Genomic Selection Models for Predicting end-use quality traits in CIMMYT spring bread wheat

Diego Jarquin, Reka Howard, Jesse Poland, Sarah Battenfield, Carlos Guzman, Jose Crossa

University of Nebraska Lincoln

Quality traits

- Not the primary breeding objective
- Secondary to yield, agronomic performance, and disease resistance
- Difficult to assess for quality
- Small population sizes can be assessed (expensive and laborious)
- Testing occurs at later stages
- Desirable test for wheat quality in earlier generations

Quality traits

- Alveograph ratio of height to length of the curve
- Alveograph work value under the curve
- Flour proteins
- Flour sodium dodecyl sulfate sedimentation
- Mixograph mix time
- Loaf volume

- A few of these traits are related with gluten composition
  - Strong pan breads
  - Medium bread and noodles
  - Weak cakes, cookies and pastries
  - Tenacious only acceptable as wheat for animal feed

Prediction of Complex Traits: The Nature of the Problem

- Many important traits/diseases in humans, plants and animals are complex

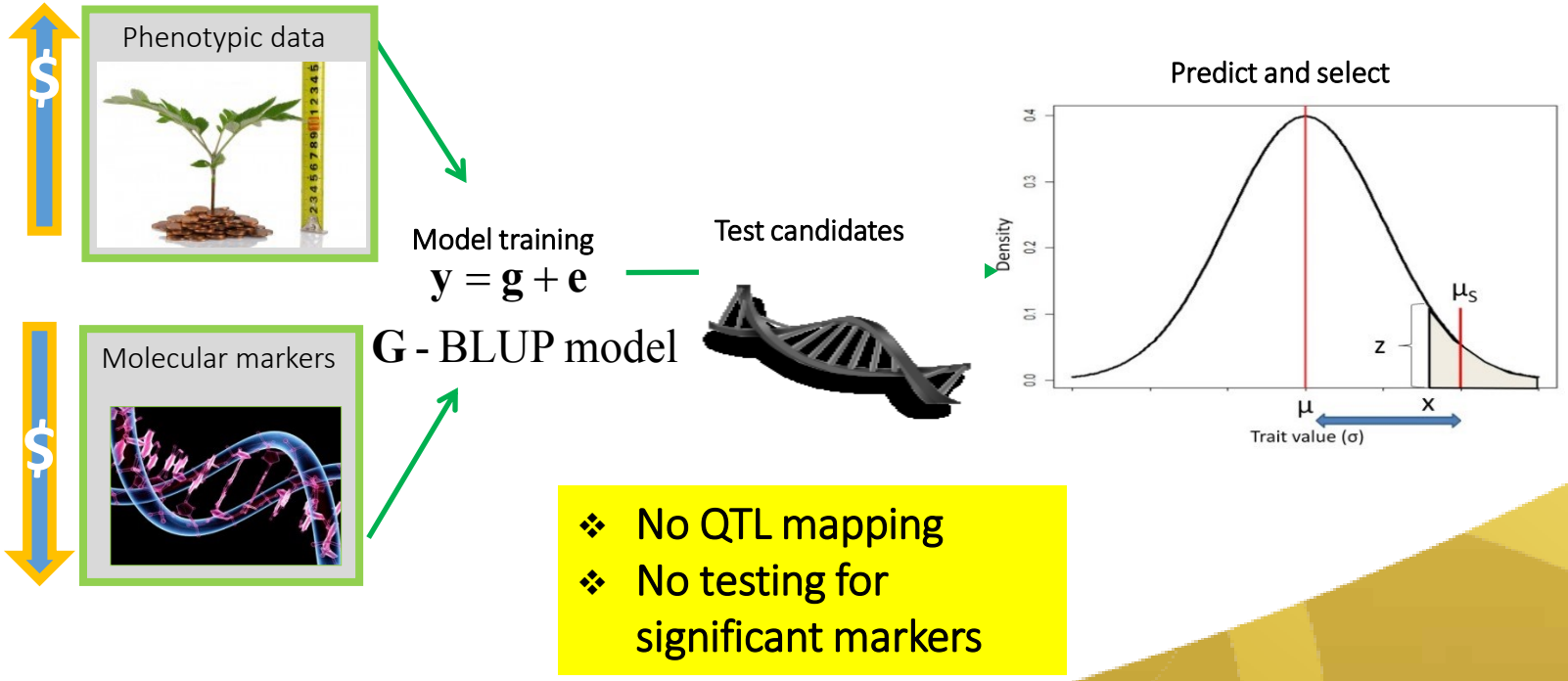
- They are affected by large numbers of small-effect genes and large number of environmental factors

