Genomic Prediction in Maize: From Diverse Collections to Single-Cross Hybrids

Christian Riedelsheimer

Institute of Plant Breeding, Seed Science, and Population Genetics
University of Hohenheim
Germany

49th Illinois Corn Breeders’ School
March. 4th, 2013
Champaign, Illinois
Need for performance prediction in hybrid maize breeding

Great technological advancements on the one side:
- Speed of inbreeding
- Phenotyping techniques
- Trait dissection

But a great bottleneck on the other side:
- Production & evaluation of (test)crosses slow and expensive
  - Line *per se* performance weak predictor of hybrid performance
  - Many agronomic traits are highly polygenic → MAS failed
Basic rationale of genomic prediction

- Use high-density genomic information to exploit precise relationship information and all QTL effects simultaneously
- No preselection of significant QTL effects for model building

Heffner et al., *Crop Sci.*, 2009
Basic rationale of metabolic prediction

Metabolic prediction

- Works with **concentration levels of metabolites** (sugars, amino acids, organic acids, phenolics...) at a given point in time
- Metabolites capture **high level of condensed information**
- Metabolites might be closer linked to traits tightly connected to plant metabolism

Goodacre, *Metabolomics*, 2005
Focus on GCA first

- Decomposition into **general and specific combining abilities**:
  \[ y_{ij} = \mu + GCA^d_i + GCA^f_j + SCA_{ij} \]

- Rel. contribution of SCA ↓ with ↑ genetic distance between pools

- Much can be gained if GCA could be predicted with high accuracy to identify superior parents to cross in factorials

- → Focus on GCA prediction first
Population

- Diversity panel
- 285 Dent lines
- Stiff Stalk and Non-Stiff Stalk, some tropical (CIMMYT) lines
- Origin: worldwide, but mainly USA and EU
- 570 testcrosses with two $F_1$ Flint testers
Field trials of testcrosses

- 2 years (2008 and 2009)
- 3 locations in South Germany
- Alpha-lattice with 2 rep.
Prescreening for GCA of biomass and quality traits

## Traits / Genotyping

### Traits

- Dry matter yield, plant height, dry matter conc., female flowering
- Quality traits predicted using NIRS (Grieder et al., JNIRS, 2011)
  - Starch
  - Sugar
  - Lignin
- Estimation of GCA values (Grieder et al., TAG, 2011)

### Genotyping

- Illumina MaizeSNP50 chip containing 56,110 SNPs
- 38,019 polymorphic SNPs remained after quality control
LD decay with distance

- $r^2 = 0.1$ at $\approx 500$ kb

LD between adjacent SNPs

- Average $r^2$ between adjacent SNPs: 0.34
Prescreening for GCA of biomass and quality traits

Metabolic profiling

1. Leaf sampling
2. Robotized grinding at -80 °C
3. Ethanolic extraction
4. Spectral analytics (GC-MS)
Structure in metabolite data

- Concentration levels of 130 metabolic compounds
- Highly correlated blocks of metabolites
- High repeatabilities (mean $w^2 = 0.73$)
- Only weak correlations with GCA
Prescreening for GCA of biomass and quality traits

Genetic architecture of metabolites

- GWA mapping precisely localized for 26 metabolites QTL which explained up to 32% of genetic variance (FDR ≤ 2.5%)

- Highly plausible candidate genes for biomass-correlated metabolites (e.g. lignin synthesis enzyme for several lignin precursors)
Prescreening for GCA of biomass and quality traits

Prediction model: RR-BLUP

Example GCA for dry matter yield

RR-BLUP: normally distributed genetic effects

Genetic architecture of agronomic traits

GWAS: no QTL which explains a reasonable amount of variance
### GCA prediction: results

<table>
<thead>
<tr>
<th>GCA</th>
<th>$h_{GCA}^2$</th>
<th>Accuracies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SNPs</td>
</tr>
<tr>
<td>Dry matter yield</td>
<td>0.89</td>
<td>0.78</td>
</tr>
<tr>
<td>Plant height</td>
<td>0.95</td>
<td>0.72</td>
</tr>
<tr>
<td>Dry matter conc.</td>
<td>0.96</td>
<td>0.80</td>
</tr>
<tr>
<td>Female flowering</td>
<td>0.98</td>
<td>0.81</td>
</tr>
<tr>
<td>Starch content</td>
<td>0.93</td>
<td>0.73</td>
</tr>
<tr>
<td>Sugar content</td>
<td>0.94</td>
<td>0.72</td>
</tr>
<tr>
<td>Lignin content</td>
<td>0.82</td>
<td>0.80</td>
</tr>
</tbody>
</table>

