Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multivariate Genomic Best Linear Unbiased Prediction (GBLUP) Can Improve Genomic Prediction in Mungbean

Mark Edward Fabreag
The University of Queensland
Shanice Van Haeften
The University of Queensland
Cassandra Pegg
Commonwealth Scientific and Industrial Research Organisation
Eric Dinglasan
The University of Queensland
Millicent Smith
The University of Queensland
Valeria Paccapelo
Queensland Department of Primary Industries
Thomas Noble
Queensland Department of Primary Industries
Merrill Ryan
Queensland Department of Primary Industries
Lee Hickey
The University of Queensland
Ben J. Hayes
b.hayes@uq.edu.au

The University of Queensland

Research Article

Keywords: genomic selection, genomic best linear unbiased prediction, multi-trait, multi-environment, Vigna radiata

Posted Date: April 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9153120/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Multivariate Genomic Best Linear Unbiased Prediction (GBLUP) Can Improve Genomic Prediction in Mungbean

Mark Edward Fabreag[a], Shanice Van Haeften[a], Cassandra Pegg[b], Eric Dinglasan[a], Millicent Smith[a,c], Valeria Paccapelo[d], Thomas Noble[d], Merrill Ryan[d], Lee Hickey[a], Ben J. Hayes[a]

[a] Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St. Lucia, QLD, Australia

[b] Agriculture and Food, CSIRO, St. Lucia, QLD, Australia

[c] School of Agriculture & Food Sustainability, The University of Queensland, St. Lucia, QLD, Australia

[d] Department of Primary Industries, Warwick, QLD, Australia

Corresponding Author: Ben Hayes, b.hayes@uq.edu.au

Authors ORCID Number:

Mark Edward Fabreag: 0009-0004-9013-3436

Shanice Van Haeften: 0000-0003-0412-3457

Cassandra Pegg: 0000-0002-6080-7047

Eric Dinglasan: 0000-0002-4824-5977

Millicent Smith: 0000-0003-1749-2884

Valeria Paccapelo:

Thomas Noble: 0000-0002-7731-5559

Merrill Ryan:

Lee Hickey: 0000-0001-6909-7101

Ben Hayes: 0000-0002-5606-3970

Keywords: genomic selection, genomic best linear unbiased prediction, multi-trait, multi-environment, *Vigna radiata*

Abstract

Genomic prediction (GP) remains underexplored in mungbean despite its potential to accelerate genetic gain in breeding programs. This study evaluated the implementation of GP in mungbean using multivariate genomic best linear unbiased prediction (GBLUP), including multi-trait, multi-environment and multi-trait multi-environment models to maximize prediction accuracy. Utilizing yield (YLD), days-to-50%-flowering (DTF), and plant height (PHT) data collected from 313 diverse mungbean accessions (including accessions released and widely grown in Australia) tested across four environments and genotyped for 3,310 SNPs, we compared multivariate GBLUP models including single-trait−multi-environment, multi-trait−single-environment, and multi-trait−multi-environment against univariate single-trait−single-environment models. For multi-trait models, YLD was the primary trait, while DTF and PHT were the secondary traits. Secondary traits are correlated traits which are easy and relatively inexpensive to measure early in crop development. Cross-validation schemes implemented assessed whether secondary traits were available at the time of prediction of YLD for the validation sets (no secondary traits, NOSEC vs secondary traits available, SEC) for multi-trait models, cross-environment data availability (no cross-environment data, Scenario 1 vs cross-environment data available, Scenario 2) for multi-environment models and their factorial combinations (NOSEC/SEC × Scenario1/2) for multi-trait multi-environment models could improve prediction accuracy. Single-trait single-environment GBLUP models provided baseline prediction accuracies of 0.33±0.02 to 0.61±0.01 for YLD, 0.39±0.02 to 0.49±0.02 for DTF, and 0.38±0.01 to 0.59±0.01 for PHT. Multivariate approaches substantially improved prediction accuracy, with gains up to +52.7% for YLD, +100.22% for DTF and +83.43% for

PHT, over baseline models. The magnitude of improvement depended on trait heritability, genetic correlation between traits and between environments, and availability of information on the validation set. These findings demonstrate that genomic prediction is feasible in mungbean and can be substantially enhanced through multivariate approaches, providing a foundation for implementing genomic selection in mungbean breeding programs.

