Importancia da Biometria no Melhoramento

XXIII International Symposium in Genetics and Plant Breeding

Fernando Toledo, Philomin Juliana, Leonardo Crespo-Herrera, Jose Crossa & Juan Burgueño

International Maize and Wheat Improvement Center

Lavras, MG – September 5th, 2019

CIMMYT MR

Introduction

"Today more people are hungry than entire population of South Asia at beginning of Green Revolution (1970)"

- World hunger **rising** in 2016 for first time this century
- **815 million** chronically undernourished – up to 38 million
- 489 million located in countries affected by **conflicts**

"Hunger and Malnutrition kill 1.5 more people than AIDS, Tuberculosis, Diabetes, Road accidents, Malaria and all natural disasters combined (US Department of State)"

CIMMYT

Wheat Breeding in CIMMYT

Challenge: *1.6% increase in global production annually; i.e. average yield to rise from 3 t/ha to 5 t/ha by 2050*

- Globally the **most important** food crop
- Food for **2.5 billion** poor ($<$ US$2) in 89 countries
- Important source of **calories** and **protein** in developing countries
- **Product** Lines as source of parents and/or for direct release
  - Over 60 million ha:
    - Climate change;
    - Depleting ground water;
    - Energy and fertilizer costs; and
    - Emerging diseases and pests

CIMMYT

Priorities Traits

- **High** and **stable** yield potential
- Durable **resistance** to *rust* fungi
- Water use **efficiency**
- **Drought** tolerance
- **Heat** tolerance
- Appropriate end-use **quality**
- **Enhanced** *Zn* and *Fe* content (South Asia)

CIMMYT MR

Genotype x Environment x Management x Trait x . . .

- Evaluate the **repeatability** of certain types of interaction
  - approach the **similarities** among test locations; and
  - identify **patterns** of interactions across years.
- **Target Definition:** concept of Mega-Environment (**ME**)
  - **Broad** area (*not contiguous* but frequently *transcontinental*);
  - **Climatic** factors;
  - Similar main **stress** (biotic/abiotic);
  - **Cropping system**; and
  - Consumer **preferences**

CIMMYT

Pattern Analysis

- Multi-environment trials spans over years (**GLY** array)

$$y_{ijk} = \mu + x_{ijk}$$

  - Assumes that genotypes in a year are **representative**
  - **Distances/Correlation** among locations within years
  - **Classification** and **Ordination** of environments across year
    - $\boldsymbol{D}_{ii'}$ are dissimilarities
    - $\boldsymbol{a}_{ii'}$ are similarities
    - $\boldsymbol{D}$s and $\boldsymbol{a}$s are complementary (Gower complements)
  - . . . **dimensional** reduction methods (Eigen decomposition)

* *DeLacy* (90's) collection of papers

CIMMYT

Mega Environments

ME	Latitude	Moisture	Weather	Season	Area
1	$< 35°$	Irrigated	Temperate	Autumn	30
2	$< 35°$	High Rainfall	Temperate	Autumn	5
3	$< 35°$	High Rainfall	Temperate	Autumn	
4A	$< 35°$	Low Rainfall	Temperate	Autumn	15
4B	$< 35°$	Low Rainfall	Temperate	Autumn	
4C	$< 35°$	Residual Rainfall	Hot	Autumn	
5A	$< 35°$	High Rainfall	Hot	Autumn	10
5B	$< 35°$	Irrigated	Hot	Autumn	
6	$> 35°$	Moderate Rainfall	Temperate	Spring	
Total					**60**

CIMMYT MR

Resistance to key Diseases

- *Septoria* leaf blight (**ME2**)
- Spot Blotch (**ME5**)
- Tan Spot (**ME4**)
- *Fusarium* – head scab and *mycotoxins* (**ME2/4/5**)
- Karnal bunt (**ME1**)
- Root rots and nematodes (**ME4**)
- Wheat blast **new** threat in South Asia (**ME5**)

CIMMYT MR

Strategy (5-year cycle)

