Implementing genomic selection and comparing it to marker-assisted selection

Alexander E. Lipka
Assistant Professor of Biometry
Department of Crop Sciences
University of Illinois
USA
Genome-wide association study (GWAS)

Association with Vitamin E Levels in Maize Grain

Marketers exhibiting peak associations with traits are potential targets for marker-assisted selection (MAS)

- Identify genomic regions associated with a phenotype
- Fit a statistical model at each SNP in genome
- Use fitted models to test $H_0$: No association with SNP and phenotype
Examples of GWAS identifying potential targets for MAS breeding efforts

- Rincker et al. (2016): Targets for brown stem rot resistance in soybean
- Lipka et al. (2013): Targets for boosting vitamin E and antioxidant levels in maize grain
- Owens/Lipka et al. (2014): Targets for boosting provitamin A and other carotenoid levels in maize grain
Genetic diversity can lead to false positives in a GWAS

- Two sources for false positives:
  - Population Structure
  - Familial Relatedness

Genetic Diversity of 2,815 Maize Inbreds

Romay et al. (2013)
Mixed models reduce false positives in GWAS

\[ Y_i = \mu + \sum_{j=1}^{3} \beta_j PC_{ji} + \alpha x_i + \text{Line}_i + \varepsilon_i \]

- **Grand Mean**
- **Marker effect**
- **Fixed effects: account for population structure**
- **Observed SNP alleles of \( i^{th} \) individual**
- **Random effects: account for familial relatedness**
- **Random error term**

- \( (\text{Line}_1, \ldots, \text{Line}_n) \sim \text{MVN}(0, 2K\sigma_G^2) \)
- \( K = \text{kinship matrix} \)
- \( \varepsilon_i \sim \text{i.i.d. } \text{N}(0, \sigma_E^2) \)

Yu et al. (2006)
Computational approaches for reducing computational burden

• The unified mixed linear model is a common approach for GWAS:
  – Effectively accounts for population structure and familial relatedness
  – Reduces false positives

• Resulting computational challenges:
  – A typical GWAS of ~2,000,000 SNPs with standard model fitting approaches can be impractical
  – Newly-developed model fitting approaches need to be used to address this challenge

GAPIT R package (Lipka et al. 2012):
• Employs computationally-efficient approaches for GWAS
• Makes it possible to perform mixed-model GWAS on an ordinary computer

  – Newly-developed model fitting approaches need to be used to address this challenge
Unified mixed linear model (MLM)

\[ Y = \mu + \sum_{i=1}^{3} \beta_i PC_i + \alpha X + Line_i + \epsilon_i \]

- **Grand Mean**
- **Marker effect**
- **Random effects:**
  - account for familial relatedness

- Variance component estimation is computationally intensive
- **GAPIT** employs two approaches to reduce this computational burden
  - \((Line_1, ..., Line_n) \sim \text{MVN}(0, \Sigma K \sigma^2_G)\)
  - \(K = \text{kinship matrix}\)
  - \(\epsilon_i \sim \text{i.i.d. N}(0, \sigma^2_E)\)

Yu et al. (2006)
Approach 1: Compressed mixed linear model

\[ Y_i = \mu + \sum_{j=1}^{3} \beta_j PC_{ji} + \alpha x_i + \text{Group}_i + \varepsilon_i \]

- Reduces computational time because it works with a smaller kinship matrix

Using kinship matrix

- \((\text{Group}_1, \ldots, \text{Group}_k) \sim \text{MVN}(0, K_C \sigma^2_G)
- K_C = \text{kinship ("compressed") kinship matrix}
- \varepsilon_i \sim \text{i.i.d. N}(0, \sigma^2_E)

Zhang et al. (2010)
Approach 2: Population parameters previously determined (P3D)

Output Summary

Lipka et al. (2012)
Zhang et al. (2010)
Accounting for multiple reps and locations

Seeds obtained from a germplasm bank

- Fit a mixed model accounting for genetic, environmental, and genetic x environmental (GxE) sources of trait variation
- Output from this model:
  - BLUPs/BLUEs trait values for each taxa
  - Estimates of trait variation attributable to each source

multiple reps and locations?
Statistical model used to obtain best linear unbiased predictions (BLUPs)

\[ Y_{iik} = \mu + G_i + E_i + (GE)_{ij} + \epsilon_{iik} \]

- **Output 1**: BLUPs of the genotype effect
- **Output 2**: Variance component estimates for calculating heritabilities

