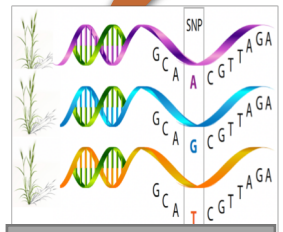
Basics of Genomic Predictions

Waseem Hussain
Postdoctoral Research Asssociate

Genomic Predictions/Selection

Reference/training population:

Phenotyping



Genotyping

Model training

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

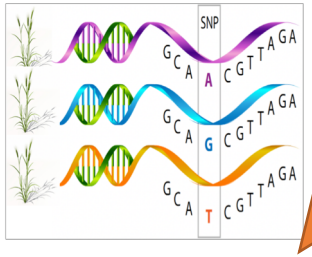
Prediction equation

Genomic Estimated Breeding values

Select the Candidates

Candidate population

Only Genotyping



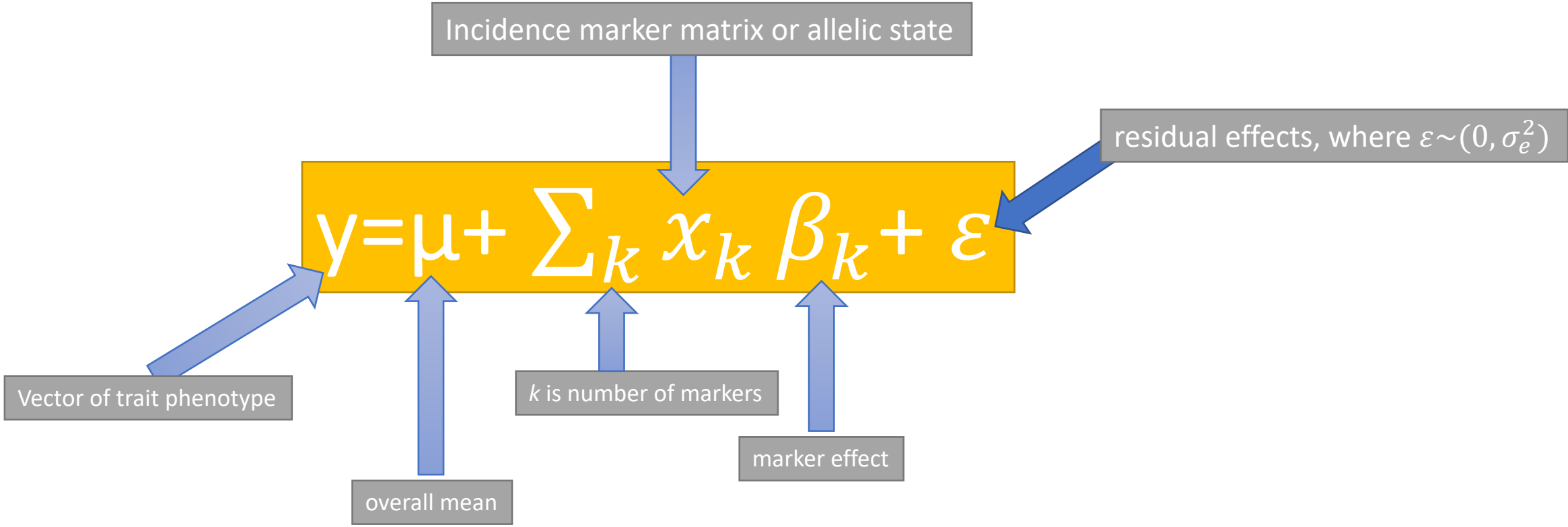
Factors effecting the prediction Accuracy

- Model selection/performances: varies based on marker assumption and effects.
- Gene effects: Additive and non-additive contributions.
- Heritability: highly heritable traits have high prediction accuracy.
- Linkage disequilibrium: reduced prediction accuracy if LD decays in advanced generations.

