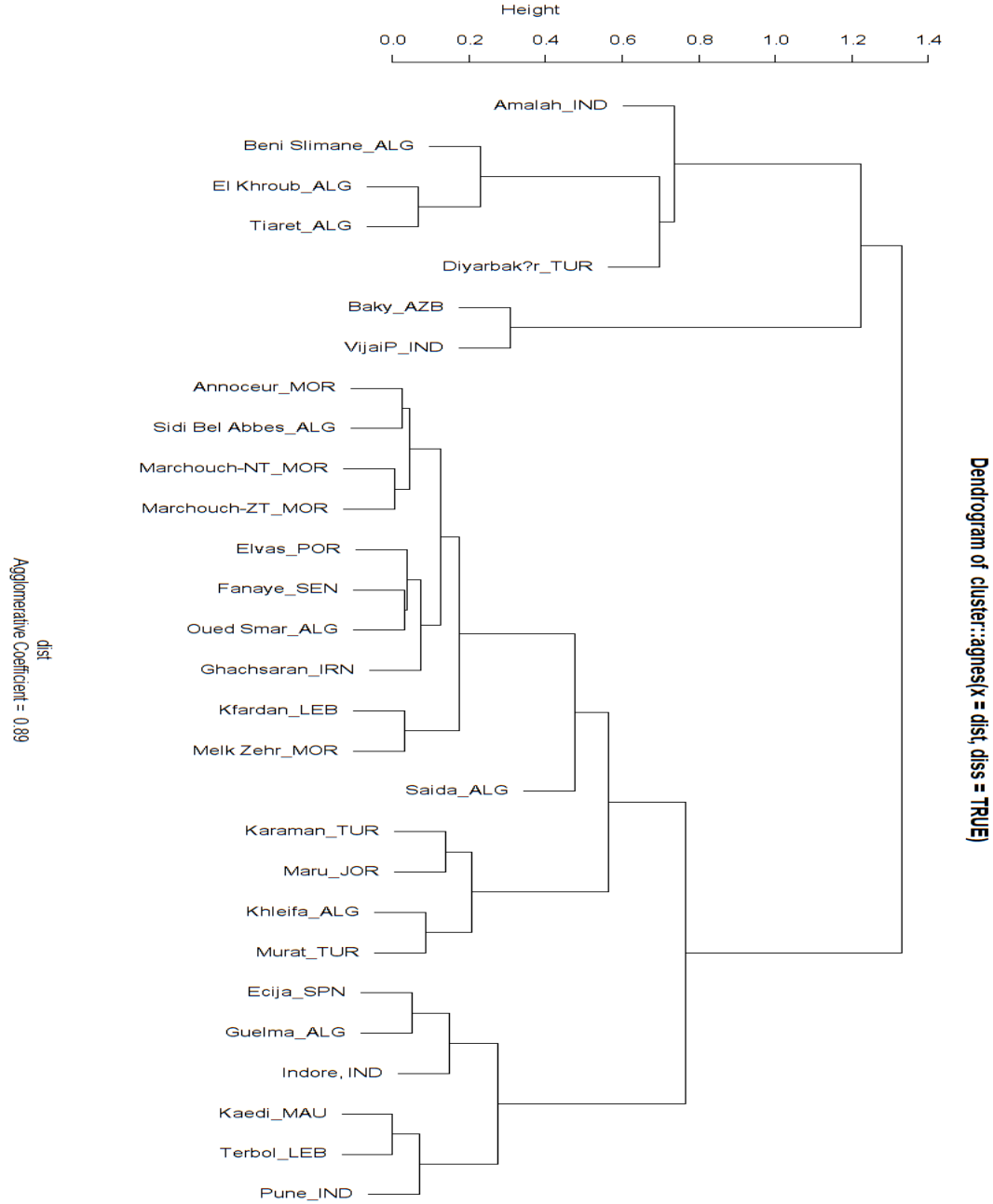
- Planning: which populations, how many, blocks
- Designing: experimental designs based on need, requirements and resources available
- Measuring: phenotypic evaluation in the field or controlled environments including diseases
- QC: spectrum of data, visualization in the field,
- Analyzing (QC as well): single field and MET analysis (outlier detection is done here not above): An outlier is a data point that has large studentized residuals from the model.
- Reporting: heritabilities, CVs, BLUEs and BLUPs, GxE correlations and latent regressions for stability

Phenotyping data management: Managing samples, germplasm lists, location lists and trial data (BrAPI)