- \( G_i \) = Random Genotype Effect
- \( E_i \) = Random Environment Effect
- \( (GE)_{ij} \) = Random Genotype x Environment Effect
Statistical model used to obtain best linear unbiased estimators (BLUEs)

\[ Y_{ijk} = \mu + G_i + E_j + (GE)_{ij} + \varepsilon_{ijk} \]

- **Output:** BLUEs of the genotype effect
  
- **Fixed Effect**
  - \( G_i \) = Fixed Genotype Effect
  - \( E_i \) = Random Environment Effect
  - \( (GE)_{ij} \) = Random Genotype x Environment Effect
  
- **Random Effect**
  - \( \mu \)
  - \( \varepsilon_{ijk} \)

Grand Mean

Fixed

Random effect
BLUPs vs BLUEs

• BLUPs:
  – Advantage: Makes more sense from a biological perspective
  – Disadvantage 1: BLUPs “shrink” values towards the mean
  – Disadvantage 2: Fitting random effects is more computationally intensive than fitting fixed effects

• BLUEs:
  – Advantage 1: BLUEs do not shrink values towards the mean
  – Advantage 2: Less computationally intensive
  – Disadvantage: Makes less sense from a biological perspective
BLUPs and BLUEs: Some Technical Notes

• In plant breeding, estimate of grand mean is added to BLUPs and BLUEs
  – Rationale: BLUEs/BLUPs will be in the same units of measurement as raw trait data
  – After adding grand mean estimate, they are still called BLUPs/BLUEs

• Consider transforming your phenotypic data before fitting statistical models:
  – Rationale: This would help with deviations from normality and constant variance assumptions
Software I used to obtain BLUPs and BLUEs

• SAS:
  – Advantage: (Relatively) simple to use
  – Disadvantage 1: Annual license fee
  – Disadvantage 2: Takes a long time to compute

• ASReml:
  – Advantage: Can fit very complicated models quickly
  – Disadvantage 1: Not simple to use
  – Disadvantage 2: Expensive annual license fee

• R:
  – Advantage: Free
  – Disadvantage: Potentially not as extensively tested as SAS and ASREML
Phenotype: kernel color visually assessed using standardized color scale

- Also included AR1xAR1 correlation structure to account for spatial variation
- Backwards elimination conducted to remove non-significant effects
- Analysis conducted in ASREML
Example: Rincker et al. (2016)

- Brown stem rot (BSR) and
  - Three genes associated with BSR resistance, $Rbs1$-3, have been identified in previous studies
  - Critical need to obtain a more precise location of these loci
  - Result in more efficient MAS for BSR resistance

Source: cornandsoybeandigest.com/
Separate GWAS performed on four association panels

Table 1. Characteristics of association panels analyzed with genome-wide association study and stepwise procedures.

<table>
<thead>
<tr>
<th>Panel</th>
<th>Data type</th>
<th>Symptoms measured</th>
<th>Accessions</th>
<th>SNP† markers</th>
<th>Box-Cox lambda</th>
<th>BSR Score†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>N-1989</td>
<td>Binary</td>
<td>Foliar and stem</td>
<td>2773</td>
<td>33,240</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>B-1997</td>
<td>Proportion 0–1</td>
<td>Foliar</td>
<td>540</td>
<td>33,486</td>
<td>log</td>
<td>0.09</td>
</tr>
<tr>
<td>B-1997</td>
<td>Proportion 0–1</td>
<td>Stem</td>
<td>540</td>
<td>33,486</td>
<td>1</td>
<td>0.38</td>
</tr>
<tr>
<td>B-2000</td>
<td>Proportion 0–1</td>
<td>Foliar</td>
<td>825</td>
<td>32,150</td>
<td>0.25</td>
<td>0.33</td>
</tr>
<tr>
<td>P-2003</td>
<td>Proportion 0–1</td>
<td>Stem</td>
<td>606</td>
<td>29,815</td>
<td>0.75</td>
<td>0.39</td>
</tr>
</tbody>
</table>

- N-1989 panel:
  - Binary phenotype: logistic regression + stepwise model selection
- Other panels:
  - Quantitative phenotype: Unified MLM + multi-locus mixed model

Rincker et al. (2016)
Unified MLM GWAS identifies signals near *Rbs1*-*Rbs3*

- Multi-locus mixed model identified two peak SNPs from this region in the final model
- GWAS was reran using these two peak SNPs as covariates

Rincker et al. (2016)
Peak SNPs from MLMM reduces explains most of \textit{Rbs1}-\textit{Rbs3} signal