Acknowledgments

MEF would like to thank CSIRO AI for Missions for the PhD Scholarship grant. This research was supported by the International Mungbean Improvement Network (ACIAR/CIM2014/079) and we gratefully acknowledge all partner countries for their contributions to the IMIN network. We also thank Prof Cheng-Ruei Lee's research group at National Taiwan University for providing access to the Crystal reference genome. MEF was supported as an Adjunct PhD student of the Australian Research Council Training Centre in Predictive Breeding for Agricultural Futures (IC230100016) and as an associated PhD student of the International Research Training Group on Accelerating Crop Genetic Gain.

Authors Contribution

MEF contributed to conceptualization of the paper, conducted the formal analysis and wrote the original draft. **SVH** contributed to phenotype data generation and preliminary analysis. **VP**, **TN** and **MR** provided the genotype and phenotype data and reviewed the manuscript. **CP** supervised the research project and reviewed the manuscript. **ED**, **MS** and **LH** conceptualized, supervised the research project and critically reviewed the manuscript. **BH** conceptualized, provided primary supervision on the project, and critically reviewed the manuscript. All authors read and approved the final manuscript.

Key Message

Genomic prediction remains underexplored in mungbean. However, a multi-trait approach incorporating correlated traits and multi-environment data improved prediction accuracy, demonstrating its potential for accelerating genetic gain in mungbean breeding.

Introduction

Mungbean is a socio-economically important grain legume crop that serves as a significant source of high-quality protein, carbohydrates, and iron for many communities (Dahiya et al., 2015). It is commonly processed and consumed as dahl, bean sprout, noodles and paste across tropical and subtropical regions (Nair & Schreinemachers, 2020). It is worth noting that during the green revolution era, cereal crops were the focus of agricultural research and production, marginalizing many other crops including mungbean (Foyer et al., 2016). Nevertheless, active efforts to improve mungbean agronomy and genetics began about five decades ago, aiming to diversify and supplement cereal-dominated diets (Shanmugasundaram et al., 2009). Since then, mungbean production, demand, and value have continued to increase (Nair & Schreinemachers, 2020; Nair et al., 2012; Shanmugasundaram et al., 2009). Despite genetic and agronomic improvements made for mungbean, it is still a relatively low yielding crop with an average of 0.75 metric tonnes per ha (mt/ha) globally (Nair & Schreinemachers, 2020) and ~0.9 mt/ha in Australia (Australian Mungbean Association, 2022). This is in contrast with soybean, a major legume crop, where average yield globally is 2.79 mt/ha (Langemeier & Zhou, 2022), and 1.78 mt/ha (average across 2019-2023) in Australia (Granwal, 2024).

While the application of advanced breeding technologies is a key driver of high rates of genetic gain in major crops, the application of these technologies has been limited in mungbean. For example, a quality reference genome for mungbean has only recently become available (Ha et al., 2021; Kang et al., 2014; Liu et al., 2022). These reference genomes have enabled the development of more abundant and higher quality SNP (single nucleotide polymorphism) marker resources. Nonetheless, in mungbean, the use of SNPs has been mostly limited to QTL (quantitative trait loci) mapping (Ha et al., 2021; Liu et al., 2022; Noble et al., 2018). These

104 significant marker-trait associations have limited utility for highly quantitative traits which are
105 controlled by a large number of QTLs with small effects, like yield (Bernardo, 2008).