Desta and Ortiz, 2014

Prediction models

Basic standard model



Ordinary least squares $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ p>>>n

$$y=\mu+\sum_k x_k \beta_k+ \varepsilon$$

estimate β through OLS

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

If the effects of markers are estimated simultaneously OLS is not valid

Problem is number of markers (p) is >>>>number of individuals (p)

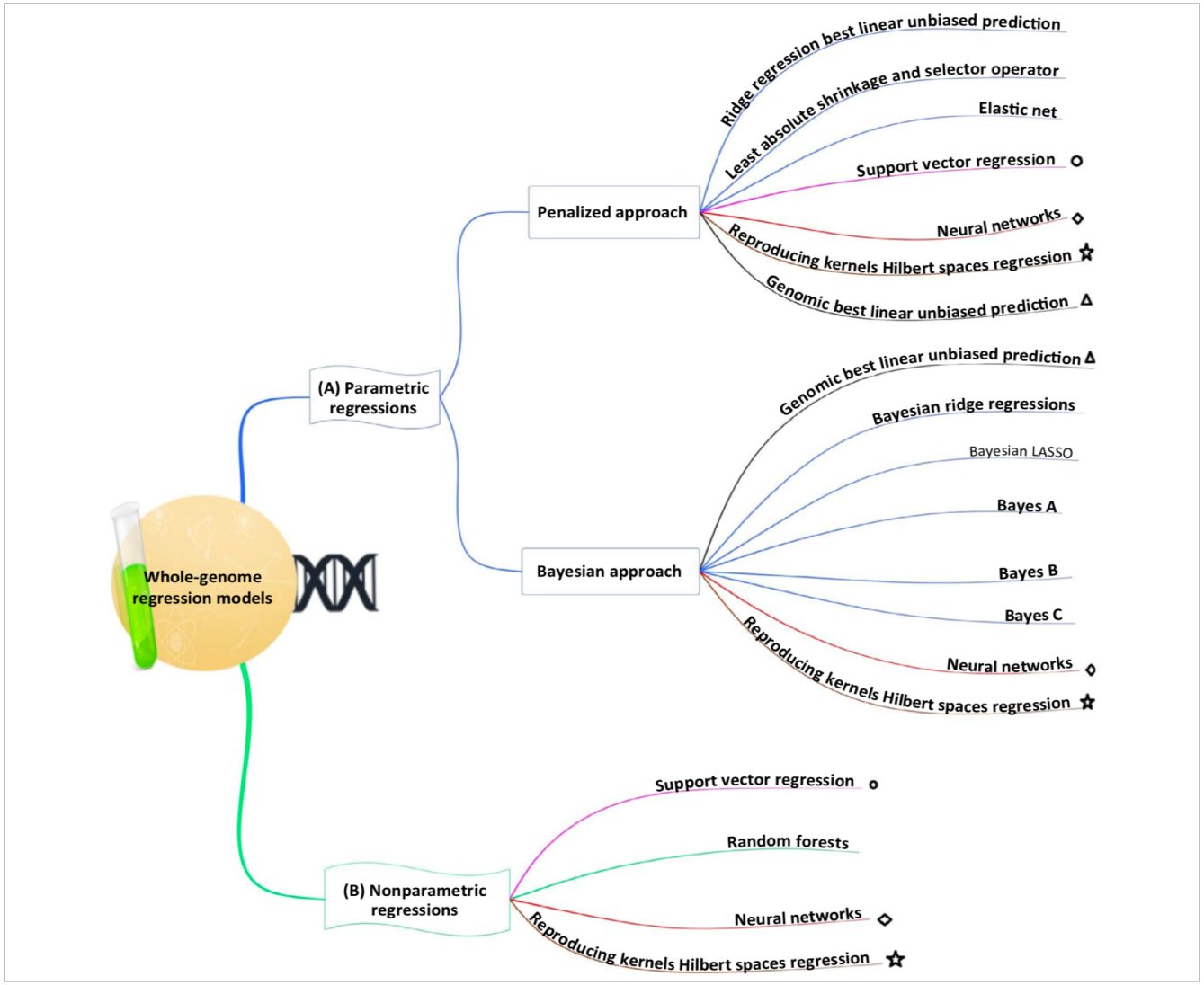
- X^TX is singular, determinant is 0, not invertible
- Multicollinearity: produces singular matrix again
- Predictors are highly correlated
- Insufficient degrees of freedom to estimate all marker effects at the same time

OLS estimates are not valid in that case.