- Similar findings were obtained in the other association panels
Breeding Ramifications

- Previous $Rbs1-Rbs3$ signals been refined to a 0.3 Mb region on Chromosome 16
- Should facilitate both MAS-based approaches and gene cloning efforts
- Demonstrates the utility of GWAS in soybean

Source: blogs.ext.vt.edu

Rincker et al. (2016)
Biofortification

• Identify target genes associated with nutrients in crops

• Increase nutritional value of local crop varieties by selecting on these target genes

• Results in increased availability of essential nutrients

Source: www.aboutharvest.com
Compounds analyzed in Lipka et al. (2013)

- **Tocochromanols**
  - Lipid-soluble antioxidants
  - Consist of *tocopherols* (T) and *tocotrienols* (T)
  - \( \alpha \)-tocopherol (\( \alpha \)T) has greatest vitamin E activity

- **Vitamin E**
  - Essential nutrient
  - Suboptimal dietary intake exists in specific population segments
  - Deficiency associated with cardiovascular disease and decreased immune function
Grain tocochromanol compositions across a maize diversity panel

Distribution of Tocochromanol Compounds

- Boost vitamin E levels by increasing α-tocopherol concentration
Data analyzed in Lipka et al. (2013)

- 281-member Goodman diversity panel
- Grown at Purdue University in 2009 and 2010 field seasons
- Compound levels quantified in grain:
  - Tocochromanols for 252 lines
Phenotypic data used for analysis

• High-pressure liquid chromatography (HPLC) used to measure tocochromanol levels in maize grain
• Mixed model accounting for field season effects fitted to each phenotype
• Best linear unbiased predictors (BLUPs) of lines from each model used as phenotypes for our GWAS

Source: www.ssi.shimadzu.com

Torbert Rocheford

• 20 tocochromanol compounds, sums, ratios, and proportions were analyzed in GAPIT
• GWAS was conducted using 294,092 SNPs with minor allele frequency ≥ 0.05 fitted to each phenotype
• Best linear unbiased predictors (BLUPs) of lines from each model used as phenotypes for our GWAS
In-class example: GWAS scan of Lipka et al. (2013) data subset

- Trait: α-tocopherol
  - Has the greatest Vitamin E activity
- Marker subset:
  - 3,093 marker set obtained from various marker technologies (i.e., the 4k marker set)
- GWAS software used: Genome association and prediction integrated tool (GAPIT)
  - Unified mixed linear model is fitted at each SNP
  - Population parameters previously determined (P3D) used to save computational time
In-class example: GWAS scan of Lipka et al. (2013) data subset

- 4K_SNPsmdp_genotype_test1_GBS_Names1.hmp.txt
  - Genotypic data: 3,093 SNPs
- alpha.tocopherol.BLUPs_No_Outliers.transformed.txt
  - Phenotypic data: α-tocopherol levels
- Scripts_Necessary_for_GAPIT
  - Folder containing scripts to be read into R
- Run_GWAS_on_alpha_tocopherol_4K_SNPs.r
  - R script for conducting the GWAS
In-class example: GWAS scan of Lipka et al. (2013) data subset
In-class example: GWAS scan of Lipka et al. (2013) data subset

- For details on running GAPIT, here is the user manual: http://zzlab.net/GAPIT/gapit_help_document.pdf
In-class example: GWAS scan of Lipka et al. (2013) data subset
In-class example: GWAS scan of Lipka et al. (2013) data subset
In-class example: GWAS scan of Lipka et al. (2013) data subset