- Genes and environmental factors may interact in complex ways

Genomic Selection (GS) in a Nutshell:

✦ Whole Genomic Prediction:

- ❖ Combines genotypic and phenotypic information to calibrate models and perform prediction on un-phenotyped individuals using molecular markers



- ❖ No QTL mapping
- ❖ No testing for significant markers

Statistical Models in Genetics

Current implemented approaches

GWAS

SNP Selection

Whole Genome Regression & Prediction

Regress Phenotypes on all available markers, jointly

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \qquad (i = 1,\ldots,n) \qquad p >> n$$

Toolkit

- Penalized
  - Ridge Regression (Shrinkage)
  - Elastic Net
  - LASSO (Shrinkage and Selection)
- Bayesian — The 'Bayesian Alphabet'
  - Bayesian Ridge Regression (shrinkage)
  - Bayes A
  - Bayesian LASSO
  - Bayes B/C (selection & shrinkage)

Simplest Case:

✧ Model

$$y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \qquad \varepsilon_i \overset{IID}{\sim} \mathrm{N}(0,\sigma^2) \qquad p >> n$$

Estimation methods

Penalized regressions

$$(\hat{\mu},\hat{\beta})_{\arg\min} =$$

$$(\hat{\mu},\hat{\beta})_{\arg\min} = \left\{\sum_{i=1}^{n}\left(y_i - \mu - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda J(\boldsymbol{\beta})\right\}$$

$$\left\{\sum_{i=1}^{n}\left(y_i - \mu - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\}$$

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right]^{-1}\mathbf{X}'(y_i - \mu)$$

RR-BLUP

Bayesian regressions

$$\beta_j \overset{IID}{\sim} \mathrm{N}(0,\sigma_\beta^2)$$

$$p(\mathbf{y},\boldsymbol{\beta},\mu \mid \sigma^2,\sigma_\beta^2) \propto$$

$$p(\mathbf{y} \mid \boldsymbol{\beta},\mu,\sigma^2,\sigma_\beta^2)\,p(\boldsymbol{\beta} \mid \sigma_\beta^2)$$

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{X} + \sigma^2\sigma_\beta^{-2}\mathbf{I}\right]^{-1}\mathbf{X}'(y_i - \mu)$$

Bayesian Ridge Regression BRR

✧ Complexities (dealing with highly dimensional matrices *p*×*p*)

GBLUP:

- An alternative parameterization of BRR:

$$y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \qquad \beta_j \overset{IID}{\sim} \mathrm{N}\left(0, \sigma_\beta^2\right) \qquad \varepsilon_i \overset{IID}{\sim} \mathrm{N}\left(0, \sigma^2\right) \qquad p >> n$$

- Replacing:

$$g_i = \sum_{j=1}^{p} x_{ij}\beta_j = \mathbf{x}_i\boldsymbol{\beta} \quad \Rightarrow \quad \mathbf{g} = \{g_i\} = \mathbf{X}\boldsymbol{\beta} \quad \Rightarrow \quad \mathrm{cov}(\mathbf{g}) = \mathrm{cov}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}\mathbf{X}'\sigma_\beta^2$$

$$\mathbf{g} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{X}\mathbf{X}'\sigma_\beta^2\right) = \mathrm{N}\left(\mathbf{0}, \mathbf{G}\sigma_g^2\right) \quad \Longleftrightarrow \quad \mathbf{G} = \frac{1}{p}\mathbf{X}\mathbf{X}' \qquad \sigma_g^2 = p\sigma_\beta^2$$

Kinship matrix

$$p(\mathbf{y}, \mathbf{g} \mid \mu, \sigma^2, \sigma_\beta^2) \propto \prod_{i=1}^{n} \mathrm{N}\left(y_i \mid \mu + g_i, \sigma^2\right) \mathrm{N}\left(\mathbf{g} \mid \mathbf{0}, \mathbf{G}\sigma_g^2\right)$$

- Posterior mode of this model is equivalent to BLUP of g:

$$\hat{\mathbf{g}} = \left[\mathbf{I} + \lambda\mathbf{G}^{-1}\right]^{-1}\left(y_i - \mu\right) = \mathbf{X}\hat{\boldsymbol{\beta}} \quad \Rightarrow \quad \lambda = \sigma^2 / \sigma_g^2$$

- Advantages (dealing with matrices of order *n*×*n*)

Empirical Assessment

13th International Gluten Workshop
LACC IGW

Empirical Assessment - CV1:

- Predicting performance of new developed lines through relationships with others

- c45 → 712
- c46 → 995

Training Size	Testing Size	Total
1138	569	1707
995	712	1707
712	995	1707

Lines	**c45**	**c46**
c45_L1	y	
c45_L2	y	
.	y	
.	y	
.	y	
c45_L712	y	
c46_L1		y
c46_L2		y
.		y
.		y
.		y
c46_L995		y