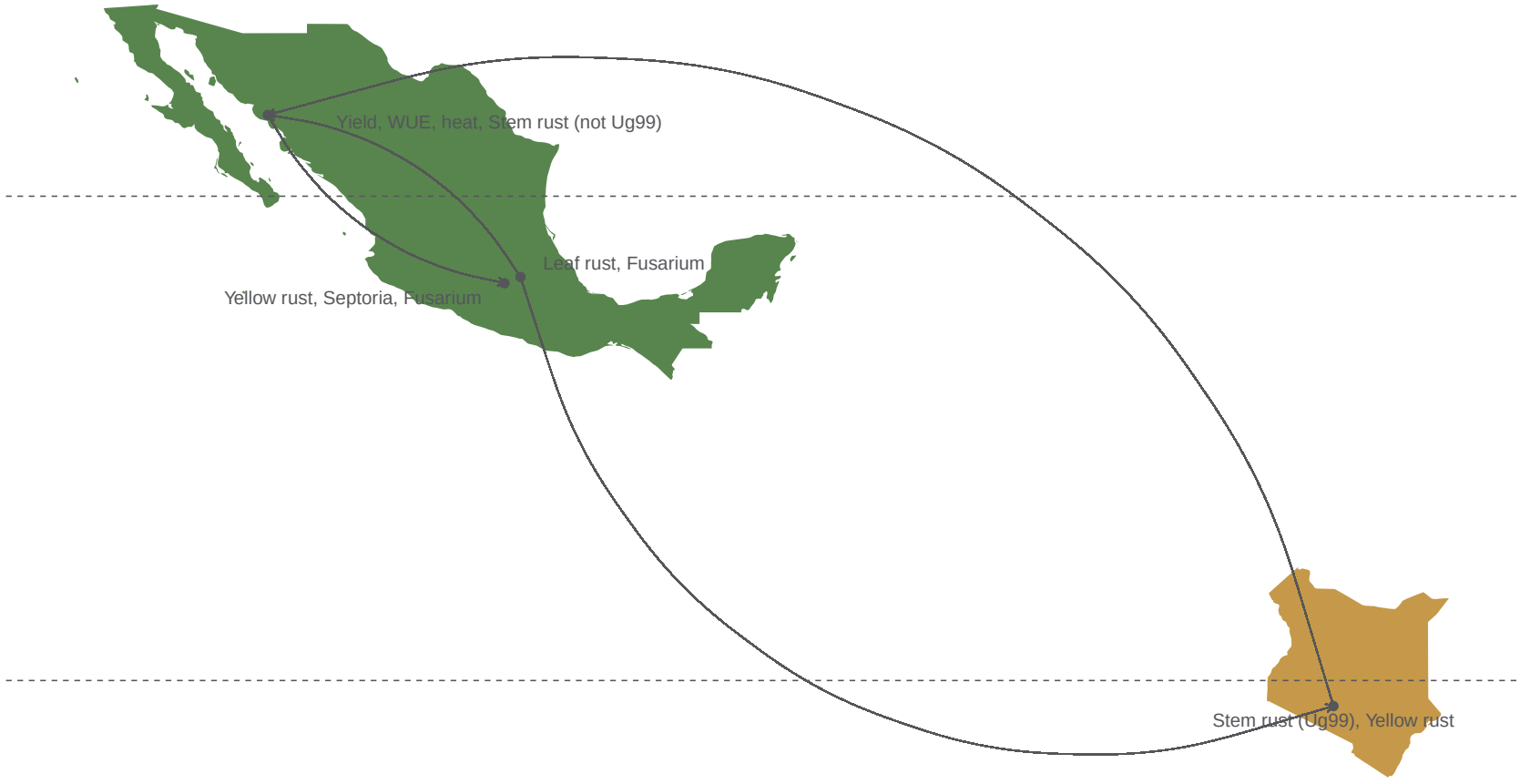
Up-scaled breeding and testing to deliver genetic gain. . .

- Parental **diversity**
  - High value parents as donors for different traits
- **Crosses:** ~1500 Biparental, ~500 Top and ~500 Back
- Targeted utilization of **new** genes, **traits** and **germplasm**
- **Large** population sizes (**depends**)
- **Selected-bulk** selection scheme
  - Selection of progenies in segregation generations

"Each selection adds to the gains for more than one trait"