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chromosome</th>
<th>Position</th>
<th>P.value</th>
<th>maf</th>
<th>nobs</th>
<th>Rsquare_without.SNP</th>
<th>Rsquare_with.SNP</th>
<th>Effect.Est</th>
<th>FDR_Adjusted_P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>PZB02283.1</td>
<td>5</td>
<td>200,367,532</td>
<td>2.47E-12</td>
<td>0.203187251</td>
<td>251</td>
<td>0.244613724</td>
<td>0.408846946</td>
<td>0.390584717</td>
<td>6.24E-09</td>
</tr>
<tr>
<td>PZB02424.2</td>
<td>5</td>
<td>200,370,309</td>
<td>2.95E-06</td>
<td>0.167330677</td>
<td>251</td>
<td>0.244613724</td>
<td>0.313756139</td>
<td>-0.25623365</td>
<td>0.003722329</td>
</tr>
<tr>
<td>PZB02002.1</td>
<td>3</td>
<td>137,231,734</td>
<td>0.000218352</td>
<td>0.199203187</td>
<td>251</td>
<td>0.244613724</td>
<td>0.287148157</td>
<td>0.1928257</td>
<td>0.183634318</td>
</tr>
<tr>
<td>PZD00015.5</td>
<td>3</td>
<td>137,229,812</td>
<td>0.001045234</td>
<td>0.24501992</td>
<td>251</td>
<td>0.244613724</td>
<td>0.277860945</td>
<td>0.15824979</td>
<td>0.553336778</td>
</tr>
<tr>
<td>PZA00732.2</td>
<td>2</td>
<td>1,187,041</td>
<td>0.001129994</td>
<td>0.101593625</td>
<td>251</td>
<td>0.244613724</td>
<td>0.277405113</td>
<td>0.21571626</td>
<td>0.553336778</td>
</tr>
<tr>
<td>PZA03188.2</td>
<td>1</td>
<td>281,708,766</td>
<td>0.001781959</td>
<td>0.37451992</td>
<td>251</td>
<td>0.244613724</td>
<td>0.274756039</td>
<td>0.137534504</td>
<td>0.553336778</td>
</tr>
<tr>
<td>PZA03322.1</td>
<td>4</td>
<td>236,407,395</td>
<td>0.00020846</td>
<td>0.348605578</td>
<td>251</td>
<td>0.244613724</td>
<td>0.273848712</td>
<td>-0.13420303</td>
<td>0.553336778</td>
</tr>
<tr>
<td>PHM5468.25</td>
<td>8</td>
<td>130,509,443</td>
<td>0.002645262</td>
<td>0.252988048</td>
<td>251</td>
<td>0.244613724</td>
<td>0.272478664</td>
<td>0.145149925</td>
<td>0.553336778</td>
</tr>
<tr>
<td>PZB02215.7</td>
<td>7</td>
<td>9,256,489</td>
<td>0.000277669</td>
<td>0.059760956</td>
<td>251</td>
<td>0.244613724</td>
<td>0.272200495</td>
<td>-0.256858949</td>
<td>0.553336778</td>
</tr>
<tr>
<td>PZA02393.2</td>
<td>1</td>
<td>16,581,396</td>
<td>0.002859692</td>
<td>0.472111554</td>
<td>251</td>
<td>0.244613724</td>
<td>0.272031681</td>
<td>-0.130129573</td>
<td>0.553336778</td>
</tr>
<tr>
<td>PZA00758.1</td>
<td>8</td>
<td>23,769,876</td>
<td>0.003061352</td>
<td>0.199203187</td>
<td>251</td>
<td>0.244613724</td>
<td>0.271641559</td>
<td>-0.145079639</td>
<td>0.553336778</td>
</tr>
<tr>
<td>PZA02060.1</td>
<td>5</td>
<td>203,205,315</td>
<td>0.003261396</td>
<td>0.472111554</td>
<td>251</td>
<td>0.244613724</td>
<td>0.271279274</td>
<td>0.134959714</td>
<td>0.553336778</td>
</tr>
<tr>
<td>PZB01947.1</td>
<td>3</td>
<td>7,600,438</td>
<td>0.003347877</td>
<td>0.183266932</td>
<td>251</td>
<td>0.244613724</td>
<td>0.27113028</td>
<td>-0.161852658</td>
<td>0.553336778</td>
</tr>
<tr>
<td>PZA00022.2</td>
<td>6</td>
<td>85,884,402</td>
<td>0.003386473</td>
<td>0.372509963</td>
<td>251</td>
<td>0.244613724</td>
<td>0.271064853</td>
<td>-0.12155698</td>
<td>0.553336778</td>
</tr>
<tr>
<td>PZB01993.3</td>
<td>5</td>
<td>7,871,700</td>
<td>0.003424801</td>
<td>0.115537857</td>
<td>251</td>
<td>0.244613724</td>
<td>0.271000634</td>
<td>0.166582084</td>
<td>0.553336778</td>
</tr>
<tr>
<td>PZB01725.2</td>
<td>1</td>
<td>267,887,581</td>
<td>0.003509072</td>
<td>0.235059761</td>
<td>251</td>
<td>0.244613724</td>
<td>0.270861986</td>
<td>-0.13178931</td>
<td>0.553336778</td>
</tr>
<tr>
<td>PZA00590.1</td>
<td>2</td>
<td>22,069,205</td>
<td>0.003904833</td>
<td>0.239043825</td>
<td>251</td>
<td>0.244613724</td>
<td>0.270253433</td>
<td>0.156250018</td>
<td>0.575777451</td>
</tr>
</tbody>
</table>
GWAS identified signals near two biosynthetic pathway genes