106

107 A more efficient use of molecular markers when breeding for highly quantitative traits (e.g.
108 yield) is through genomic selection (GS) (Meuwissen et al., 2001). In GS, all markers across
109 the genome are used, and all marker effects are simultaneously estimated for genomic
110 predictions. The underlying assumption for GS is that each of the many QTL affecting the
111 complex trait will be in linkage disequilibrium (LD) with at least one marker. As a result, a
112 higher proportion of genetic variance is captured. GS has been shown to accelerate genetic gain
113 in major crops, especially for highly quantitative and complex traits (Bernardo & Yu, 2007;
114 Beyene et al., 2019; Beyene et al., 2015; Crossa et al., 2017; Massman et al., 2013; Sweeney
115 et al., 2021).

116

117 Effective implementation of genomic selection relies critically on the accuracy of the genomic
118 prediction models, alongside other contributing factors. Multivariate approaches such as multi-
119 trait genomic prediction models, which simultaneously use information from multiple traits,
120 have been shown to have improved prediction accuracies as compared to univariate
121 counterparts in some situations (Calus & Veerkamp, 2011; Jia & Jannink, 2012; Neyhart et al.,
122 2019). For example, genetic improvement of the primary trait of interest can be challenging
123 when it is of low heritability, difficult or expensive to phenotype. Multi-trait prediction can be
124 especially beneficial for such traits, when it incorporates secondary traits. Secondary traits are
125 traits that are correlated to the primary trait, are highly heritable, can be easier and cheaper to
126 phenotype, and can be observed earlier. Rutkoski et al. (2016) showed for example that yield
127 prediction in wheat can be enhanced by incorporating high-throughput data (such as canopy

temperature, red and green normalized difference vegetation index) as secondary traits in a multi-trait genomic prediction framework. Likewise, Watson et al. (2019) utilized glasshouse traits (including plant height) observed under speed breeding conditions in a multi-trait GBLUP to improve prediction of yield in spring wheat. The benefit of the multi-trait approach has also been shown in breeding for end-use quality traits in wheat (Hayes et al., 2017). Phenotyping end-use quality traits are costly and often necessitate large quantities of flour, which are only available in the later stages of testing when field plots are larger. To circumvent these problems, Hayes et al., (2017) used near infrared and nuclear magnetic resonance methods to predict for end-use quality traits. These methods are cheaper to implement and require smaller amounts of flour, enabling selection at an earlier stage of the breeding cycle.

Multi-environment trial (MET) data can also be analyzed using multi-trait genomic prediction frameworks, where performance in each environment is treated as a distinct trait. This is equivalent to the multivariate genotype-by-environment interaction model with unstructured genetic variance-covariance, which allows for a more realistic assumption that the genetic variance across environments, and genetic correlations between environments are heterogeneous. Such multi-environment genomic prediction models has been shown to enhance prediction of the performance of breeding lines within and across target environments (Burgueño et al., 2012; Hunt et al., 2020; Montesinos-López et al., 2016; Montesinos-López et al., 2019).

In MET, multiple traits are commonly measured. This allows implementation of multi-trait multi-environment (multi-trait multi-environment) genomic prediction frameworks which leverage genetic correlations across three dimensions: (1) between different traits within the

same environment, (2) within the same trait across different environments, and (3) between different traits across different environments. Multi-trait multi-environment genomic prediction models have been shown to enhance prediction accuracies substantially (Covarrubias-Pazaran et al., 2018; Gill et al., 2021; Saludares et al., 2024). When fitting unstructured genetic variance-covariance matrices, however, the number of parameters to be estimated increases quadratically with the number of traits and/or environments (*i.e.* $p(p-1)/2$), creating computational challenges for multi-trait, multi-environment and multi-trait multi-environment models. To address this, Smith, Cullis and Thompson (2001) proposed a factor analytic approach which involves fitting k latent variable factors to model variance-covariance, resulting in a parsimonious model with $p(k+1)-k(k-1)/2$ parameters to be estimated. The linear mixed model with factor analytic variance structures has been increasingly applied to model and study genotype-by-environment interactions (Cowling et al., 2023; Hunt et al., 2020; Smith, Cullis, & Gilmour, 2001; Smith & Cullis, 2018; Tolhurst et al., 2022).