CIMMYT

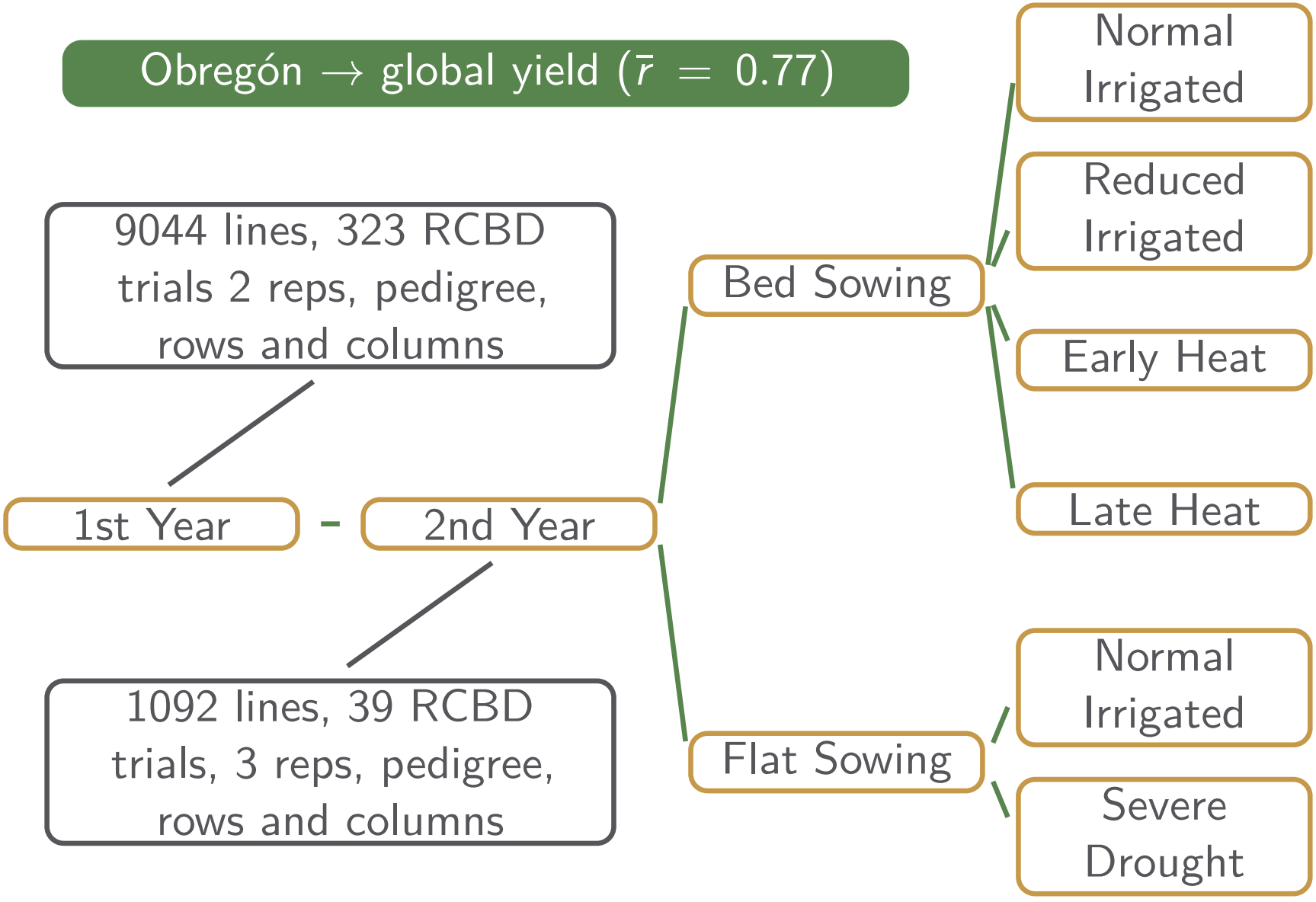
Borlaug's Shuttle Breeding



Braun *et al* (1996) doi: 10.1007/BF00022843

CIMMYT MR

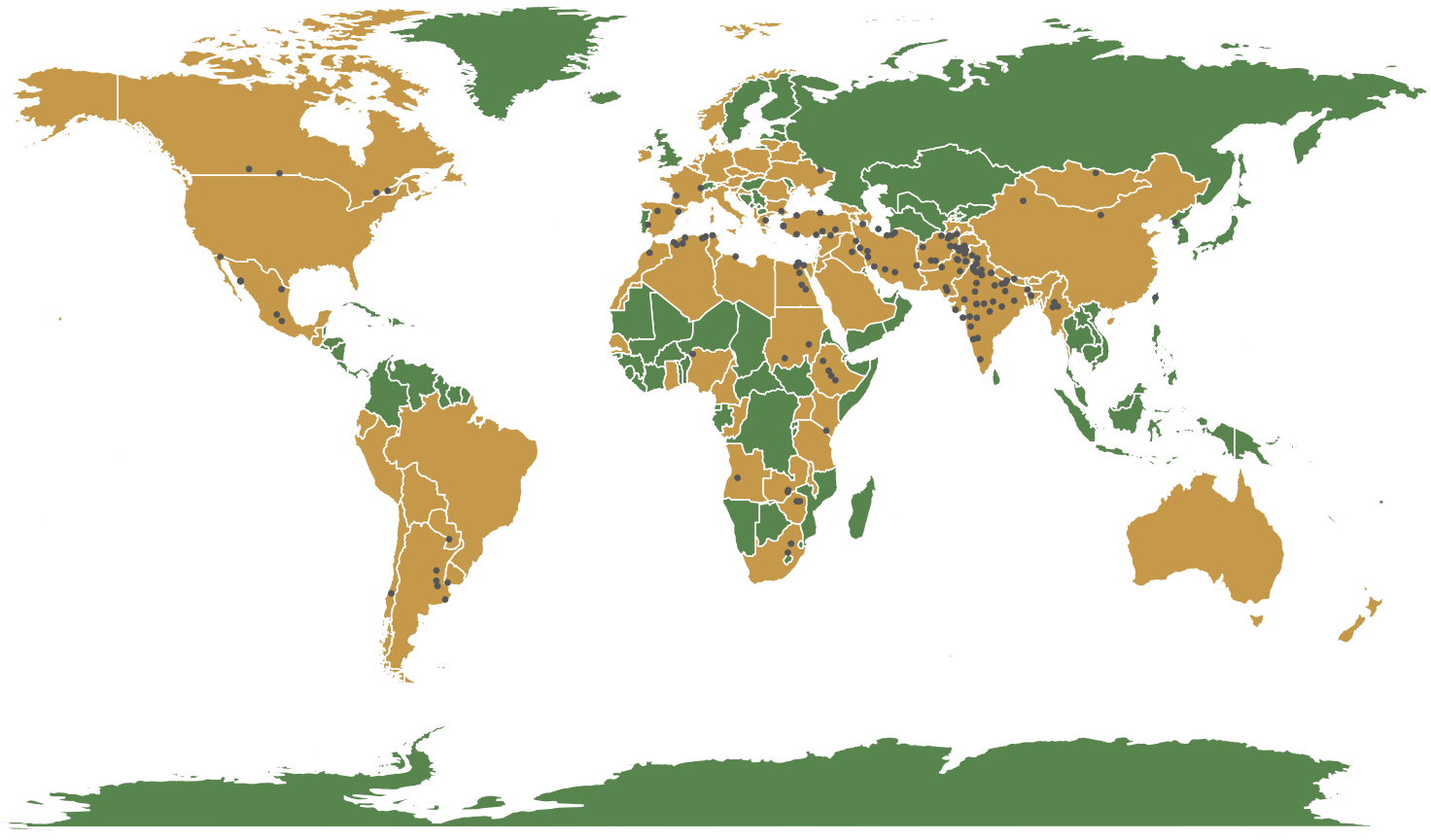
Phases and Analysis



CIMMYT

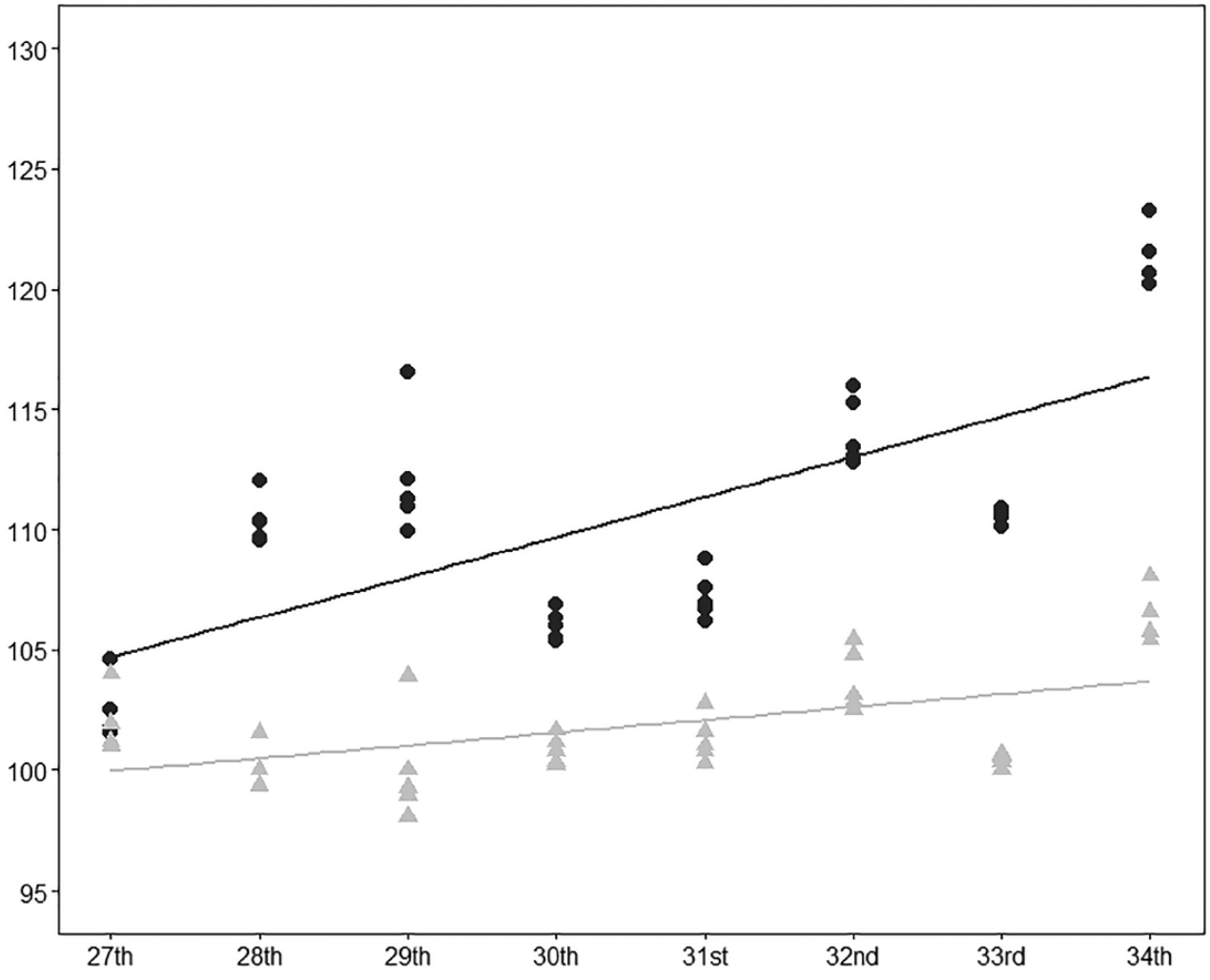
International Trials

- Elite Nursery annually distributed by CIMMYT to collaborators
  - ~200 sites with 50 lines (+**checks**) in $\alpha-$lattice



CIMMYT

Genetic Progress (gains)



Crespo-Herrera *et al* (2017) doi: 10.2135/cropsci2016.06.0553

CIMMYT

Genomic Selection

- Biparental **QTL** has low power for marker–trait
- Conventional pedigree does not account for **Mendelian Sampling**
- **Complications:**
  - **size** and **diversity** of training populations;
  - **heritability** of the target trait;
  - dimensionality of data ($p \gg n$); and
  - **multicolinearity** among markers

$$R = \frac{i \times h \times \sigma_A}{t}$$

Crossa *et al* (2017) doi: 10.1016/j.tplants.2017.08.011

CIMMYT

Accessing Accuracy

- Four basic **scenarios**:
  - **Tested** and **Untested** Lines (observed/unobserved)
  - **Tested** and **Untested** Environments (observed/unobserved)
- Predict lines in environments where they were not tested (**CV1**)
- Predict lines in some environments but not in others (**CV2**)
- Predict lines in untested environments (**CV0**)
- Try to mimic *sparse* testing

Burgueño *et al* (2012) doi: 10.2135/cropsci2011.06.0299

Rapid Cycling

- **GS** 13.4% of gains against checks
- **GS** higher gains than pedigree (7.3%)
- **GS** (drought) **2x** higher gains than others
- Alternative considering:
  - **cost** markers *vs* phenotyping
  - difficulties to phenotype **stress**
  - opportunity to study **inheritance**

Beyene *et al* (2014) doi: 10.2135/cropsci2014.07.0460

CIMMYT

"Genomic" by Environment Interaction I

***all** marker's **ECs** interactions becomes infeasible to manage*

$$y \sim \mu + w_{ij} + g + wg + \epsilon \text{ with } wg \overset{i.i.d.}{\sim} N(0, \boldsymbol{Z}_g \boldsymbol{G} \boldsymbol{Z}_g' \circ \Omega_{gw})$$