To address this issues various models has been proposed

- Constrain possible effects
- fit the markers as random effects

Overview of Prediction Models



Adapted from manuscript Desta and Ortiz, 2014

Ridge-Regression BLUP (rr-BLUP)

$$y=\mu+\sum_k x_k \beta_k+ \varepsilon$$

estimate β by adding positive constraint

$$\boldsymbol{\beta_{ridge} = (X^T X)^{-1} + \lambda I)\, X^T y}$$

$$\beta_k \sim N(0, \sigma_e^2) \qquad \lambda = \frac{\sigma_e^2}{\sigma_\beta^2}$$

Ridge regression induces homogeneous shrinkage and it depends upon allele frequency
Assumes all markers have same variance with small but non-zero effect.

- We add constant to diagonal, thus it is invertible
- Degree of shrinkage depends upon $\boldsymbol{\lambda}$, larger the $\boldsymbol{\lambda}$ larger is the shrinkage

$$\frac{\boldsymbol{\beta_{OLS}}}{\boldsymbol{1+\lambda}}$$

- We basically shrink OLS estimates towards 0
- $\lambda \mathbf{I}$ term reduces collinearity and prevents the matrix $\mathbf{X}^T\mathbf{X}$ from becoming singular.

Cross validation

evaluate prediction performance

- take model uncertainty into account
- divide data into training and testing sets
- train the model in the training set
- evaluate predictive performance in the testing set
- predictive correlation: $r = cor(y, y_{predicted})$

K-fold Cross-validation

Cross-validation for rrBLUP

- Divide the data into training and testing set

Training set $\in (y_{training}, X_{training})$

Testing set $\in (y_{testing}, X_{testing})$

$$y_{training} = X_{training}\beta_{training} + \varepsilon_{training}$$

- Perform cross-validation

$$\acute{y}_{testing} = X_{testing}\beta_{testing}$$

$$cor\,(y_{testing}, \acute{y}_{testing}) = cor\,(y_{testing}, X_{testing}\beta_{testing})$$

Where,

$\acute{y}$= predicted value

X= n x m matrix

Dimension Reduction methods

Principle Component Regression

Singular or eigen value decomposition: Provides additional in site into nature of Ridge regression

$$y=\mu+\sum_k x_k \beta_k+ \varepsilon$$

replace y with eigenvectors and regress phenotypes directly on eigenvectors or principle components

$$x_{k=UDV^T}$$

$$\boldsymbol{\beta_{OLS} = V D^{-1} U^T y}$$

Reason is to avoid inverse
Suitable when predictors are correlated
And calculations are tedious

U= n x m orthogonal matrix
D= n x m diagonal matrix with singular values
V= n x n orthogonal matrix

- Ordinary least squares (OLS) when we know QTL or only one marker tag one QTL.
- What if we have dense marker system and account marker associations or LD between markers?

Compute Genomic relationship matrix (GRM)
Mixed model approach

Mixed model approach

$$y=\mu+\sum_k x_k \beta_k+Zu+\varepsilon$$

$$\tilde{\beta} = (X'V^{-1}X)^{-} X'V^{-1}y, \quad \text{with} \quad \text{BLUE}(X\beta) = X\tilde{\beta}$$

BLUE

$y \sim N(X\beta, Zu, R\sigma^2 e)$

V= $Zu, R\sigma^2 e$

Henderson mixed model equation

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'ZG^{-1}\frac{\sigma^2 e}{\sigma^2 u} \end{bmatrix} \begin{bmatrix} \tilde{\beta} \\ \tilde{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

β_k is fixed effect

u is random effect with $u \sim N(0, G\sigma^2 u)$

and $\varepsilon \sim N(0, R\sigma^2 e)$, typically R is identity matrix

BLUP

$$\tilde{u} = G\,Z'V^{-1}(y - X\beta) = \text{BLUP}(u)$$

Genomic BLUP (gBLUP)

$$y=\mu+ +Zu+ \varepsilon$$

GRM to account for mendelian sampling

$$\hat{u} = \left[\mathbf{I} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_u^2}\right]\mathbf{y}$$

Equivalence between rrBLUP and gBLUP

For gBLUP the $Var(y) = \mathbf{ZGZ}'\sigma_u^2 + \mathbf{I}\sigma_e^2$

For rrBLUP the $Var(y) = \mathbf{XX}'\sigma_\beta^2 + \mathbf{I}\sigma_e^2$

Construction of G matrix

$$y = Zu + \varepsilon$$

$$u \sim N(0, G\sigma^2 a)$$

$\sigma^2 a$ is additive variance = $\sum_{i=1}^{m} 2p_j\,(1-p_j$

3 steps

1. Create a centered X matrix
2. Create the cross product
3. Divide it by $\sum_{i=1}^{m} 2p_j(1\text{-}p_j)$

$$G = \frac{XX^T}{\sum_{i=1}^{m} 2p_j\,(1-p_j)}$$

First G matrix, VanRanden (2008)