Breeding Management
bms.icarda.org:48080/ibpworkbench/main
BREEDING ACTIVITIES
Manage Germplasm
Manage Studies
Manage Samples
INFORMATION MANAGEMENT
STATISTICAL ANALYSIS
PROGRAM ADMINISTRATION
FABA BEAN BREEDING
My Programs
kehel
MANAGE STUDIES
PYT Dis Low Latt
Save
Return to Manage Studies
BASIC DETAILS
Actions
Settings
Germplasm & Checks
Environments
Experimental Design
Measurements
Define Measurement Details
Add
TRAITS
Name
Description
Input Variables
Days_to_flowering
Flowering - count days after sowing (number)
Plant_height
Plant height - soil to tip at maturity (cm)
Days_to_maturity
Maturity - count days after sowing (number)
Grain_yield
Grain yield -dry and weigh (kg/ha)
Select All
Remove
Measurements
Select Environment:
1 - Unspecified Location
Records per page:
100
Showing 1 to 84 of 84 entries
Show Categorical Description
REP_NO
ENTRY_NO
PLOT_NO
GID
ENTRY_TYPE
BLOCK_NO
DESIGNATION
Days_to_flowering
Plant_height
Days_to_maturity
Grain_yield
1 9 21101 30072 Test entry 1 SelKFSH/013/237-1 109 77 168 1823.9
1 16 21102 30079 Test entry 1 SelKFSH/013/317-1 109 79 168 2473.79
1 5 21103 30069 Test entry 1 SelKFSH/013/878-1 112 74 168 2138.36
1 26 21104 30089 Test entry 1 SelKFSH/013/268-1 112 80 168 2430.56
1 1 21105 30065 Test entry 1 SelKFSH/013/680-1 109 93 168 2935.01
1 24 21106 30087 Test entry 1 SelKFSH/013/47-4 103 78 166 2243.19
1 21 21107 30085 Test entry 1 SelKFSH/013/988-1 106 75 166 2557.65
1 2 21108 40971 Test entry 2 SelKFSH/013/758-1 103 74 166 2222.22
1 19 21109 30082 Test entry 2 SelKFSH/013/622-4 112 79 168 2808.3
1 8 21110 40972 Test entry 2 Super Aguadulce 112 78 166 2809.22
1 14 21111 30077 Test entry 2 SelKFSH/013/203-2 112 73 168 1474.15

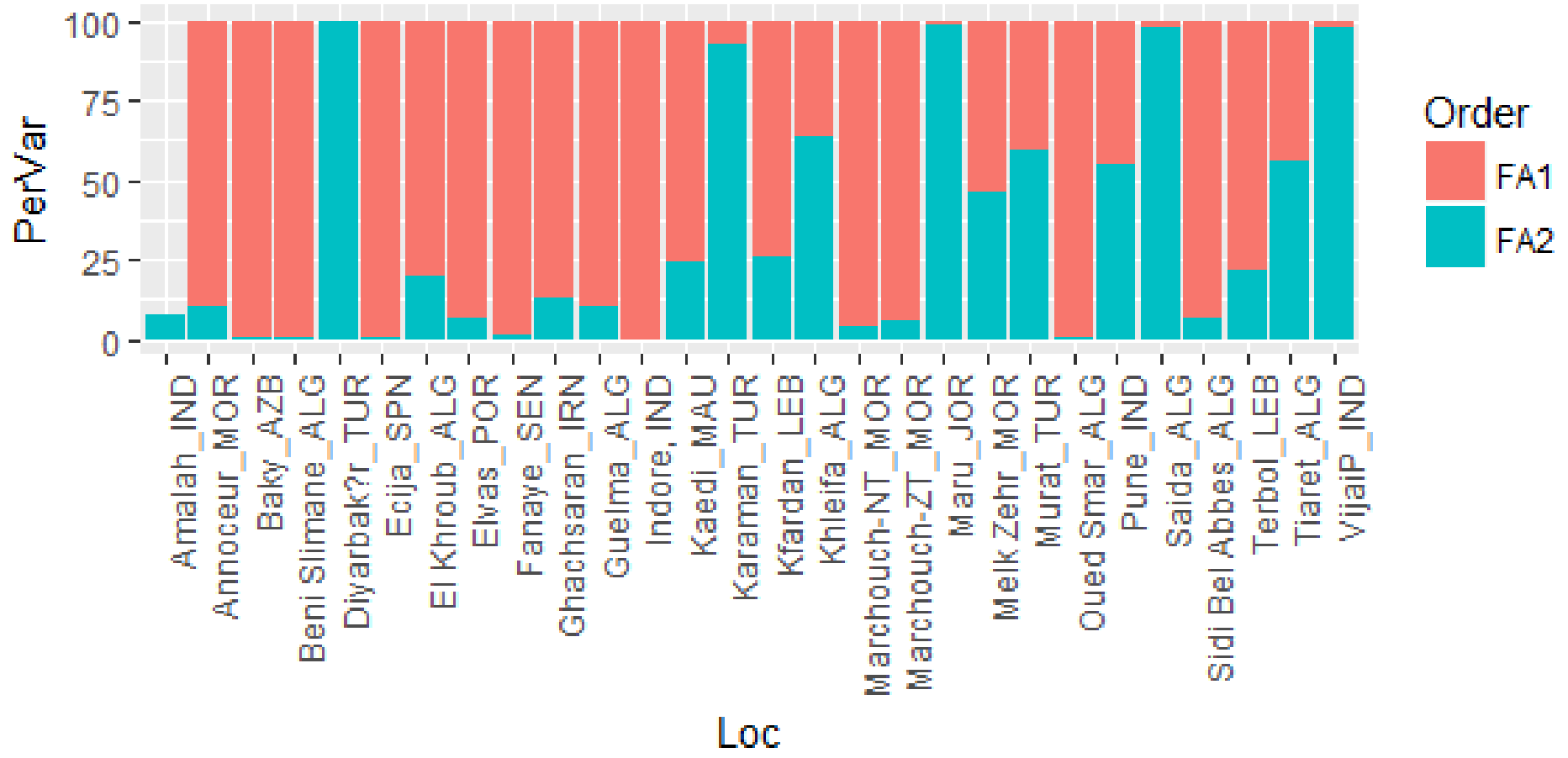
Location genetic correlation matrix for 40