- With interactions (wg):
  - **Accuracy:** 35/21% better predictions (CV1/CV2)
  - **Agreement:** 29/45% on top 20% (CV1/CV2)
  - **Variances:** reduces error of about 33%

  the proposed model can be useful for breeding as well as for providing agronomic recommendations tailored to conditions

Jarquin *et al* (2014) doi: 10.1007/s00122-013-2243-1

CIMMYT

"Genomic" by Environment Interaction II

marker effects: *stratified, across or interaction ?*

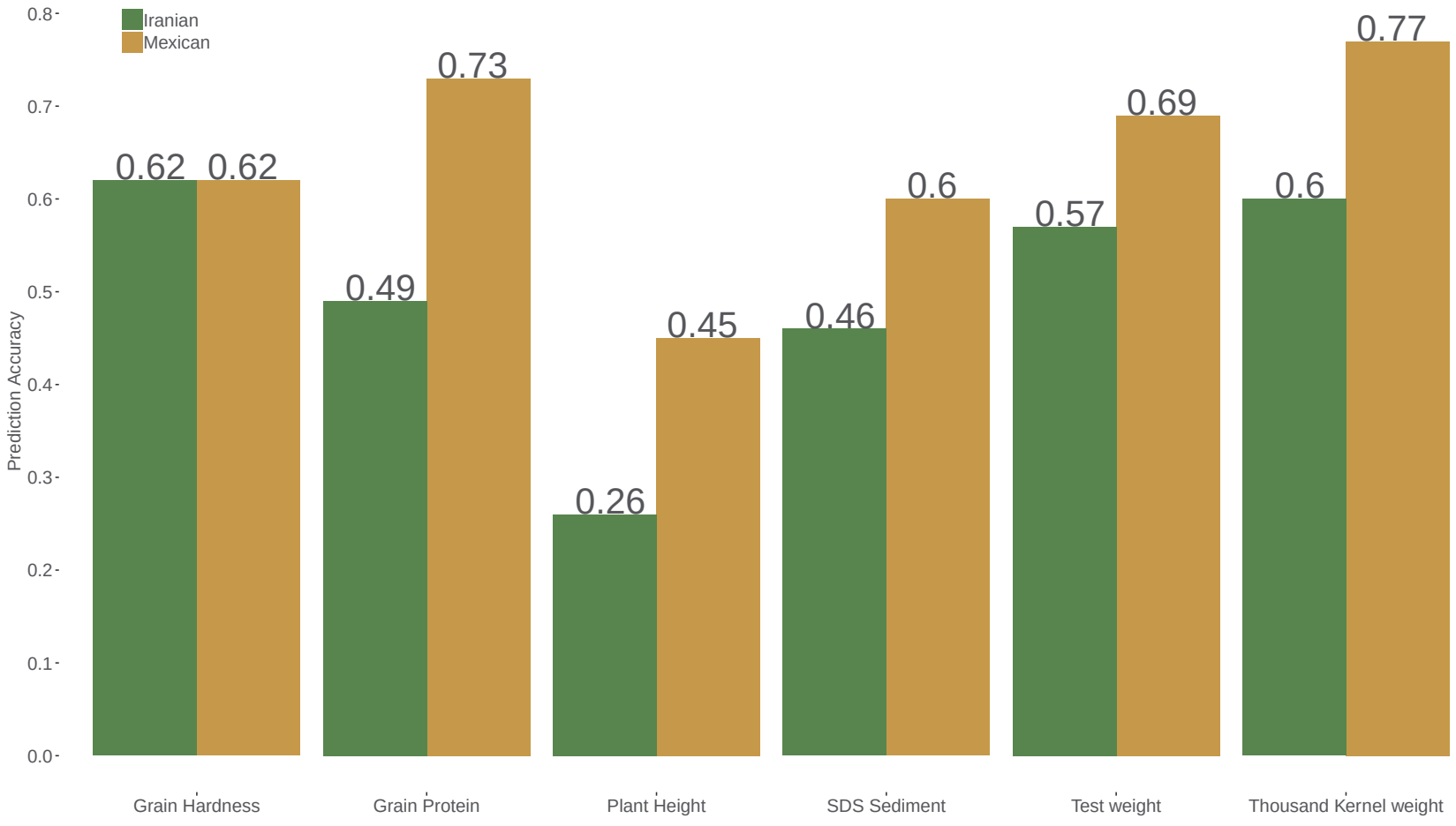
$$y \sim \mu + x(\beta_0 + \beta_i) + \epsilon$$

Environment	Stratified	Across	Interaction	Change (%)	
1	0.471	0.234	0.438	-7	88
2	0.425	0.356	0.413	-3	16
3	0.509	0.386	0.489	-4	27
4	0.451	0.396	0.442	-2	12