Bayesian approach

Estimates variance components and solve the mixed model equation in single framework

$$P(\theta|y) = \frac{P(\theta, y)}{P(y)}$$

y= observed value

θ= parameter(unobserved value)

Bayesian models differ only with respect to prior, pick prior that provides best fit
Shrinkage inducing mechanism is included in the model mainly by specifying an appropriate prior density for regression coefficients

Bayesian approaches

Bayesian ridge regression
Induces homogeneous shrinkage of all marker effects towards zero and yields a Gaussian distribution of marker effects
Bayesian LASSO
Combines shrinkage and variable selection methods
Has an exponential prior on marker variances resulting in a double exponential (DE) distribution.
The density distribution has a higher mass density at zero and heavier prior tails compared with a Gaussian distribution
Bayes A
Utilizes an inverse chi-square (x^2) on marker variances yielding a scaled t-distribution for marker effects
it shrinks tiny marker effects towards zero and larger values survive
Has a higher peak of mass density zero compared with the DE distribution
BayesB
Fraction of markers with zero effect
BayesC
assumes t-distribution one with large variance for SNP fraction and other with small variance

More details on The Bayesian Alphabet can be found in Gianola, 2013, 10.1534/genetics.113.151753

https://doi.org/10.1016/j.tplants.2014.05.006

Prediction studies in Literature

Genomic Selection in Plant Breeding: A Comparison of Models

Add to Binder | View My Binders | View Comments

Printer-friendly PDF

This article in CS

Vol. 52 No. 1, p. 146-160

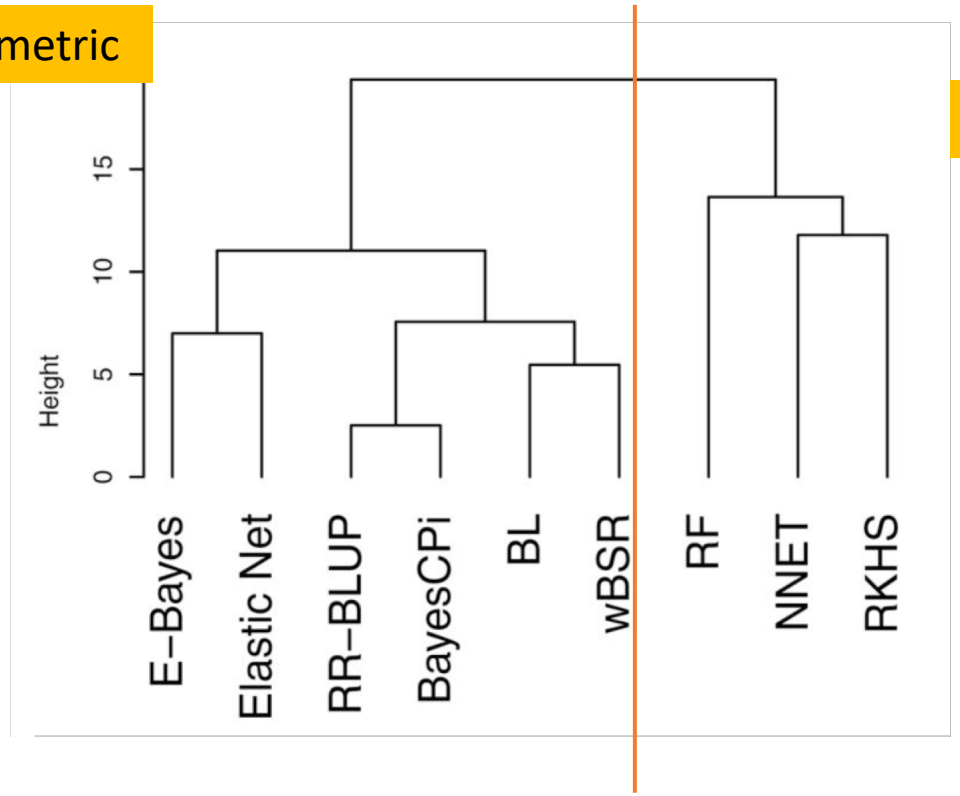
Received: June 1, 2011
Published: Jan, 2012

