Importância 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
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)”
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
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)
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**
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
  - \( D_{ii'} \) are dissimilarities
  - \( a_{ii'} \) are similarities
  - \( D_s \) and \( a_s \) are complementary (Gower complements)
  - …dimensional reduction methods (Eigen decomposition)

* DeLacy (90’s) collection of papers
## Mega Environments

<table>
<thead>
<tr>
<th>ME</th>
<th>Latitude</th>
<th>Moisture</th>
<th>Weather</th>
<th>Season</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&lt; 35°</td>
<td>Irrigated</td>
<td>Temperate</td>
<td>Autumn</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>&lt; 35°</td>
<td>High Rainfall</td>
<td>Temperate</td>
<td>Autumn</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>&lt; 35°</td>
<td>High Rainfall</td>
<td>Temperate</td>
<td>Autumn</td>
<td></td>
</tr>
<tr>
<td>4A</td>
<td>&lt; 35°</td>
<td>Low Rainfall</td>
<td>Temperate</td>
<td>Autumn</td>
<td>15</td>
</tr>
<tr>
<td>4B</td>
<td>&lt; 35°</td>
<td>Low Rainfall</td>
<td>Temperate</td>
<td>Autumn</td>
<td></td>
</tr>
<tr>
<td>4C</td>
<td>&lt; 35°</td>
<td>Residual Rainfall</td>
<td>Hot</td>
<td>Autumn</td>
<td></td>
</tr>
<tr>
<td>5A</td>
<td>&lt; 35°</td>
<td>High Rainfall</td>
<td>Hot</td>
<td>Autumn</td>
<td>10</td>
</tr>
<tr>
<td>5B</td>
<td>&lt; 35°</td>
<td>Irrigated</td>
<td>Hot</td>
<td>Autumn</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>&gt; 35°</td>
<td>Moderate Rainfall</td>
<td>Temperate</td>
<td>Spring</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>Total</strong></td>
<td></td>
<td><strong>Total</strong></td>
<td><strong>60</strong></td>
</tr>
</tbody>
</table>
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)
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”
Phases and Analysis

Obregón → global yield ($\bar{\rho} = 0.77$)

1st Year
- 9044 lines, 323 RCBD trials, 2 reps, pedigree, rows and columns

2nd Year
- 1092 lines, 39 RCBD trials, 3 reps, pedigree, rows and columns

Bed Sowing
- Normal Irrigated
- Reduced Irrigated
- Early Heat
- Late Heat

Flat Sowing
- Normal Irrigated
- Severe Drought
International Trials

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

Crespo-Herrera et al. (2017) doi: 10.2135/cropsci2016.06.0553
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
- multicollinearity among markers

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

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

Rapid Cycling

- **Pop** ($F_{2:3}$)
  - **Ped**$_{C0}$
  - **GS**$_{C0}$
  - **Ped**$_{F3:4}$
  - **GS**$_{C1}$
  - **Ped**$_{F4:5}$
  - **GS**$_{C2}$
  - **Ped**$_{F5:6}$
  - **GS**$_{C3}$
  - **Testcrosses**

  - **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**

all marker’s ECs interactions becomes infeasible to manage

\[ y \sim \mu + w_{ij} + g + wg + \epsilon \text{ with } wg \sim N(0, Z_g GZ'_g \odot \Omega_{gw}) \]

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.

**marker effects:** stratified, across or interaction?

\[ y \sim \mu + x(\beta_0 + \beta_i) + \epsilon \]

<table>
<thead>
<tr>
<th>Environment</th>
<th>Stratified</th>
<th>Across</th>
<th>Interaction</th>
<th>Change (%)</th>
<th>Change (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.471</td>
<td>0.234</td>
<td>0.438</td>
<td>-7</td>
<td>88</td>
</tr>
<tr>
<td>2</td>
<td>0.425</td>
<td>0.356</td>
<td>0.413</td>
<td>-3</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>0.509</td>
<td>0.386</td>
<td>0.489</td>
<td>-4</td>
<td>27</td>
</tr>
<tr>
<td>4</td>
<td>0.451</td>
<td>0.396</td>
<td>0.442</td>
<td>-2</td>
<td>12</td>
</tr>
</tbody>
</table>

Genomic Selection I

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 ($r$):

<table>
<thead>
<tr>
<th></th>
<th>within</th>
<th>across</th>
<th>Others</th>
<th>within</th>
<th>across</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bed 5IR</td>
<td>0.59</td>
<td>0.15</td>
<td>Flour yield</td>
<td>0.61</td>
<td>0.43</td>
</tr>
<tr>
<td>Flat 5IR</td>
<td><strong>0.60</strong></td>
<td>0.05</td>
<td>Loaf volume</td>
<td><strong>0.72</strong></td>
<td>0.50</td>
</tr>
<tr>
<td>Bed 2IR</td>
<td>0.59</td>
<td>0.14</td>
<td>Septoria</td>
<td>0.57</td>
<td>0.17</td>
</tr>
<tr>
<td>Flat drip</td>
<td>0.59</td>
<td>0.09</td>
<td>Spot blotch</td>
<td>0.55</td>
<td>0.24</td>
</tr>
<tr>
<td>Late heat</td>
<td><strong>0.60</strong></td>
<td>0.17</td>
<td>Stem rust</td>
<td><strong>0.79</strong></td>
<td><strong>0.60</strong></td>
</tr>
</tbody>
</table>
Fine tuning... 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
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

Multi-trait

three way interaction (genotype $\times$ trait $\times$ 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)

Generalized Linear Models

appropriate genomic models for data rather than gaussian

- Able to analyse scales, binary and ordinal, counts and $\beta$ 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

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 accessible **keras/TensorFlow**

44% of the identified QTLs **coincided** with previous reports.

Some regions were consistent **across** environments (stability?).

**Selection** stronger than drift in driving frequencies (not shown)

**Additional** results for quality and diseases resistance
High Throughput Phenotyping I

\[
NDVI = \frac{NIR - \text{Red}}{NIR + \text{Red}}
\]

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

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 → 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
  ...
Success & Technology Adoption

“international” performance of genotypes guide new crossings

- giant recurrent selection scheme

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

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**
**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
Thank you for your interest!

f.toledo@cgiar.org