Lopez-Cruz *et al* (2015) doi: 10.1534/g3.114.016097

CIMMYT MR

Genomic Selection I



Crossa *et al* (2016) doi: 10.1534/g3.116.029637

CIMMYT

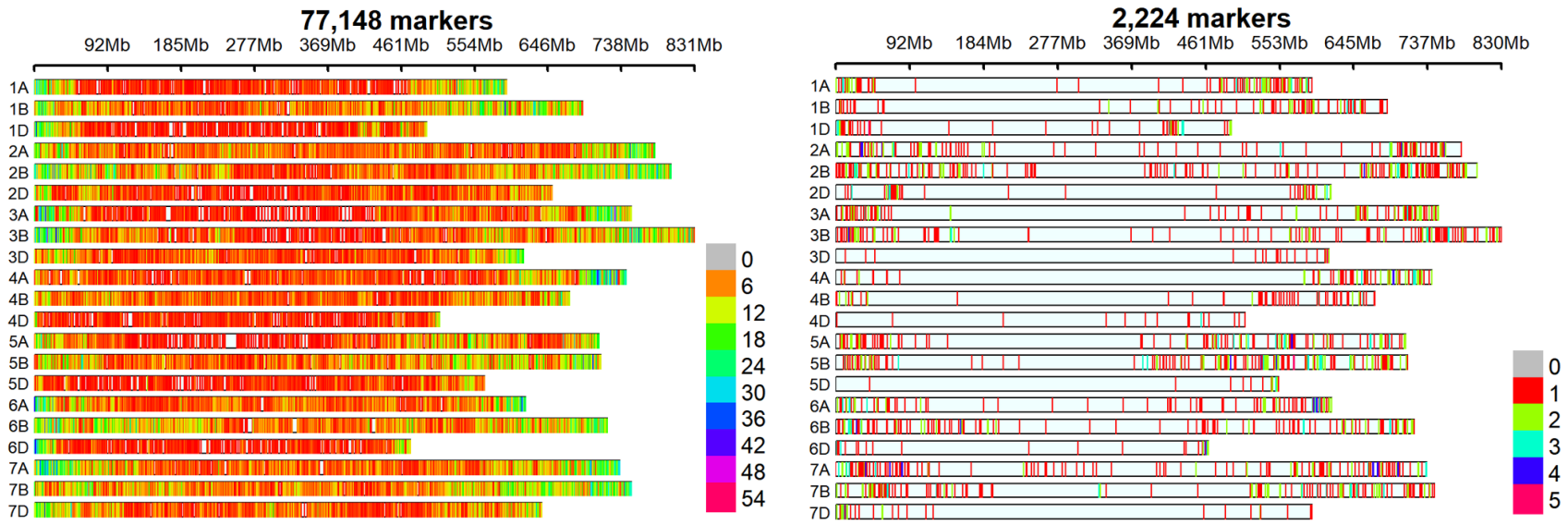
Genomic Selection II (~46,000 lines from 2013 to 18)

- 1st year yield trials average accuracies (r):
  - **within** Yield (0.67) and Stem rust (0.60)
  - **across** Yield (0.42) and Stem rust (0.50)
- 2nd year yield trials accuracies (r):

Yield			Others		
	within	across		within	across
Bed 5IR	0.59	0.15	Flour yield	**0.61**	0.43
Flat 5IR	**0.60**	0.05	Loaf volume	**0.72**	0.50
Bed 2IR	0.59	0.14	Septoria	0.57	0.17
Flat drip	0.59	0.09	Spot blotch	0.55	0.24
Late heat	**0.60**	0.17	Stem rust	**0.79**	**0.60**

CIMMYT

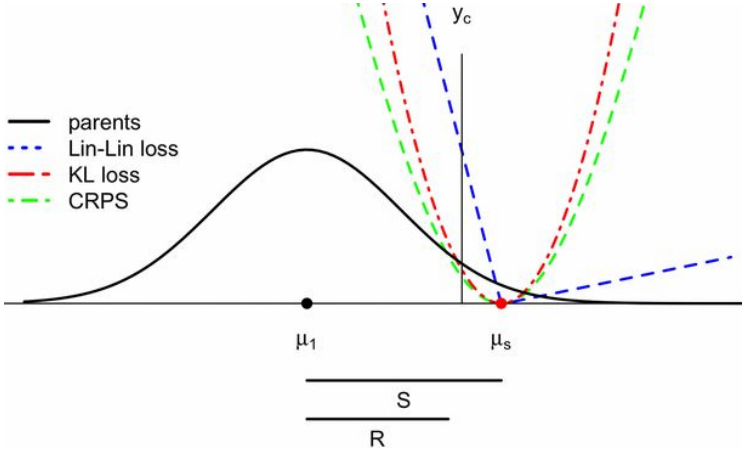
Fine tunning . . . cost (density) $\times$ Accuracy



- high-coverage (less missing) an average increase of 0.02
- another filter **pairwise** correlation
  - $\rho = 0.5$ decreases of 0.05
  - $\rho = 0.3$ decreases of 0.23

CIMMYT

Loss Functions



- Usually, selection is by **truncation**. . .
- **Minimize** risk & **Maximize** gains
- **Proposition:** assess the cost of decision (**loss function**)
  - better performance in long-term selection for single-trait
  - gains for all traits in multi-trait (even with - correlations)
  - no differences among loss functions w.r.t variance components

Villar-Hernández *et al* (2018) doi: 10.1534/g3.118.200430

CIMMYT

Multi-trait

three way *interaction (genotype x trait x environment)*

- Results vary according to the type of prediction (**CV1/CV2**)
- When traits are highly correlated $\rightarrow$ high prediction accuracy
  - **Unstructured** $\gg$ **Diagonal** $\gg$ **Identity**
- **Realistically** mimic the data in plant breeding programs
- Work under development e.g., include other structures (**FA**)

Montesino-Lopez *et al* (2016) doi: 10.1534/g3.116.032359

CIMMYT MR

Generalized Linear Models

appropriate genomic models for data rather than gaussian

- Able to analyse **scales**, **binary** and **ordinal**, **counts** and β data
- Transformations/Approximations $\rightarrow$ **bias** with **low power**
- Account the nonlinear **relationship** between responses
- **Specificities:** discreteness, non-negativity, and overdispersion
- Superior **performance** in terms of prediction **accuracy**