Nicolas Heslot[ac], Hsiao-Pei Yang[b], Mark E. Sorrells[a] and Jean-Luc Jannink [*b]

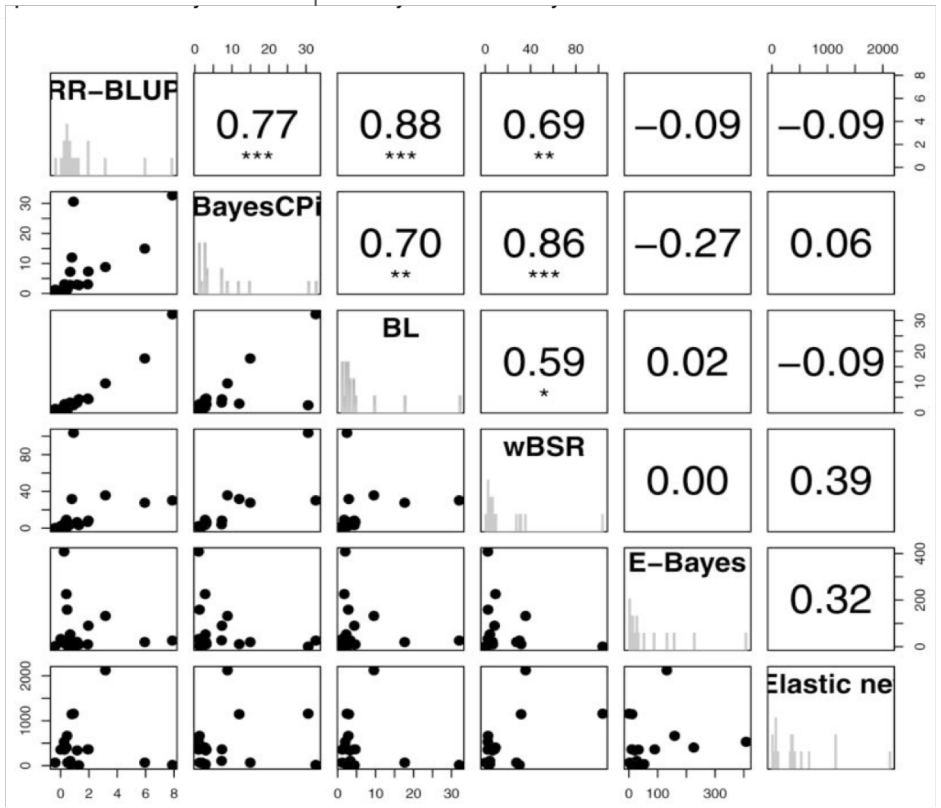
Author Affiliations

doi:10.2135/cropsci2011.06.0297

parametric



Non-parametric



Comparison and correlation between models

Molecular Breeding
October 2017, 37:117 | Cite as

Practical application of genomic selection in a doubled-haploid winter wheat breeding program

Authors Authors and affiliations

Jiayin Song, Brett F. Carver, Carol Powers, Liuling Yan, Jaroslav Klápště, Yousry A. El-K... Charles Chen