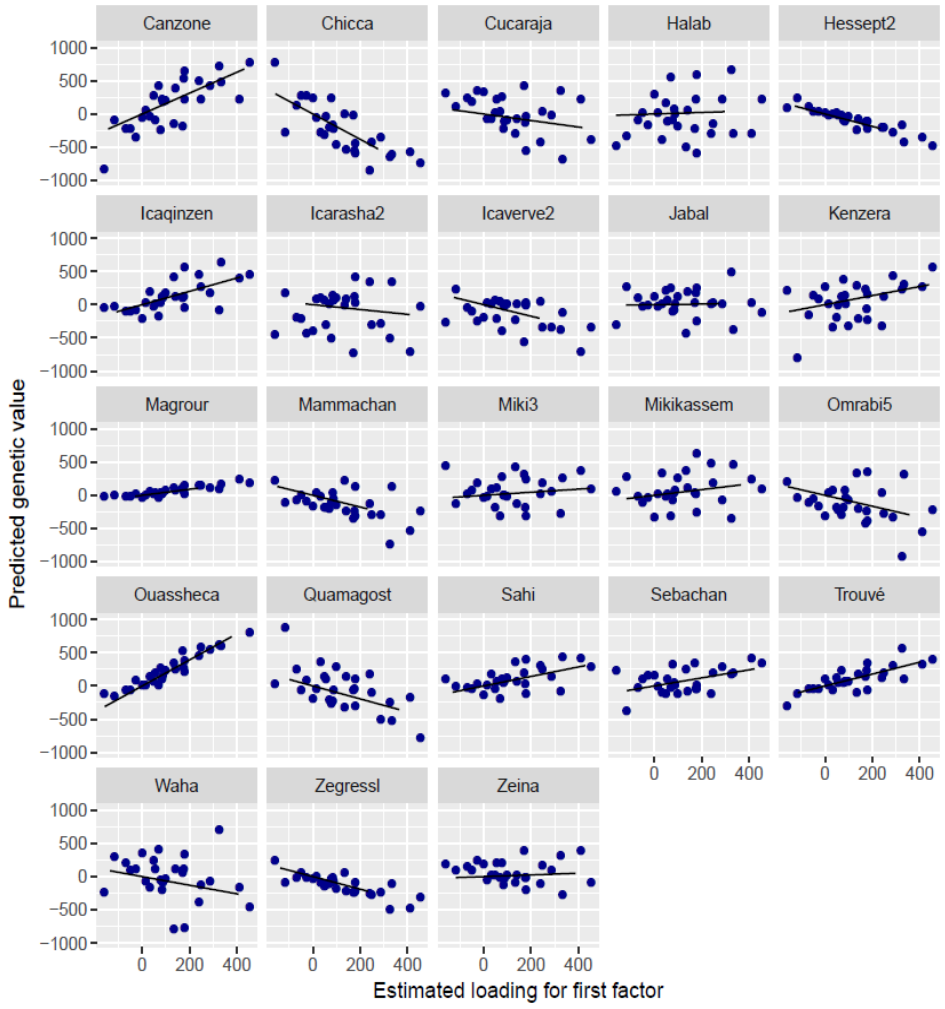


Cluster based on genetic correlations

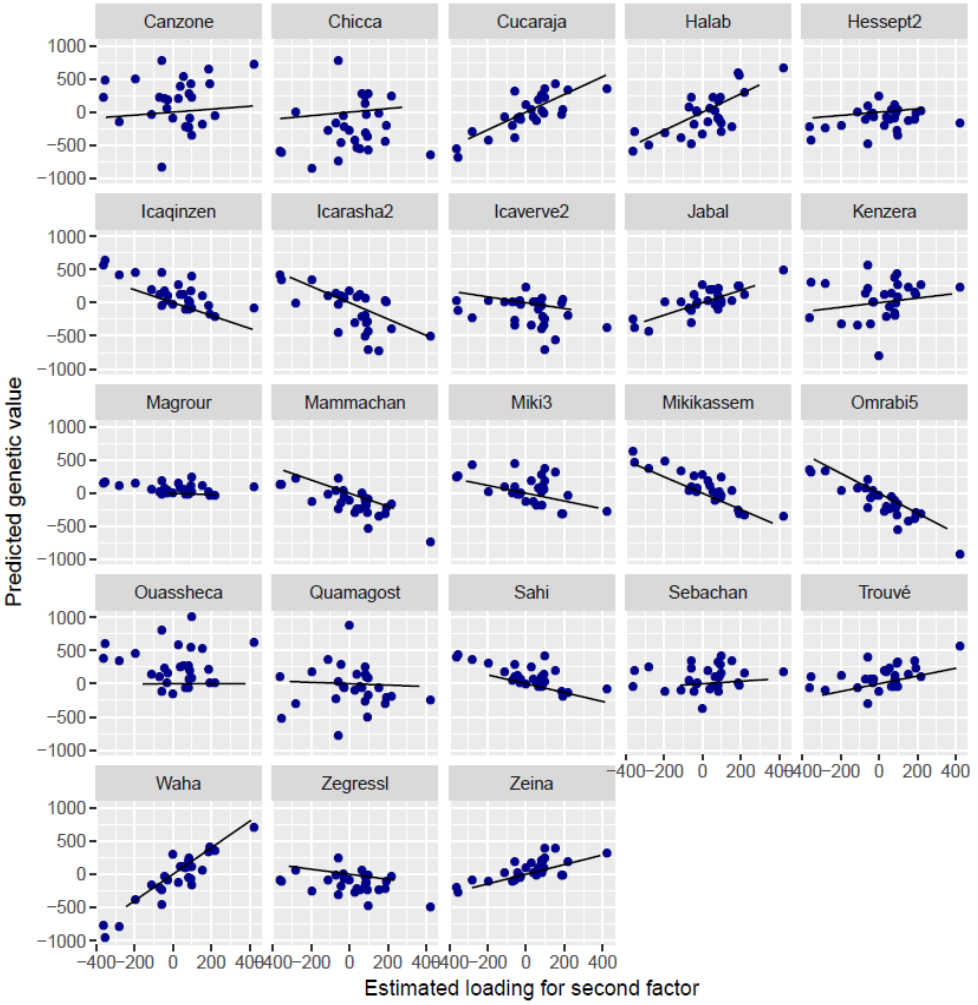
Variance explained by the two first FA model for 40



