Comparison of strategies for discovery and utilization of useful genetic material from maize landrace populations

Gregor Gorjanc, Janez Jenko, Sarah J Hearne, John M Hickey
Seeds of Discovery project (SeeD)
Breeders “avoid” landraces

Historical U.S. Corn Grain Yields
1866 to date

Data source: USDA-NASS
2012 yield est. as of Aug 2012

Scope

Discovery → Improvement

Scope of the study

Landraces → Pre-bridging germplasm → Bridging germplasm → Elite germplasms
The plan

1) Genotype gene bank accessions

2) Collect test-cross phenotypes

3) Genomic evaluation and selection

4) Create synthetic population with useful genetic material for further improvement
How to handle this hairball?

- Decide on a handful of possible approaches
- Utilize simulation to identify the best approach and key influential factors

Gorjanc et al. BMC Genomics (2016) 17:30
DOI 10.1186/s12864-015-2345-z

RESEARCH ARTICLE

Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations

Gregor Gorjanc¹,²*, Janez Jenko²,³, Sarah J. Hearne⁴ and John M. Hickey²
Simulation setup

- Mimic the SeeD project with AlphaSim

Maize genome

Landraces

Proposed approaches
Simulation – factors?

- **Effective population size** \(\Rightarrow\) genomic variance \((N_e=1K, 100K)\) #2
- **Landrace variance** \((F=0.3, 0.9)\) #2
- **Heritability** \((h^2=0.25, 0.50)\) #2
- **Approach** \((\text{Landrace, LandraceDH, LandraceElite})\) #3
- **Genotyping platform** \((\text{GBS10x10K, GBS1x100K})\) #2
- **Landrace seeds in training** \((n_{Seed}=1, 3, 5)\) #3
- **Selection of landraces** \((n=40, 80, k=10, 20, 40)\) #6
- **Retrain in cycles** \((n=0, 20, 40, 60)\) #4

Scenarios: \(2\times2\times2\times3\times2\times3\times6\times4 = 2592 \times 10\) replicates
Maize genome

- Coalescent simulation per chromosome
  - two heterotic groups (A and B)
  - allocate QTLs and SNPs
  - rank haplotypes by true breeding value (TBV)
- the best for testers (A, B → A×B)
- lowest half for landraces
Maize genome II

- Coalescent simulation per chromosome
  - Effective population size ($N_e=1K, 100K$)
  - Mutation rate $1.0 \times 10^{-8}$
  - Recombination rate $1.0 \times 10^{-8}$

- $N_e=100K$
  - Seg. sites: $27 \times 10^6$
  - Disk usage: 152 GB

- $N_e=10K$
  - Seg. sites: $10 \times 10^6$
  - Disk usage: 100 GB

- $N_e=1K$
  - Seg. sites: $5 \times 10^6$
  - Disk usage: 28 GB
• 3000 landraces via gene dropping

  – global (between) variance already due to large $N_e$
  – between and within variance controlled by genetic drift (random mating of $2N_p$ individuals for $t$ generations)

\[ N_p = \frac{1}{4(1-(1-F_t)^{(1/t)})} \]

<table>
<thead>
<tr>
<th>$F_t$</th>
<th>$2N_p$</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>0.9</td>
<td>2</td>
<td>9</td>
</tr>
</tbody>
</table>
Test-cross genotypes

- Additive polygenic model (2000 QTL per chromosome from Gaussian dist.)

- Cross with a hybrid-tester and obtain individual true breeding values and phenotypic values

- Phenotype = mean test-cross performance with heritability (h²=0.25, 0.50)
Genomic selection model

- Genomic evaluation using a model linking phenotypic values to genomic markers to obtain a genomic estimate of a breeding value (gEBV)

- Ridge regression

\[
y | \mu, b, \sigma_e^2 \sim N (1\mu + Xb, I\sigma_e^2) \\
b | \sigma_m^2 \sim N (0, I\sigma_m^2)
\]
Proposed approaches - Landrace