To date, only a single report has been published suggesting that genomic selection can be used in mungbean (Fumia et al., 2023). These authors applied genomic best linear unbiased prediction (GBLUP) on mungbean Mini-Core, a collection of diverse mungbean accessions (Schafleitner et al., 2015), to predict 100-seed weight, seeds per pod, pods weight and total seed weight. The authors fitted single-trait single-environment, and single-trait multi-environment genomic prediction models. However, the study reported only the prediction accuracies of single-trait single-environment GBLUP models, which ranged from ~0.6 to ~0.98. Moreover, their single-trait multi-environment analyses modelled the genetic variance-covariance across environments using a compound symmetry structure with heterogeneous genetic variances, rather than more flexible unstructured or factor analytic structure.

Our objective in this study was to assess whether multivariate genomic prediction approaches such as multi-trait, multi-environment, and multi-trait multi-environment models, could enhance prediction accuracy for mungbean beyond baseline models. The baseline performance of genomic prediction in mungbean was evaluated using single-trait single-environment GBLUP models across diverse germplasm and multiple environments. For multi-trait models, we focused on using secondary traits, which are correlated traits, highly heritable, easier and cheaper to phenotype, and can be observed earlier in the growing season, to improve yield predictions.

Materials and Methods

Plant Materials

In this study, we used a total of 335 mungbean accessions, which included of 293 accessions from the Mini-Core Panel, 13 commercial varieties released in Australia, and additional accessions and breeding lines from various countries. The mungbean Mini-Core Panel was collected from 19 countries across Asia, Americas, Australia, and Africa, and represents the larger set of diverse mungbean germplasm collection conserved in the World Vegetable Center gene-bank (Schafleitner et al., 2015). The entire set will be referred to hereafter as Mini-Core Plus Panel. Note that only 313 accessions with good quality genotype data were used for genetic diversity analysis and genomic predictions.

Genotype Data

The SNP data of Mini-Core Plus Panel was generated using DArT (Diversity Arrays Technology) SNP markers. Genotyping was done following an integrated DArT and genotype-by-sequencing pipeline. The SNPs were identified by aligning the sequence data generated to the Crystal-based mungbean reference genome (https://avrdc.org/dna-sequence-data-download/). A total of 3,310 SNP markers were retained after quality checking with parameters of 1) minimum allele frequency of 5%, 2) heterozygous ≤10%, and 3) marker missing data ≤20%. Lines with missing data of >20% were also removed, resulting to a final 313 accessions.

Genetic Diversity and Population Structure Analysis

The population structure of the Mini-Core Plus panel was analyzed using the Discriminant Analysis of Principal Component (DAPC) approach (Jombart et al., 2010), implemented in the R package $adegenet$ (Jombart, 2008). The optimal number of clusters was determined using the "elbow" method. The ideal number of principal components (PCs) for discriminant analysis was determined through cross-validation approach using the $xval$ function of $adegenet$. The genetic differentiation between the resulting clusters were also measured using Weir and Cockerham (1984) fixation index (F_{ST}). F_{ST} was calculated using the $hierfstat$ package (Goudet, 2005) in R. The genetic diversity within cluster was also assessed by calculating the expected heterozygosity for each cluster as $H_e = \frac{2\sum_i^m p_i(1-p_i)}{m}$, where p_i is the minor allele frequency at locus i.

The pairwise SNP intrachromosomal linkage disequilibrium (LD) (r^2) across the entire genome was calculated for the entire Mini-Core Plus Panel. The mean LD between SNP markers were

then calculated per window of 100kb pairwise SNP distance. LD decay was assessed using three thresholds: 1) $LD_{0.5max}$ was the physical distance at which r^2 is equivalent to half of the maximum r^2 value observed, 2) LD_{20} which was the physical distance at which $r^2 = 0.20$ on average and 3) LD_{10} was the physical distance at which $r^2 = 0.10$ on average. LD within clusters were also calculated and evaluated in the same manner.