Latent variable to factor 1 for 40



Latent variable to factor 2 for 40

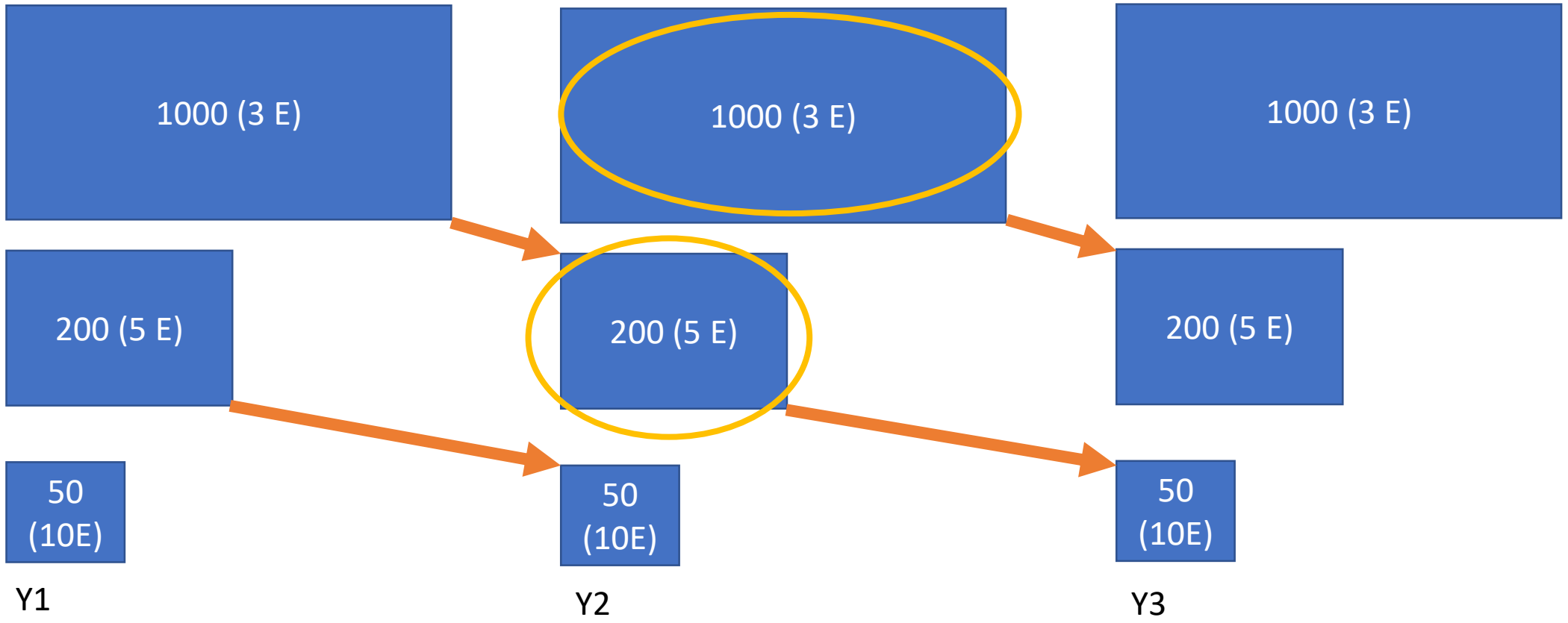


Rotated loading for envs

Env	fac_1	fac_2
Beni Slimane_ALG	-162.03	-58.46
VijaiP_IND	-118.01	-1.35
Baky_AZB	-69.98	81.86
Tiaret_ALG	-50.71	65.40
El Khroub_ALG	-28.10	97.20
Diyarbak?r_TUR	-0.39	219.21
Kfardan_LEB	14.23	-31.52
Maru_JOR	31.97	-111.88
Indore, IND	48.41	82.49
Sidi Bel Abbes_ALG	54.48	87.70
Pune_IND	68.77	192.00
Amalah_IND	77.38	80.73
Fanaye_SEN	83.82	-71.75
Annoceur_MOR	86.22	-27.82
Khleifa_ALG	98.08	-44.16
Saida_ALG	133.14	-281.74
Guelma_ALG	139.93	37.35
Elvas_POR	169.94	152.45
Ecija_SPN	175.39	54.49
Kaedi_MAU	177.63	185.97
Karaman_TUR	178.81	-363.96
Murat_TUR	238.73	-196.15
Marchouch-ZT_MOR	248.02	26.75
Ghachsaran_IRN	285.80	93.59
Terbol_LEB	325.08	422.44
Melk Zehr_MOR	332.14	-354.42
Marchouch-NT_MOR	410.56	98.24
Oued Smar_ALG	453.77	-59.19

Opportunities from Big data

- Looking for patterns across time and region
- Most of scientist don't use data beyond the end of the project/experiment
- There are massive opportunities of enhancing analysis using historical data