- Peak SNP within \textit{ZmVTE4} (\( P \)-value = \( 7.36 \times 10^{-14} \))
- \textit{ZmVTE4} has been previously identified

- Peak SNP located 70 bp from \textit{ZmVTE1} start site (\( P \)-value = \( 1.29 \times 10^{-7} \))
- We are the first to identify \textit{ZmVTE1} in a maize association panel

\textit{Lipka et al. (2013)}
ZmVTE4 and ZmVTE1 are important genes

- Possible to develop maize grain with enhanced vitamin E and antioxidant levels via marker-assisted selection of ZmVTE4 and ZmVTE1
Elucidating the association between αT and ZmVTE4

- Short-range LD decay with peak SNP
- Significant GWAS signals up to 4,000,000 bp away from ZmVTE4

Lipka et al. (2013)
Stepwise model selection identified two other \( ZmVTE4 \) SNPs associated with \( \alpha T \)

\( \gamma \)-tocopherol methyltransferase (\( ZmVTE4 \))

\( \uparrow \) = SNP identified in GWAS

\( \downarrow \) = SNP identified in stepwise model selection (developed in Segura et al., 2012)

- \( ZmVTE4 \) signal explained by three SNPs
- 5.76-fold change in \( \alpha T \) levels between most and least favorable haplotypes of these three SNPs

Lipka et al. (2013)
Including three \textit{ZmVTE4} SNPs as covariate removes signal

Three \textit{ZmVTE4} SNPs explain the complex association signals in this region

\cite{lipka2013}
Targeting vitamin A deficiency through biofortification

- Vitamin A deficiency (VAD):
  - Affects 17-30% of children under 5
  - 250-500,000 children become blind every year
  - Infant morbidity and mortality

- Maize is a primary food source in many vitamin A deficient regions

- Biofortification: breed locally-adapted maize lines for increased provitamin A levels in grain

Source: en.wikipedia.org
Work in maize provitamin A biofortification prior to Owens/Lipka et al. (2014)

- Candidate gene studies identified loci in maize (Harjes et al., 2008; Vallabheneni et al., 2010; Yan et al. 2010)

Owens/Lipka et al (2014):
1.) Conduct an GWAS to identify new candidate genes
2.) Determine a minimal marker set to accurately predict carotenoid levels

Pleiotropy identified among metabolite QTL (Kandianis et al., 2013)
Data analyzed in Owens/Lipka et al. (2014)

- Maize lines with white kernels do not produce measurable carotenoids
- We only analyzed a subset of 201 lines that range from light yellow to dark orange kernel color

- Compound levels quantified in grain:
  - Carotenoids for 252 lines
GWAS found significant marker-trait associations near carotenoid pathway genes

- Adjusting for multiple testing at the genome-wide level was conservative
- We also conducted a pathway-level analysis, where only markers near 58 *a priori* genes were considered

Owens/Lipka et al. (2014)
GWAS found significant marker-trait associations near carotenoid pathway genes

- This work identified potential targets for marker-assisted selection (MAS)
- Are selecting for these target loci sufficient for improving provitamin A content in maize grain?

Owens/Lipka et al. (2014)
Targeted marker subsets for estimating kinship

• Suppose we are testing SNP 1 on chromosome 1 for an association with a trait.

• K_chr model has greater power to detect marker-trait associations in high-LD regions.

  - Similar “leave one chromosome out” approach used for other chromosomes.

Rincent et al. (2014)
Re-evaluated associations using K_chr model

• Previously published GWAS results from two maize diversity panels:
  – Mendelian: Sweet vs. starchy corn
  – Polygenic: Carotenoids and tocochromanols
  – Complex: Flowering time and plant height

• Compared results of the K_chr model to the unified MLM:
  – Did the K_chr model identify signals in “novel genomic regions”?
  – Did the K_chr model identify more statistically significant associations in high LD regions?

Chen and Lipka (2016)
K_chr identified signals in “novel genomic regions”

Four tocochromanol traits in Goodman diversity panel

Chen and Lipka (2016)
K_chr identified stronger associations in high LD regions

Associations with tocotrienol ratio in vicinity of ZmVTE1

Chen and Lipka (2016)