3,000 landraces

Ne = 1,000 or 100,000
F within landraces = 0.3 or 0.9

Select the best 40 or 80 landraces

Plant a random selection of 1, 3, or 5 seeds per landrace for training

GENOTYPE GBS = 10K@10x or 100K@1x

Landrace 1

Landrace 2

Landrace 3,000

Elite hybrid

Select at random 10, 20, or 40 seeds from each selected landrace

GENOTYPE GBS = 10K@10x or 100K@1x

ESTIMATE BREEDING VALUE

Advance the best 10 seeds – one per landrace

Plant a random selection of 10 seeds in two replicates

PHENOTYPE $h^2 = 0.25$ or 0.50

TRAIN THE MODEL (regress "testcross" phenotypes on the landrace seed genotypes) → ESTIMATE BREEDING VALUE OF LANDRACES
Approaches - discovery

- Landrace
- LandraceDH
- LandraceElite
Approaches - improvement

1  2  …  10

Advance  Train

M  L  T  P
Metrics of interest?

• Metrics:
  – genetic merit
    (mean of true breeding values scaled between 0 (landrace average) and 1 (elite hybrid))
  – kinship with the elite hybrid
    (proportion of genome equal to the genome of a tester)
  – accuracy
    (correlation between the true and estimated value)

• Where (stage of simulation):
  – ???
Stage of simulation

1. Selecting accessions (A)
2. Sel. seeds within accessions (S)
3.-6. Synthetic population cycles (C1, C2, C3, and C4)

0. Training (T)
How to present results?

• Results
  – 2592 scenarios × 10 replicates
  – 6 stages of simulation
  \[25,920 \times 6 \approx 150,000\] values per metric

• Show the effect of **individual** factors

• Show the effect of the most influential factors and their **interactions** with the use of **regression trees**
Genetic merit

Approach
- Landrace
- LandraceDH
- LandraceElite

Stage
- T
- A
- S
- C1
- C2
- C3
- C4
Genetic merit

\[ N_e = 1,000 \quad N_e = 100,000 \]

\[ F = 0.3 \quad F = 0.9 \]

\[ h^2 = 0.25 \quad h^2 = 0.50 \]
Kinship with elite
Kinship with elite
(assume no recombination for a demo)

4/16 matches
→ Probability that gametes are equal = ¼
### Accuracy

<table>
<thead>
<tr>
<th>Approach</th>
<th>Stage</th>
<th>$F=0.3$</th>
<th>$F=0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lamheno</td>
<td>T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LamhenoElite</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LamhenoElite</td>
<td>S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LamhenoElite</td>
<td>C1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LamhenoElite</td>
<td>C2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LamhenoElite</td>
<td>C3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LamhenoElite</td>
<td>C4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### $N_e=1,000$ & $F=0.3$ & $h^2=0.25$

#### $N_e=1,000$ & $F=0.9$ & $h^2=0.50$

#### $N_e=100,000$ & $F=0.3$ & $h^2=0.25$

#### $N_e=100,000$ & $F=0.9$ & $h^2=0.50$
Fig. S9. Decision tree for accuracy of selecting accessions in the discovery phase with the Landrace approach, high diversity of the founding population ($Ne=100,000$), and high diversity within accessions ($F=0.3$).
Decision tree for accuracy in the final stage of improvement phase with the Landrace approach, high diversity of the founding population ($N_e=100,000$), and high diversity within accessions ($F=0.3$).
Conclusions (recommendations)

• **Approach**
  - do not use the LandraceElite approach
  - more gain with the LandraceDH approach but additional seasons → not worthwhile

• **Genotyping platform**
  - larger no. of markers at lower coverage seems to be better due to a large $N_e$

• **The way landraces are selected not that influential**
  - #Landraces and #individuals/landrace screened

• **Ideally if done again test more seeds per landrace**
  - 3 per landrace likely enough (given the resources)

• **Need to retrain when developing synthetic pop.**
  - 40 phenotyped individuals per cycle likely enough
Comparison of strategies for discovery and utilization of useful genetic material from maize landrace populations

Gregor Gorjanc, Janez Jenko, Sarah J Hearne, John M Hickey