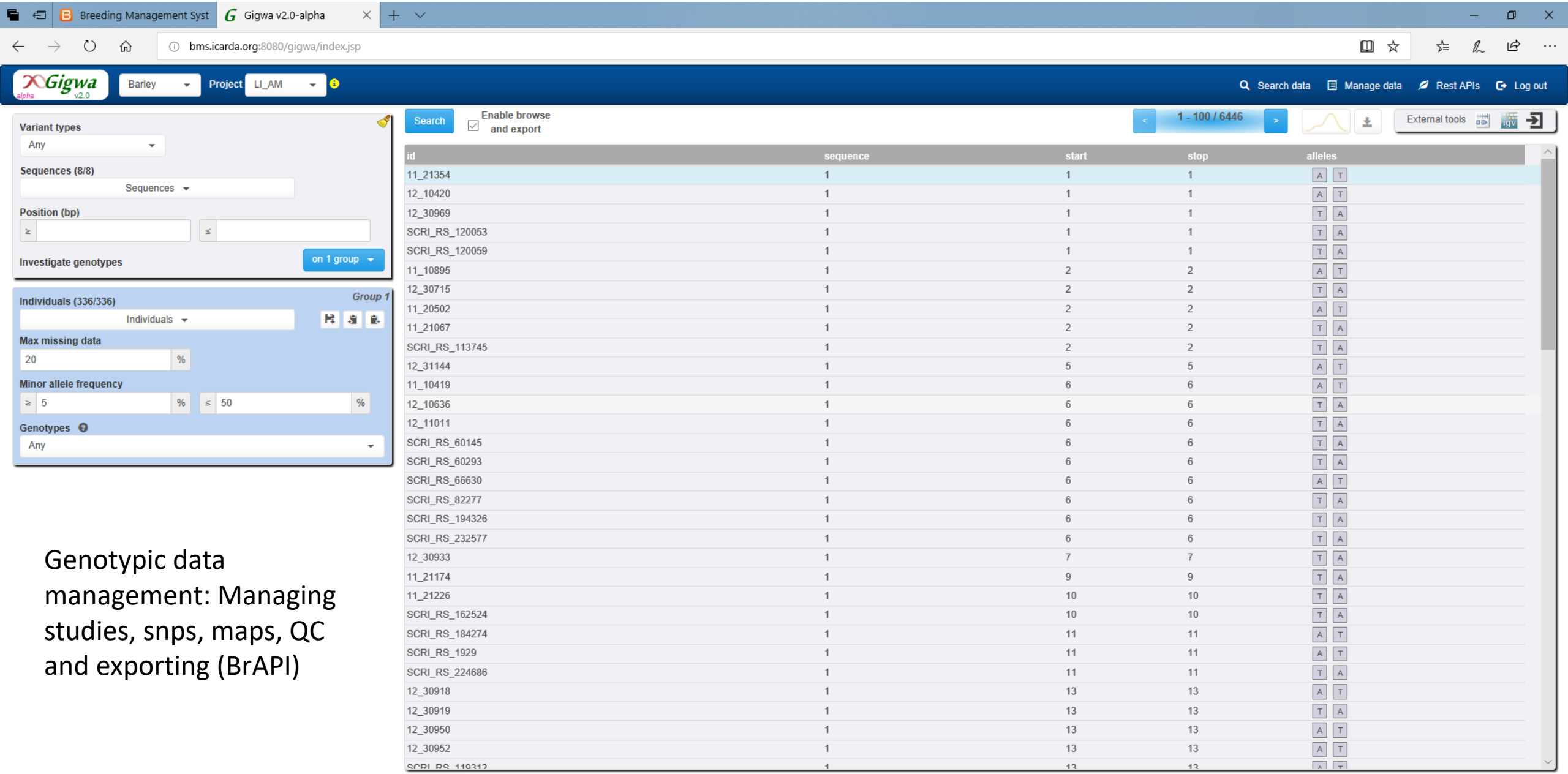
Data Cycle at BCI (Geno)

- Planning: nature of genotyping based on need and objectives of the study (high versus low density)
- Sampling and tracking: get samples to go for genotyping from leaves (we send leaves or DNA). Every sample has an ID that is linked to the trial and GID (Sample tracking is mandatory)
- Genotyping: Send samples with their ID in the genotyping platform format
- Data management: That where the raw data is managed (GIGWA in ICARDA case)
- QC: Can also happen within the data management system
- Analyzing: Diversity, MT association, pop genetics, GS