- Prediction accuracies decline only up to ≈ 20 % if all closely related lines are removed to get an unstructured 'core set'
Influence of genetic architecture

- Same population
- **6 model traits with contrasting genetic architecture**
  - 3 polygenic agronomic traits (line per se)
  - 3 metabolites each having different major QTL which explain 22 - 30 % of genetic variance
- 5 prediction models with different assumptions:
  - RR-/G-BLUP
  - LASSO
  - Elastic net
  - RKHS
  - BayesB
Influence of genetic architecture

![Graphs showing SNP effects across chromosomes for RR-BLUP, LASSO, and Elastic net methods](image-url)
**Table:** Examples of prediction accuracies of traits with polygenic (blue) and oligogenic (red) genetic architecture

<table>
<thead>
<tr>
<th>Trait</th>
<th>RR-/G-BLUP</th>
<th>LASSO</th>
<th>Elastic net</th>
<th>RKHS</th>
<th>BayesB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Per se yield</td>
<td>0.61</td>
<td>0.51</td>
<td>0.56</td>
<td>0.61</td>
<td>0.59</td>
</tr>
<tr>
<td>Dopamine</td>
<td>0.74</td>
<td>0.79</td>
<td>0.79</td>
<td>0.74</td>
<td>0.75</td>
</tr>
</tbody>
</table>

- Best performance
  - **Polygenic:** RR-/G-BLUP
  - **Oligogenic:** LASSO or Elastic net

- **Largest differences in accuracies across methods:** only 0.05 - 0.14
Model guidance if genetic architecture is unknown

- What if genetic architecture is unknown or unclear?
- **Genome partitioning of genetic variance:**
  \[ y = 1\mu + Q\beta + \sum_{c=1}^{10} (Sg_c) + e \]
  - \( Q \) contains first 10 PCs to correct for pop. structure
  - For chromosome \( c \): Genotypic effects \( g_c \sim N(0, G_c\sigma^2_{gc}) \) with \( G_c = Z_cZ_c^T/(\#SNPs)\), \( S \) maps \( y \) to \( g_c \)
• 100 Dent + 97 Flint lines
• Trait: NCLB with $h^2 = 0.76$ for Dent and 0.64 for Flint
Combining heterotic groups?

Combining heterotic pools?

- High proportion of markers with equal linkage phase across heterotic groups up to several Mb
**Table:** Mean accuracies after 100 repetitions with $N_t = 75$

<table>
<thead>
<tr>
<th>Training set</th>
<th>Prediction set</th>
<th>$r_{gg}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dent</td>
<td>Dent</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>Flint</td>
<td>0.29</td>
</tr>
<tr>
<td>Flint</td>
<td>Dent</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>Flint</td>
<td>0.61</td>
</tr>
<tr>
<td>Combined</td>
<td>Dent</td>
<td><strong>0.71</strong></td>
</tr>
<tr>
<td></td>
<td>Flint</td>
<td><strong>0.69</strong></td>
</tr>
</tbody>
</table>
Conclusions: Predicting GCA in diverse collections

• High accuracies for predicting GCA of diverse but elite lines of one heterotic pool
• Foundation: large genetic variance!
• Metabolites: highly repeatable, often genetically simple intermediated phenotypes with high predictive value
• Prediction models differ only slightly, irrespective of the underlying genetic architecture of the target trait
• Guideline if trait architecture is unclear or unknown: → Inspect curvature of cumulative distribution of genetic variance over chromosomes
• If linkage phases are consistent, pools can be combined into one training population
• Factorials are incomplete
• SNPs connect tested with untested hybrids
• → 'Fill up the gaps'
Different types of untested hybrids depending on no. of parents represented in the training population:

- **T2**: Both parents
- **T1**: Only one parent
- **T0**: No parents
We used existing genotypes as the basis for simulations

- **100 Dent** *(red)* and **100 Flint** *(blue)* lines from breeding program of University of Hohenheim
- Illumina MaizeSNP50 BeadChip
Data simulation

- **All possible 10,000 interpool hybrids generated *in silico***

- Assume that we have $75 \times 75 = 5625$ evaluated T2 hybrids (*green*)

- **Training set**: 800 randomly drawn T2 hybrids (**)

- **Validation set**: Remaining T2 (*green*), T1 (*blue*) and T0 (*red*) hybrids
Introduction

Diverse collections

Hybrid prediction

Simulation

Data simulation

Genotypic values

• Sample 300 SNPs based on 2 contrasting scenarios concerning allele frequencies:
  • **convergent pools**: \(|p^{Dent} - p^{Flint}| < 0.05\)
  • **divergent pools**: \(|p^{Dent} - p^{Flint}| > 0.60\)