Empirical Assessment – CV00:

- Predicting performance of unobserved lines in unobserved environments

Lines	**c45**	**c46**
c45_L1	y	
c45_L2	y	
.	y	
.	y	
.	y	
c45_L712	y	
c46_L1		y
c46_L2		y
.		y
.		y
.		y
c46_L995		y

Lines	**c45**	**c46**
c45_L1	y	
c45_L2	y	
.	y	
.	y	
.	y	
c45_L712	y	
c46_L1		y
c46_L2		y
.		y
.		y
.		y
c46_L995		y

- Leaving one environment out at the time

Empirical Assessment – CV00:

- Final goal

Predict c49 (1345)

Using c45, c46, c47 and c48 trials as training set (4975)

Lines	c45	c46	c47	c48	c49
c45_L1	y				
c45_L2	y				
.	y				
.	y				
.	y				
c45_L712	y				
c46_L1		y			
c46_L2		y			
.		y			
.		y			
.		y			
c46_L995		y			
c47_L1			y		
c47_L2			y		
.			y		
.			y		
.			y		
c47_L1622			y		
c48_L1				y	
c48_L2				y	
.				y	
.				y	
.				y	
c48_L1622				y	
c49_L1					y
c49_L2					y
.					y
.					y
.					y
c49_L1345					y

Results

13th International Gluten Workshop
LACC IGW

CV1: Predicting new lines in observed environments

- c45 → 712
- c46 → 995

Lines	c45	c46
c45_L1	y	
c45_L2		
.	y	
.		
.	y	
c45_L712	y	
c46_L1		y
c46_L2		
.		y
.		y
.		
c46_L995		y

GBLUP Model

TRAINING / TESTING	1138 / 569	995 / 712	712 / 995
Flour Protein	0.604	0.602	0.589
Flour SDS	0.666	0.666	0.661
Mixograph Mix Time	0.718	0.715	0.707
Alveograph W	0.697	0.695	0.683
Alveograph P/L	0.476	0.474	0.466
Loaf Volume	0.638	0.634	0.625

Empirical Assessment – CV00:

Lines	c45	c46
c45_L1	y	
c45_L2	y	
.	y	
.	y	
.	y	
c45_L712	y	
c46_L1		y
c46_L2		y
.		y
.		y
.		y
c46_L995		y

Lines	c45	c46
c45_L1	y	
c45_L2	y	
.	y	
.	y	
.	y	
c45_L712	y	
c46_L1		y
c46_L2		y
.		y
.		y
.		y
c46_L995		y

- c45 → 712
- c46 → 995

TRAINING / TESTING	c45 with c46	c46 with c45
Flour Protein	0.394	0.284
Flour SDS	0.433	0.461
Mixograph Mix Time	0.535	0.499
Alveograph W	0.512	0.475
Alveograph P/L	0.323	0.278
Loaf Volume	0.358	0.333

Empirical Assessment – CV00:

- Final goal

Predict c49 (1345)

Using c45-c48 trials

Lines	c45	c46	c47	c48	c49
c45_L1	y				
c45_L2	y				
.	y				
.	y				
.	y				
c45_L712	y				
c46_L1		y			
c46_L2		y			
.		y			
.		y			
.		y			
c46_L995		y			
c47_L1			y		
c47_L2			y		
.			y		
.			y		
.			y		
c47_L1622			y		
c48_L1				y	
c48_L2				y	
.				y	
.				y	
.				y	
c48_L1622				y	
c49_L1					y
c49_L2					y
.					y
.					y
.					y
c49_L1345					y

Models:

GBLUP
Gaussian Kernel
ELNET Elastic Net
RF Random Forest

TRAINING / TESTING	GBLUP	Gauss Kernel	PLSR	ELNET	RF	Average
Flour Protein	0.525	0.53	0.489	0.531	0.425	0.505
Flour SDS	0.578	0.574	0.541	0.582	0.516	0.562
Mixograph Mix Time	0.652	0.651	0.643	0.667	0.636	0.653
Alveograph W	0.646	0.518	0.459	0.491	0.436	0.485
Alveograph P/L	0.492	0.643	0.641	0.656	0.59	0.658
Loaf Volume	0.503	0.518	0.412	0.511	0.39	0.496

Conclusions

The GS models showed sufficient accuracy predicting processing and end-use quality traits

Predictive ability improves over time

Positive effects of increase sample size and training sets

GS can enable better selection for quality traits

GS is heavily favored as a selection tool by increasing selection intensity (it allows screen all available materials)

Thank you