	n	n0	n1	n2	p	maf	mgf	mono	loh	hwe.chisq	hwe.chisq.p	hwe.fisher	hwe.fisher.p
101032366_A_G_42	452	0	10	442	0.011061947	0.011061947	0	FALSE	FALSE	0.056554009	0.812027789	0	1
101032387_T_G_25	454	10	1	443	0.023127753	0.023127753	0.002202643	FALSE	FALSE	410.8170314	0	Inf	1.19E-19
101032400_G_A_25	378	234	92	52	0.740740741	0.259259259	0.137566138	FALSE	FALSE	50.72575802	1.06E-12	5.717051605	9.18E-12
101032420_A_G_60	386	10	1	375	0.027202073	0.027202073	0.002590674	FALSE	FALSE	349.135051	0	Inf	6.11E-19
101032426_A_G_13	441	7	45	389	0.066893424	0.066893424	0.015873016	FALSE	FALSE	14.7057884	0.00012566	5.344068031	0.00179741
101032452_A_G_5	440	4	3	433	0.0125	0.0125	0.006818182	FALSE	FALSE	230.5230805	0	649.1946425	4.85E-08
101032486_C_T_37	454	6	27	421	0.042951542	0.042951542	0.013215859	FALSE	FALSE	34.73998597	3.77E-09	13.65747393	6.26E-05
101032514_G_T_55	429	4	1	424	0.01048951	0.01048951	0.002331002	FALSE	FALSE	338.0652464	0	Inf	3.59E-09
101032544_C_T_67	395	80	17	298	0.224050633	0.224050633	0.043037975	FALSE	FALSE	303.267407	0	313.0844685	2.35E-63
101032548_T_C_28	371	6	1	364	0.017520216	0.017520216	0.002695418	FALSE	FALSE	315.1795035	0	Inf	2.01E-12
101032558_C_T_47	447	180	39	228	0.446308725	0.446308725	0.087248322	FALSE	FALSE	303.1103211	0	105.3569953	2.41E-77
101032572_A_G_48	441	7	28	406	0.047619048	0.047619048	0.015873016	FALSE	FALSE	39.69	2.98E-10	14.26249024	1.47E-05
101032575_T_C_39	442	8	0	434	0.018099548	0.018099548	0	FALSE	TRUE	442	0	Inf	2.95E-17

	101032366_A_G_42	101032387_T_G_25	101032400_G_A_25	101032420_A_G_60	101032426_A_G_13
SEEDICAR10048	AA	GG	GA	AA	AA
SEEDICAR10052	AA	GG	AA	AA	NA
SEEDICAR10187	AA	GG	GG	NA	AA
SEEDICAR10189	AA	GG	GG	AA	AA
SEEDICAR10191	AA	GG	GA	AA	AA
SEEDICAR10197	AA	GG	NA	AA	AA
SEEDICAR10262	AA	GG	NA	AA	AA
SEEDICAR10438	AA	GG	GA	AA	AA

	SEEDICAR10048	SEEDICAR10052	SEEDICAR10187	SEEDICAR10189	SEEDICAR10191	SEEDICAR10197	SEEDICAR10262	SEEDICAR10438
101032366_A_G_42	2	2	2	2	2	2	2	2
101032387_T_G_25	2	2	2	2	2	2	2	2
101032400_G_A_25	1	2	0	0	1	NA	NA	1
101032420_A_G_60	2	2	NA	2	2	2	2	2
101032426_A_G_13	2	NA	2	2	2	2	2	2
101032452_A_G_5	2	2	2	2	2	2	2	2
101032486_C_T_37	2	2	2	2	2	2	2	2
101032514_G_T_55	2	2	2	2	2	2	2	2

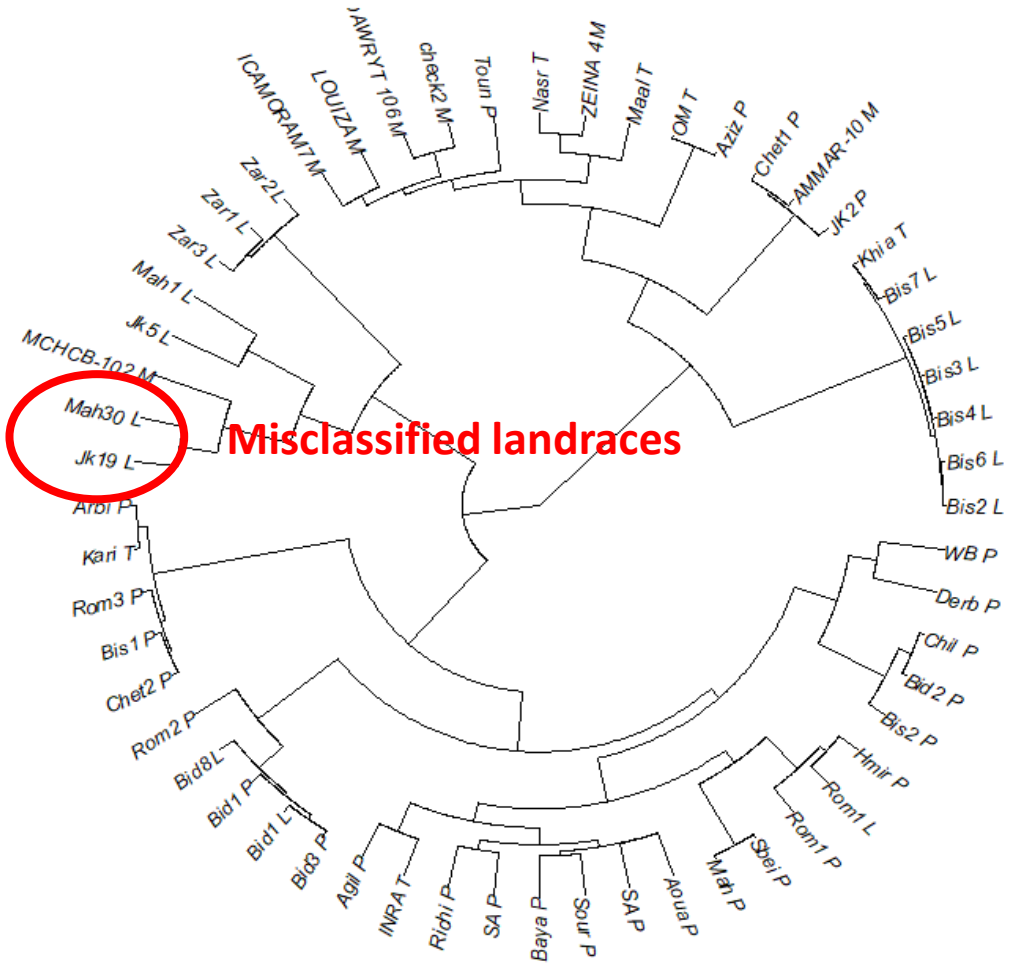


Genotypic data management: Managing studies, snps, maps, QC and exporting (BrAPI)