Montesino-Lopez *et al* (2016) doi: 10.1534/g3.116.028118
Montesino-Lopez *et al* (2017) doi: 10.1534/g3.117.039974

CIMMYT

Artificial Intelligence

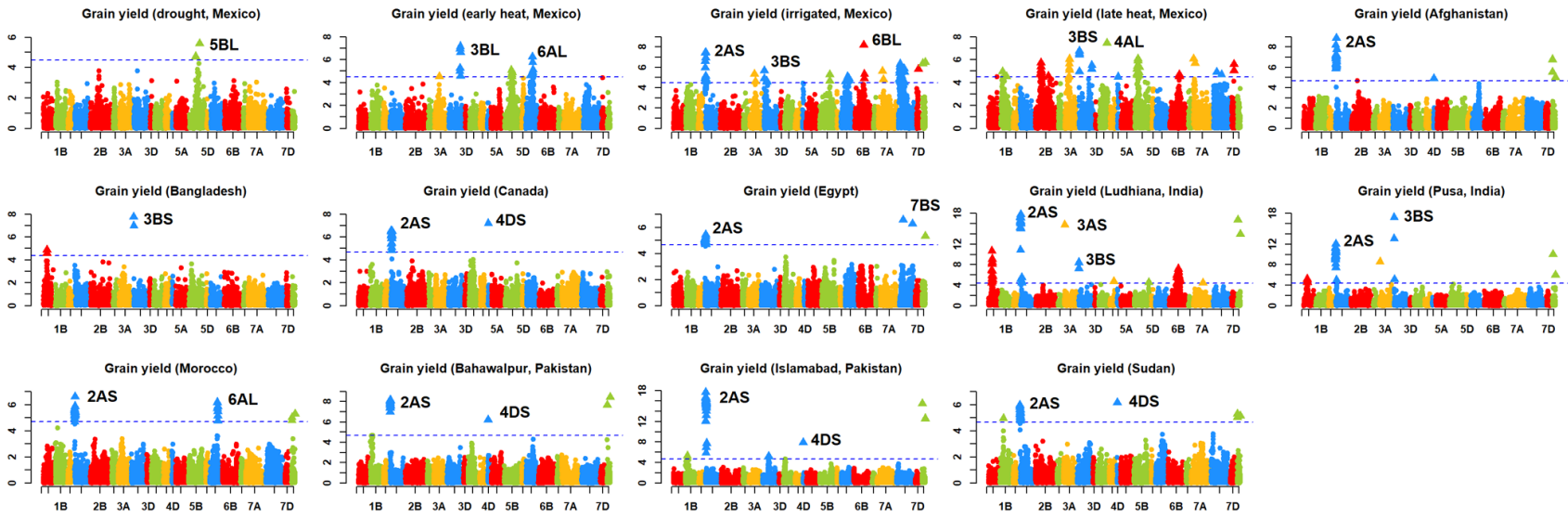
"All models are wrong, but some are useful" (Box)

- **Neural Networks:** parallel chain of **GLMs**
- Only aims to predict new data as accurately as possible
- One *layer* is **close** to penalized regression
- **AI** will be better than parametric whenever the model is wrong
- Possibility to merge non-standard phenotypes e.g., images
- Easily acessible **keras**/**TensorFlow**

Montesino-Lopez *et al* (2019) doi: 10.1534/g3.119.300585
Pérez-Enciso & Zingaretti (2019) doi: 10.3390/genes10070553

CIMMYT

Genomic Association (Inheritance ?)