Multi-Environment Trial Data

The Mini-Core Plus Panel was yield tested in four field trials conducted across three years in Queensland, Australia: at Emerald in 2017 (EME17) and 2018 (EME18), at Hermitage in 2018 (HER18), and at Gatton (GAT23) in 2023. These locations represent areas where mungbean is widely grown in Australia. Moreover, the Emerald and Hermitage locations are Australian mungbean breeding program testing sites. All the field trials followed a randomized complete block design with two replicates. Both EME17 and EME18 were rainfed, while HER18 and GAT23 were irrigated. In addition, the average minimum and maximum temperature throughout the growing duration at both EME17 (21.1^0C, 33.8^0C) and EME18 (21.3^0C, 33.5^0C) were higher than that of HER18 (16.7^0C, 30.2^0C) and GAT23 (18.1, 32.2) trials (Table 1). For HER18 and EME18, the Mini-Core Plus was divided into three sets based on maturity i.e. early, mid, and late, to facilitate agronomic operations. Despite this division, the maturity ranges of the three sets overlapped. In both EME17 and GAT23 trials, the entire Mini-Core Plus were tested as one set. EME17, EME18 and HER18 were manually harvested while GAT23 was machine harvested. Grain yield (YLD), days-to-50%-flowering (DTF) and plant height (PHT) data were collected from all the trials. DTF was captured when 50% of the plot had flowered. Plant height was recorded at maximum vegetative stage. The average seed length, average seed width, thousand seed weight, and seed coat luster were collected only from GAT23. The

number of unique accessions per trial and number of common accessions between trials is summarized in Table 2. Each field trial was considered a unique environment.

246 **Table 1.** Summary of field trial information

Trial Code	Location	Year	Latitude/ Longitude	Set	Date of Planting	In-crop average Minimum/Max Temperature (^{0}C)	Rainfall (mm)	Water Management
EME17	Emerald	2017	-23.53, 148.18	All	25-Jan-17	21.1/33.8	278	Rainfed
EME18	Emerald	2018	-23.53, 148.18	Early	30-Jan-18	21.3/33.5	242	Rainfed
EME18	Emerald	2018	-23.53, 148.18	Mid	30-Jan-18	21.3/33.5	242	Rainfed
EME18	Emerald	2018	-23.53, 148.18	Late	29-Jan-18	21.2/33.4	242	Rainfed
HER18	Hermitage	2018	-28.21, 152.10	Early	11-Jan-18	16.6/30.1	285	Irrigated
HER18	Hermitage	2018	-28.21, 152.10	Mid	11-Jan-18	16.5/30	285	Irrigated
HER18	Hermitage	2018	-28.21, 152.10	Late	20-Dec-17	16.9/30.7	341	Irrigated
GAT23	Gatton	2023	-27.56, 152.34	All	22-Dec-22	18.1/32.2	115	Irrigated

247

Table 2. Number of mungbean accessions evaluated in each environment and accessions shared across environments[a]

	EME17	EME18			HER18			GAT23
		(Early)	(Medium)	(Late)	(Early)	(Medium)	(Late)	
EME17	306/285							
EME18 Early	75/68	77/70						
EME18 Medium	93/87	3/3	107/100					
EME18 Late	137/133	3/3	3/3	142/138				
HER18 Early	75/68	77/70	3/3	3/3	77/70			
HER18 Medium	93/87	3/3	107/100	3/3	3/3	107/100		
HER18 Late	138/134	3/3	3/3	142/138	3/3	3/3	143/139	
GAT23	273/270	68/67	90/90	135/133	68/67	90/90	136/134	300/297