Enhancing conservation and use of genebank accessions using molecular markers

Conservation

- Identification of duplicates
- Identification of genetically close accessions
- Identification of misclassified accessions at the taxa and name levels
- Genetic structuring to inform on gaps in collections and collecting



Use

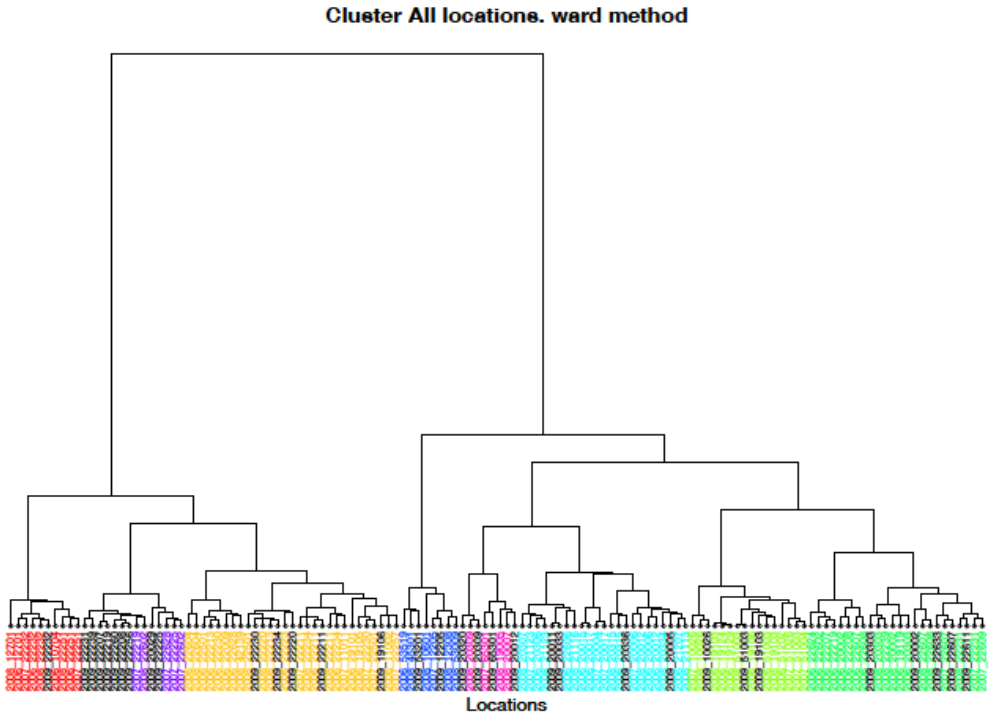
- Increasing FIGS efficiency to identify trait best-sets
- Linking trait to accessions using genomic prediction
- Trait-marker association (GWAS)
- Adaptation mechanisms thru marker-environment association (EWAS)

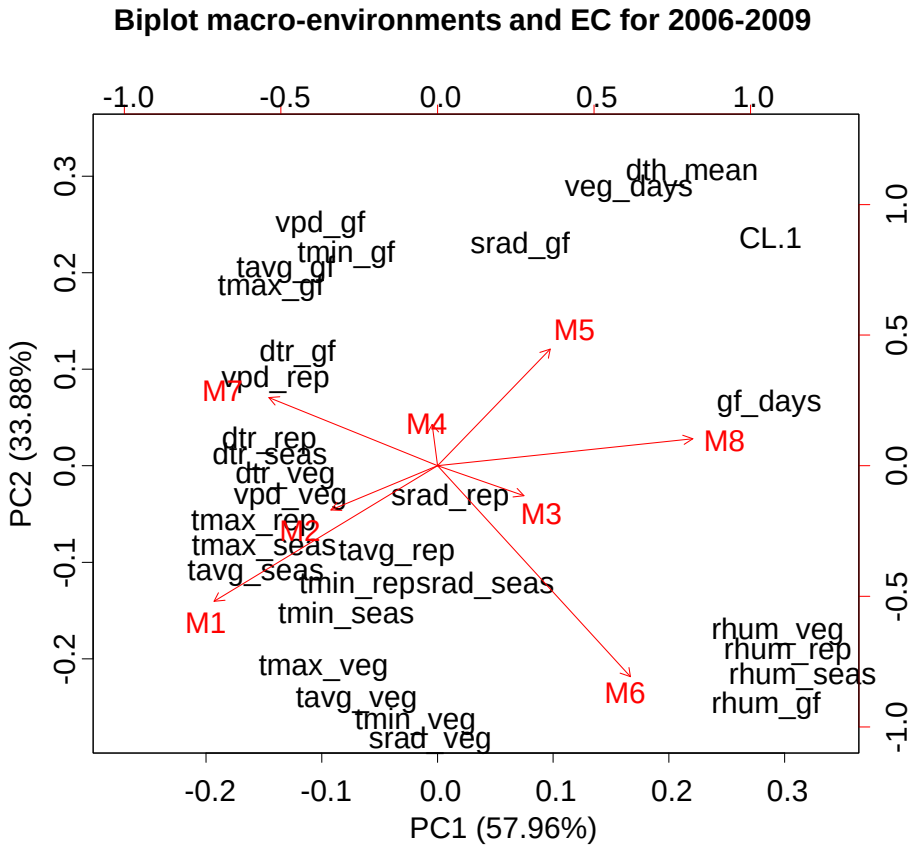
ICARDA
Science for resilient livelihoods in dry areas