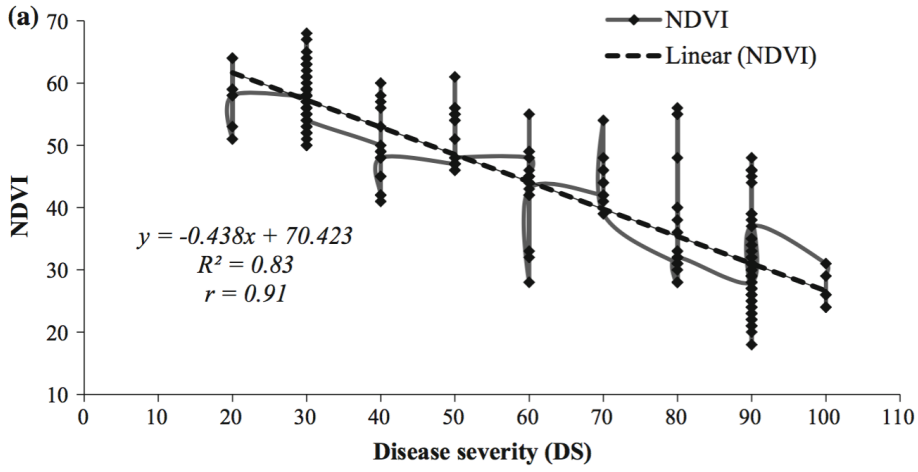


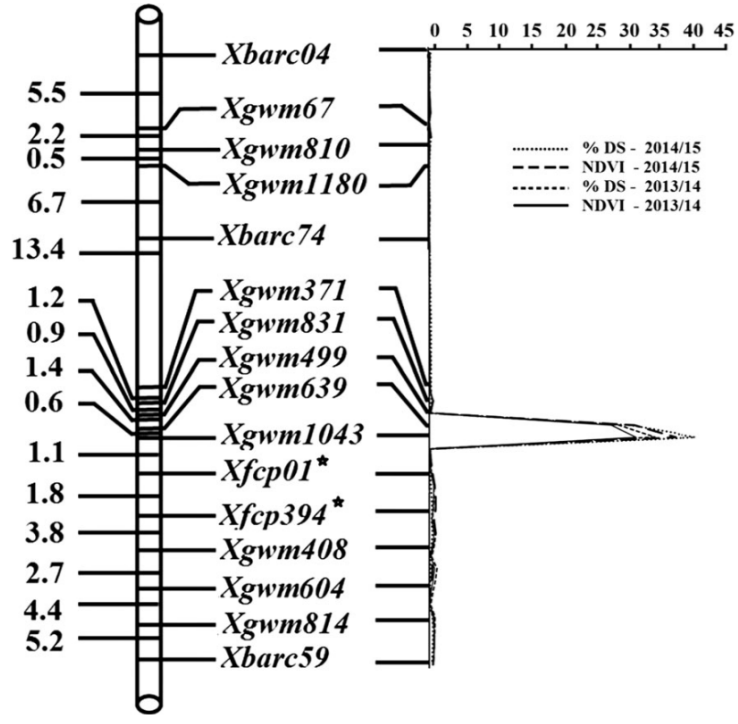
- 44% of the identified QTLs **coincided** with previous reports
- Some regions were consistents **across** environments (stability?)
- **Selection** stronger than drift in driving frequencies (not shown)
- **Additional** results for quality and diseases resistance

CIMMYT

High Throughput Phenotyping I

$$NDVI = \frac{NIR - Red}{NIR + Red}$$



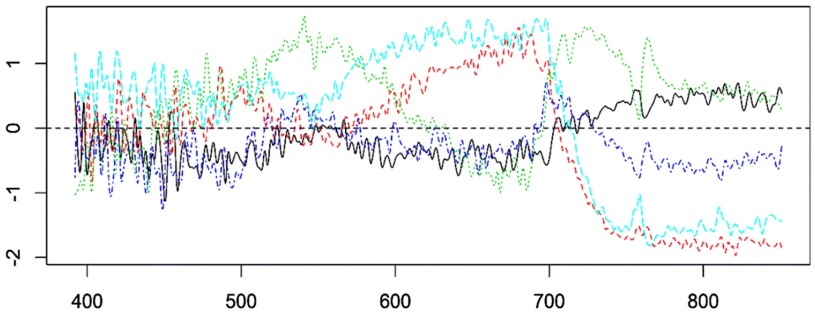


Kumar *et al* (2016) doi: 10.1007/s11032-016-0515-6

CIMMYT

High Throughput Phenotyping II

Data:



Model:

$$y_i \sim \int x_i(t)\beta(t)dt + \epsilon_i$$

*Prediction of yield and other traits by means of **FRA** using hyperspectral images can provide similar and even better accuracies than conventional techniques*

Montesino-Lopez *et al* (2016) doi: 10.1186/s13007-016-0154-2
Montesino-Lopez *et al* (2017) doi: 10.1186/s13007-017-0212-4
Montesino-Lopez *et al* (2018) doi: 10.1186/s13007-018-0314-7

CIMMYT

In Silico Quantitative Genetics

- Much of the **learning** process is **try & error**
- Usually we need to analyse **real** data under clear **scenarios**
- Maybe, **new** methods apply to **new** scenarios
- Several of what was shown was tested by **simulations**
  - Capacity to efficiently represent **full genomes**
  - **Integrates** simulation & analysis (R environment)
  - **low-** and **high-level** interfaces $\rightarrow$ great flexibility

Toledo *et al* (2019) doi: 10.1534/g3.119.400373

Data and Software Availability

- CIMMYT institutional repository of datasets and software:

  https://data.cimmyt.org

- Almost all mentioned **data** and **software** can be found there
  - allowed to use for **research**, **teaching** and **publications**
- **Collection** of softwares for **common** analysis in breeding
  - Multi-environment Trial analysis;
  - Genotype by Environment Interaction analysis
  - ...

CIMMYT

Success & Technology Adoption

- **"international"** performance of genotypes guide **new** crossings
  - giant **recurrent** selection scheme

	Release	CGIAR	%	Area	Yield	Quality
China	226	121	54	28	**78**	17
Europe	2,205	1,225	56	**82**	49	46
Former URSS	318	154	48	25	45	20
Latin American	630	455	72	78	**50**	**50**
South Asia	320	293	**92**	**98**	30	21
Sub-Saharan	291	211	73	**97**	47	15
W. Asia & N. Africa	614	434	71	**98**	47	30
World	4,604	2,893	63	**71**	48	35

Lantican *et al* (2016) isbn: 978-607-8263-55-4

CIMMYT

Final Remarks

- **Biometry** findings are changing breeding operations **daily**
- Proper **analysis** increase genetic **gains** and **understanding**
- **REPL** *read-eval-print* loop for new tools and methods
- Importance of **wide** evaluations under **target** environment
- **Impactful** international collaboration:

  Without this unprecedented cooperation none of this work would have been possible.

- Interdisciplinary research: **computer science**, **mathematics**, **statistics**, **quantitative genetics** and **bioinformatics**

CIMMYT

Acknowledgements

- Bill and Melinda Gates Foundation & **DFID**:
  - **DGGW** Project **HarvestPlus** Project - (**CRP A4NH**)
- Governments:
  - **ACIAR** - Australia
  - **BMZ** - Germany
  - **ICAR** - India
  - **SADER** - Mexico
  - **USAID** - USA
- Farmers' organizations:
  - **Agrovegetal** - Spain
  - **GRDC** - Australia (**ACRCP** & **CAIGE** Projects)
  - **Patronato-Sonora** - Mexico

CIMMYT

Thank you for your interest!

f.toledo@cgiar.org