Imputation

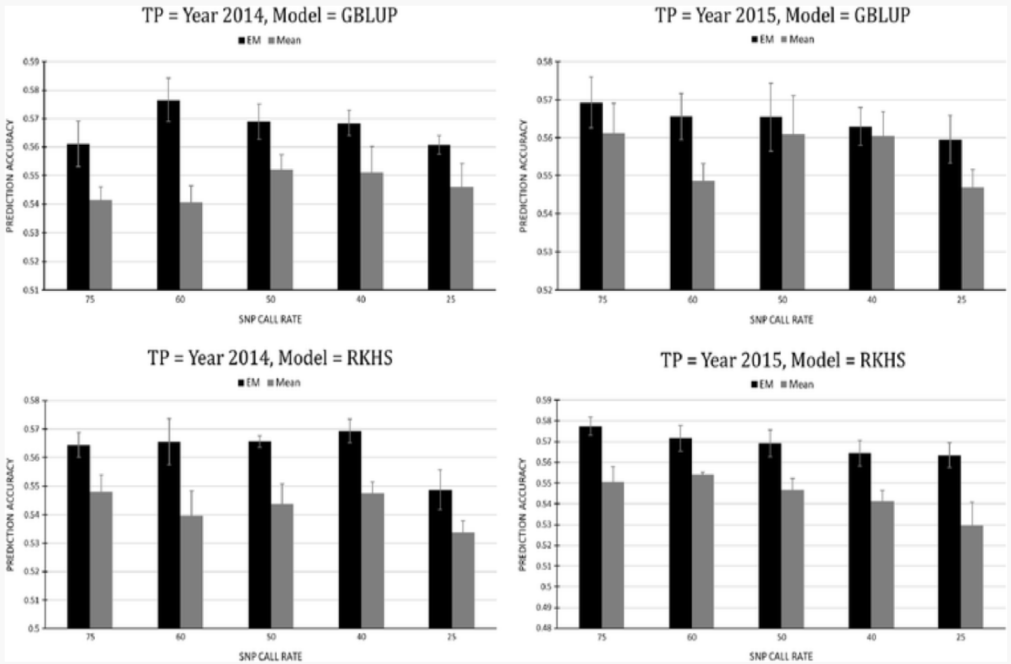


Fig. 1
Comparison of two missing data imputation methods, EM and mean, based on the predictive ability from the GBLUP (above) and RKHS (below) cross-validation models (with SNP effect only) across a gradient of SNP call rate; TP training population; bandwidth parameter was set to 0.1 for all RKHS models

Cross-validation

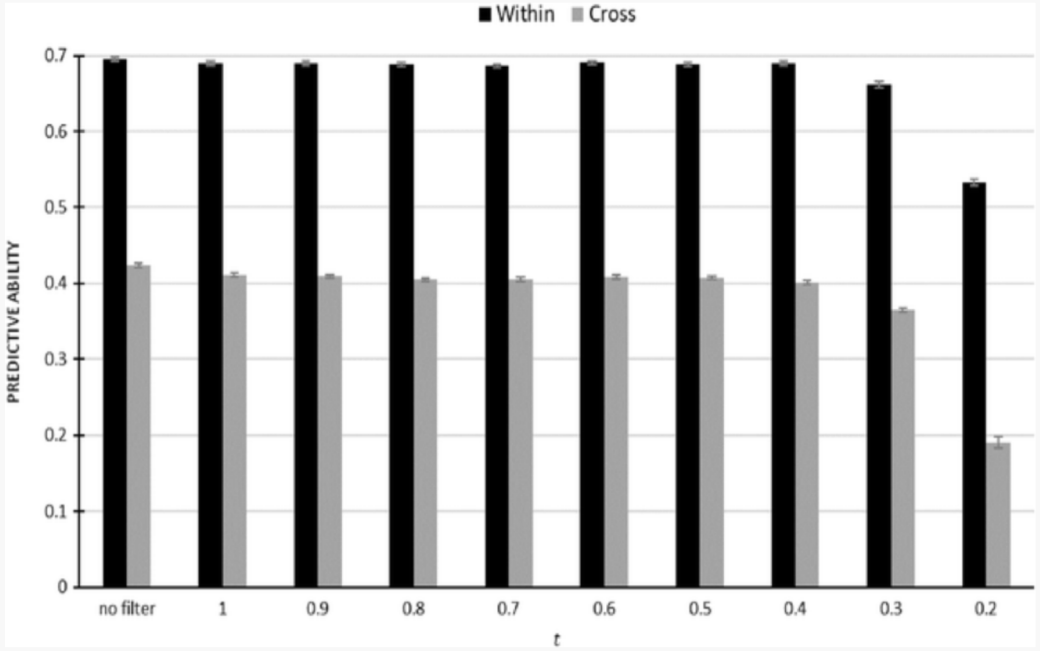


Fig. 2
Predictive ability from the best within-year cross-validation model (within: year 2015 RKHS model with the marker effect and both heading date and disease index as covariates) and the best cross-year prediction model (cross: year 2014 predicting 2015 RKHS model with the marker effect and heading date as covariate) across subsets of marker filtered by absolute pairwise correlation threshold (t)

G3 Genes | Genomes | Genetics

GENOMIC SELECTION

Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy

Adam Norman, Julian Taylor, James Edwards, and Haydn Kuchel
School of Agriculture, Food & Wine, University of Adelaide
ORCID ID: 0000-0002-0794-4907 (A.N.)