[a] The diagonal contains the total number of accessions within an environment, while the off diagonal refers to the number of accessions common between environments

[b] Numbers are reported as total number of accessions tested/total number of accessions tested with genomic data

Phenotype Data Analysis

Best linear unbiased estimates (BLUEs) for YLD, DTF and PHT were calculated per trial following the linear mixed model:

$$\mathbf{y} = \mathbf{X_g g} + \mathbf{X_b b} + \mathbf{Z_r u_r} + \mathbf{Z_c u_c} + \mathbf{e}, \text{ (Eq. 1)}$$

where $\mathbf{y}$ is the vector of observations of length n; $\mathbf{g}$ is the vector of line fixed effects of length n_g, with $\mathbf{X_g}$ as $n \times n_g$ design matrix; $\mathbf{b}$ is the vector of experimental design-associated fixed effects (i.e. replication and/or maturity set) of length n_b, with $\mathbf{X_b}$ as $n \times n_b$ design matrix; $\mathbf{u_r}$ is the vector of row random effects of length n_{u_r}, with $\mathbf{Z_r}$ as $n \times n_{u_r}$ design matrix with $\mathbf{u_r} \sim \mathrm{N}\left(0, \sigma^2_{\mathrm{u_r}} \mathbf{I}_{\mathrm{u_r}}\right)$; $\mathbf{u_c}$ is the vector of column random effects of length n_{u_c}, with $\mathbf{Z_c}$ as $n \times n_{u_c}$ design matrix with $\mathbf{u_c} \sim \mathrm{N}\left(0, \sigma^2_{\mathrm{u_c}} \mathbf{I}_{\mathrm{u_c}}\right)$; $\mathbf{e}$ is the vector of residuals of length n with $\mathbf{e} \sim \mathrm{N}(0, \mathbf{R})$. To further account for the spatial trend and noise, $\mathbf{R}$ was modelled to have an order one autoregressive covariance structure and independent and identically distributed random error, such that $\mathbf{R} = \sigma^2_{\mathrm{s}} \boldsymbol{\Sigma}_{\mathrm{r}}(\rho_{\mathrm{r}}) \otimes \boldsymbol{\Sigma}_{\mathrm{c}}(\rho_{\mathrm{c}}) + \sigma^2_{\mathrm{e}} \mathbf{I}_{\mathrm{n}}$, where σ^2_{s} is the autoregressive scaling component, $\boldsymbol{\Sigma}_{\mathbf{r}}$ is the $n_r \times n_r$ matrix with row autocorrelation ρ_{r}, $\boldsymbol{\Sigma}_{\mathbf{c}}$ is the $n_c \times n_c$ matrix with column autocorrelation ρ_{c}, and σ^2_{e} independent and identically distributed random error variance. Datapoints with extreme residuals based on the fitted model were deemed outliers. These were removed and the model was re-fitted as necessary.

Weights for the BLUEs were derived from Eq. 1 as:

$$\mathbf{w} = \mathrm{diag}(\boldsymbol{\Omega}^{-1})$$

$$\boldsymbol{\Omega} = \mathrm{var}\left(\begin{bmatrix}\mathbf{BLUE_g} \\ \mathbf{BLUE_b}\end{bmatrix}\right) = \left(\begin{bmatrix}\mathbf{X_g}\ \mathbf{X_b}\end{bmatrix}' \mathbf{V}^{-1} \begin{bmatrix}\mathbf{X_g}\ \mathbf{X_b}\end{bmatrix}\right)^{-1}$$

$$\mathbf{V} = \mathrm{var}(\mathbf{y}) = \mathbf{Z}_{\mathrm{u_r}} \mathbf{I}_{\mathrm{u_r}} \sigma^2_{\mathrm{u_r}} \mathbf{Z}'_{\mathrm{u_r}} + \mathbf{Z}_{\mathrm{u_c}} \mathbf{I}_{\mathrm{u_c}} \sigma^2_{\mathrm{u_c}} \mathbf{Z}'_{\mathrm{u_c}} + \mathbf{R}$$

where $\mathbf{w}$ is the vector of w_{ij} weights of i^{th} fixed effect at j^{th} trial; $\mathbf{BLUE_g}$ and $\mathbf{BLUE_b}$ are the vectors of the estimates of $\mathbf{g}$ and $\mathbf{b}$, respectively. These weights were used in fitting the succeeding genomic best linear unbiased prediction (GBLUP) models.