Predictions of IN using GS

SAWYT - Correlation between the predicted values of 17 models trained with data from three years to predict the four different macro-environments in 2008, 2007, and 2006.				
	Model	2008	2007	2006
0	ML	0.1540	0.0693	0.2268
1	MA	0.1608	0.0455	0.2289
2	MAW	0.3541	0.1863	0.3317
3	MAW+AW	0.2953	0.1377	0.3173
4	MAW+AW+MA	0.2327	0.1473	0.3383
5	MAS(W)	0.6140	0.5579	0.4854
6	MAS(W)+ML	0.5918	0.4981	0.5098
7	MAS(W)+MA	0.6495	0.5375	0.4115
8	MG	0.1250	0.0565	0.2230
9	MGW	0.2746	0.1186	0.3282
10	MGW+GW	0.1726	0.1521	0.2937
11	MGW+GW+MG	0.2351	0.0327	0.2937
12	MGS(W)	0.6669	0.5347	0.4123
13	MGS(W)+ML	0.2883	0.5136	0.4485
14	MGS(W)+MG	0.5898	0.5267	0.4396
15	MAGS(W)+ML	0.6343	0.5114	0.4931
16	MAGS(W)+MA+MG	0.6612	0.5261	0.4918

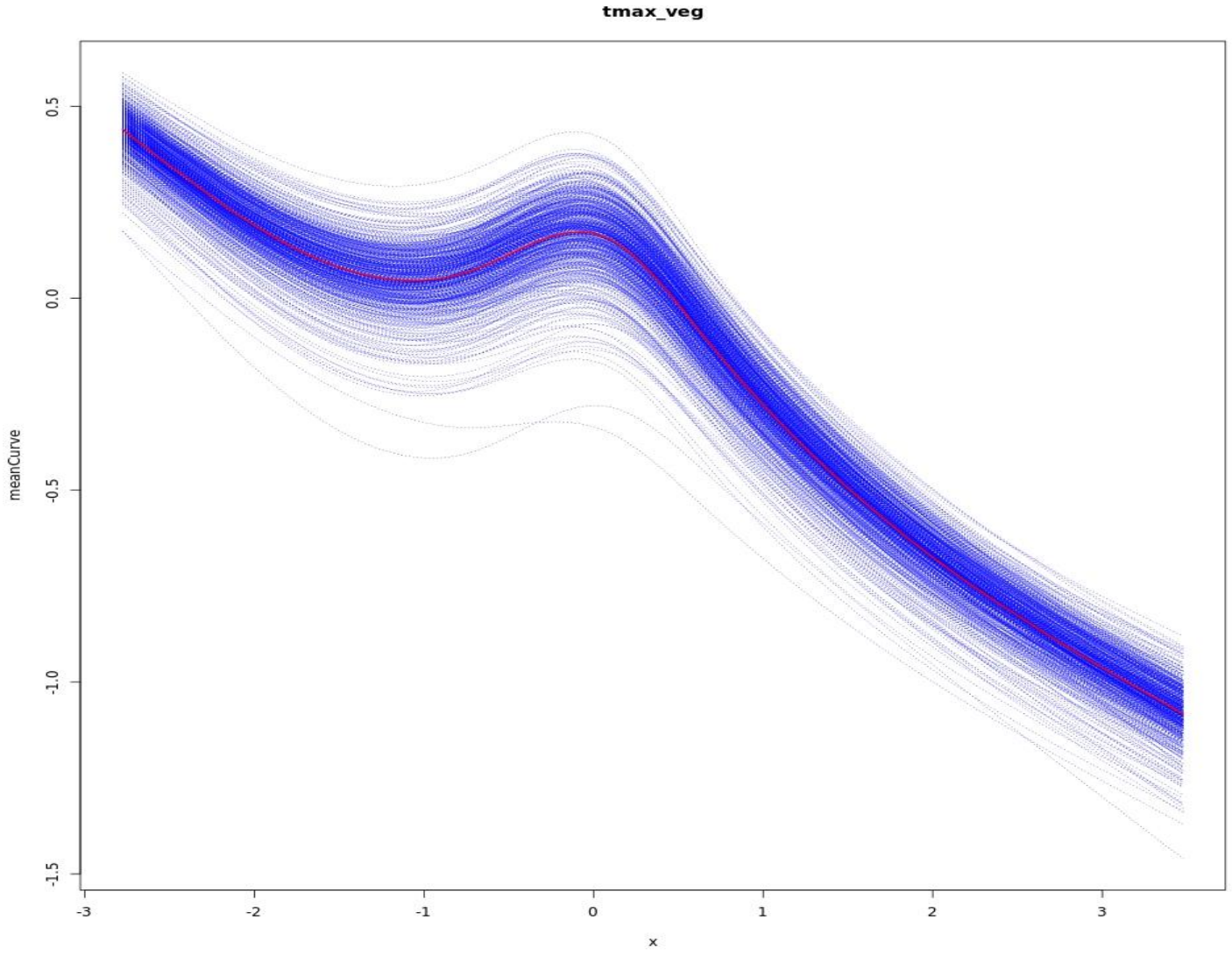
Understanding target environments in IN





8 macro environments

Using historical data to inform breeding strategies using GS



Thank you!

Questions and comments are welcome!