Cross-validation for k-mean clustering

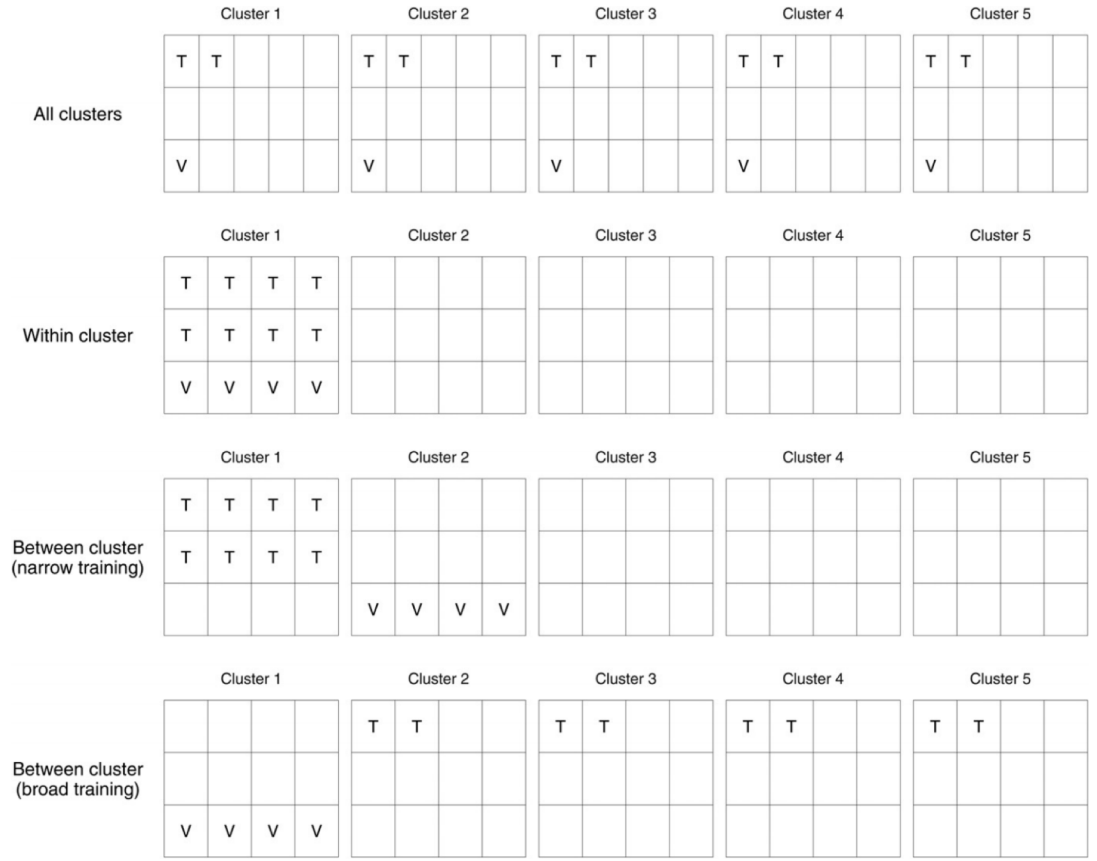


Figure 1 Description of the four cross-validation designs used to assess the impact of underlying population structure. The partitions within each cluster were formed by randomly sampling without replacement. Replication was achieved by rotating partitions within each design to provide all combinations of partitions and clusters. All designs had consistent training and validation set sizes of 1,000 and 500 respectively.