• Assign for 250 SNPs:
  • **Additive effects** drawn from $\text{Gamma}(k = 0.4, \theta = 1.66)$
  • **Dominance effects** = Additive effects $\times$ degree of dominance drawn from $\text{N}(1.0, 0.75)$

• Pure dominance effects assigned to remaining 50 SNPs

Phenotypic values

• Random noise to arrive at $h^2 = 0.75$

• Complete simulation process repeated 50 times
Hybrid prediction model

Translation classical GCA-SCA model → SNP based hybrid prediction model

Classical GCA-SCA model:

\[ y_{ij} = \mu + GCA_d^i + GCA_f^j + SCA_{ij} \]

SNP based hybrid prediction model

\[ y = 1_\mu + Z_d u_d + Z_f u_f + D_{df} d_{df} + \epsilon \]

where:

- \( Z_d, Z_f \) and \( D_{df} \) are design matrices for SNP effects
- \( u_d, u_f \) and \( d_{df} \) are the SNP effects
Hybrid prediction model

\[ y = 1 \mu + Z_d u_d + Z_f u_f + D_{df} + e \]

- \( Z_d \) corresponds to **Dent parent** of hybrid
- \( Z_f \) corresponds to **Flint parent** of hybrid
- \( D_{df} \) codes **heterozygosity** of Dent \( \times \) Flint combination
- Model fitted using RR-/G-BLUP and BayesB

\[
Z_d = \begin{bmatrix}
1 & -1 & -1 \\
1 & -1 & -1
\end{bmatrix}
\]

\[
Z_f = \begin{bmatrix}
-1 & -1 & 1 \\
1 & -1 & -1
\end{bmatrix}
\]

\[
D_{df} = \begin{bmatrix}
1 & 0 & 1 \\
0 & 0 & 0
\end{bmatrix}
\]

- one row per hybrid
- one column per marker
## Results: BayesB vs. RR-/G-BLUP

<table>
<thead>
<tr>
<th>Pools</th>
<th>Group</th>
<th>Method</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BayesB</td>
<td>RR-/G-BLUP</td>
<td></td>
</tr>
<tr>
<td>divergent</td>
<td>T2</td>
<td>0.95</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td></td>
<td>T1</td>
<td>0.90</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td></td>
<td>T0</td>
<td>0.84</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>convergent</td>
<td>T2</td>
<td>0.91</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td></td>
<td>T1</td>
<td>0.84</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td></td>
<td>T0</td>
<td>0.76</td>
<td>0.74</td>
<td></td>
</tr>
</tbody>
</table>
## Results: RR-/G-Blup with dominance yes/no

<table>
<thead>
<tr>
<th>Pools</th>
<th>Group</th>
<th>Dominance ignored</th>
<th>Dominance incorporated</th>
</tr>
</thead>
<tbody>
<tr>
<td>divergent</td>
<td>T2</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>T1</td>
<td>0.88</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>T0</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>convergent</td>
<td>T2</td>
<td>0.86</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>T1</td>
<td>0.79</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>T0</td>
<td>0.71</td>
<td>0.74</td>
</tr>
</tbody>
</table>
Conclusions: Hybrid prediction

- **Accuracies generally T2 > T1 > T0**
- BayesB gives slightly higher accuracies than G-BLUP
- Incorporating dominance effects improves accuracies more under convergent than under divergent pools
  **Reason:** Contribution of SCA variance much larger under convergent pools
- Incorporating pool specific effects improves accuracies only under low number of SNPs
  **Reason:** With 5,000 markers, still high LD across pools and consistent linkage phases despite > 500 years of separation
## Summary

### Take-home messages

- High accuracies for predicting GCA of diverse lines from one pool with large genetic variance
- Metabolites interesting for both trait dissection and prediction
- If linkage phases are consistent, pools can be combined into one training population
- Prediction of single-cross hybrids may reduce necessity of testcrossing
- Sophisticated models only pay off under special circumstances, e.g. extreme genetic architecture, high SCA-to-GCA ratio


• Riedelsheimer, C. et al. Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines *BMC Genomics* 13:452 (2012)


• Technow, F. et al. Genomic prediction of Northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. *G3:Genes/Genomes/Genetics* 3:197-203 (2013)
Acknowledgements

- **UHOH**: A. E. Melchinger, C. Grieder, F. Technow, T. Schrag, H.F. Utz
- **MPI-MP**: M. Stitt, L. Willmitzer, A. Czedik-Eysenberg, R. Sulpice, J. Lisec
- **IPK**: T. Altmann