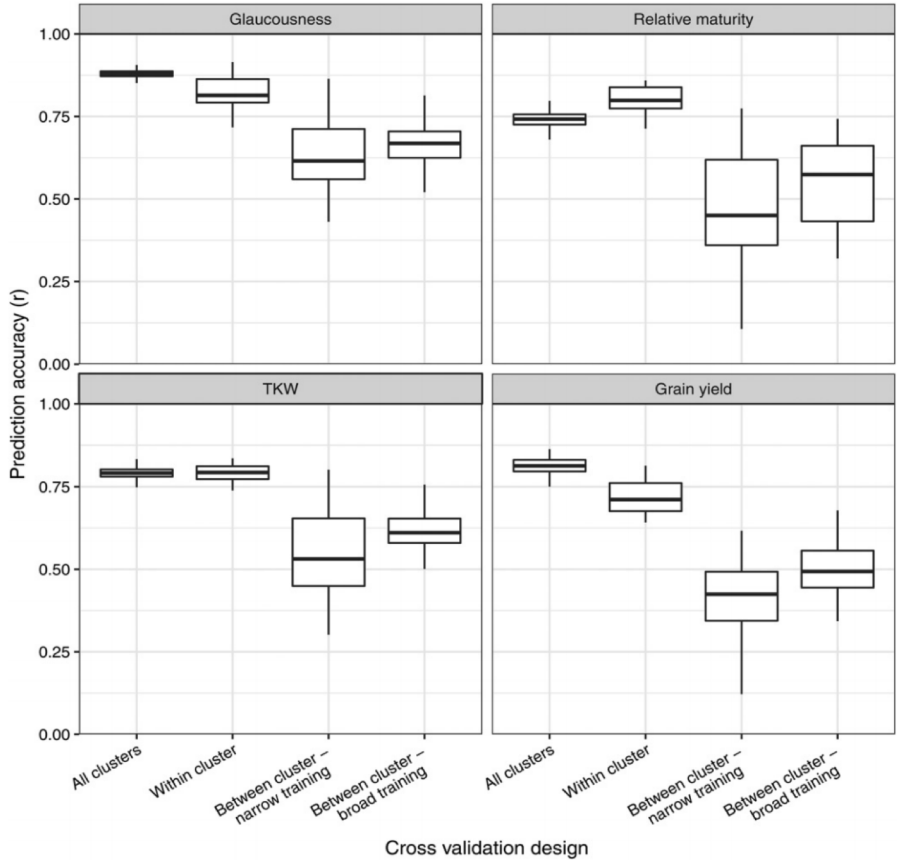


Figure 4 Boxplots showing prediction accuracies from the K-means clustering method for each category of training and validation set combinations, detailed in section 2.5.1. Prediction accuracy was calculated by correlating predictions of the validation set to the corresponding additive GBLUP values from the full model with all lines included. TKW represents thousand kernel weight.

G3 Genes | Genomes | Genetics

GENOMIC SELECTION

Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy

Adam Norman, Julian Taylor, James Edwards, and Haydn Kuchel
School of Agriculture, Food & Wine, University of Adelaide
ORCID ID: 0000-0002-0794-4907 (A.N.)

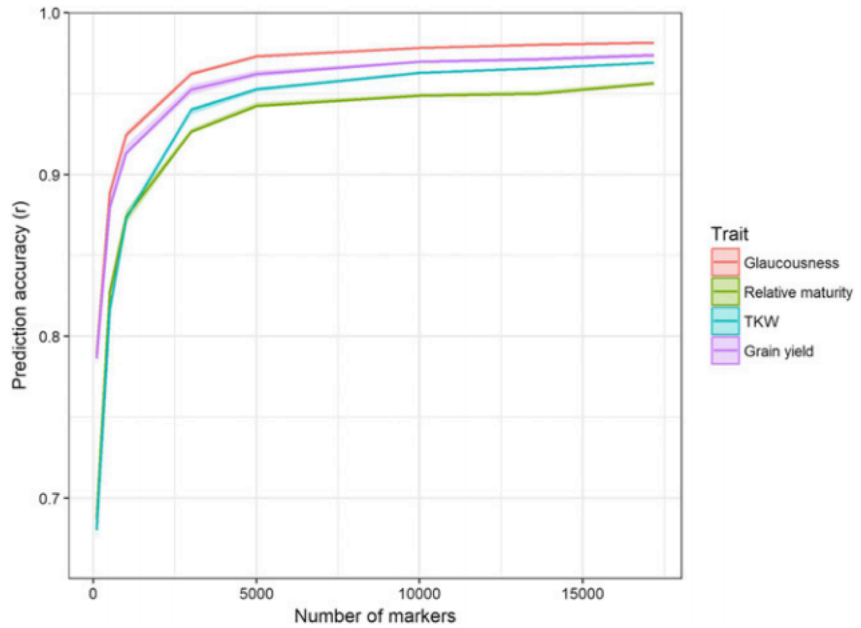


Figure 6 Plot showing the effect of marker density on prediction accuracy for each trait. Prediction accuracy was assessed by performing random five-fold cross-validation for each selection of markers, and correlating predictions of the validation set to the corresponding additive GBLUP values from the full model with all lines included. Marker subsets were selected to be evenly distributed over the genome and to have high minor allele frequency.

Demonstration in R