Genomic Prediction Models

The univariate model, that is, single trait in a single environment, hereinafter referred to as single-trait single-environment, was fitted following the linear mixed model:

$$\mathbf{BLUE_g} = \mathbf{1_n}\mu + \mathbf{Zu} + \mathbf{e}$$

where $\mathbf{BLUE_g}$ is the vector of the fixed effect estimates of accessions of length n_g, derived from Eq. 1; $\mathbf{1_n}$ is a vector of 1s of length n_g; μ is the overall mean; $\mathbf{u}$ is the vector of genomic estimated breeding values (GEBVs) of length n_g, with $\mathbf{Z}$ as $n_g \times n_g$ design matrix; $\mathbf{u} \sim N(0, \sigma_u^2 \otimes \mathbf{A})$; $\mathbf{A}$ is the SNP-based relationship matrix; $\mathbf{e}$ is the vector of residuals of length n_g with $\mathbf{e} \sim N(0, \mathbf{R})$. $\mathbf{R}$ is assumed to be known from the first stage and given as $\mathbf{R} = \mathbf{W}$, where $\mathbf{W}$ is a diagonal matrix with elements $\mathbf{w}$, the weights described above (Smith, Cullis, & Gilmour, 2001; Smith et al., 2015).

A multivariate model for a single trait across multiple environments (single-trait multi-environment) was modelled as follows:

$$\begin{bmatrix} \mathbf{BLUE_{g_1}} \\ \vdots \\ \mathbf{BLUE_{g_l}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_{n_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{1}_{n_l} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_l \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{Z}_l \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_l \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_l \end{bmatrix} \quad \text{(Eq. 2)}$$

where, for each of l environments, $\mathbf{BLUE_g}$ is the vector of accession fixed effects estimate, $\mathbf{1_n}$ is a vector of 1s, μ is the mean estimate, $\mathbf{Z}$ is the design matrix, $\mathbf{u}$ is the vector of GEBVs and

295 $\mathbf{e}$ is the vector of residuals. $[\mathbf{u}_1 \cdots \mathbf{u}_l]' \sim \mathrm{N}([0_1 \cdots 0_l]', \mathbf{L} \otimes \mathbf{A})$, where $\mathbf{A}$ is the SNP-based
296 relationship matrix and $\mathbf{L}$ is the unstructured variance-covariance matrix of breeding values

297 $$\mathbf{L} = \begin{bmatrix} \sigma^2_{\mathrm{u}_1} & \cdots & \sigma_{\mathrm{u}_{1,l}} \\ \vdots & \ddots & \vdots \\ \sigma_{\mathrm{u}_{l,1}} & \cdots & \sigma^2_{\mathrm{u}_l} \end{bmatrix}$$

298 with σ^2_{u} and σ_{u} is the variance for each environment and covariance between environment
299 pairs, respectively. $[\mathbf{e}_1 \cdots \mathbf{e}_l]' \sim \mathrm{N}([0_1 \cdots 0_l]', \mathbf{R})$, where $\mathbf{R}$ is as described above, with $\mathbf{R} =$
300 $\bigoplus_1^l \mathbf{W}_1$, where $\mathbf{W}$ is a diagonal matrix with elements $\mathbf{w}$, described above.
301
302 Multi-trait model configurations were YLD-DTF, YLD-PHT and Full (YLD-DTF-PHT), with
303 YLD the primary trait and DTF and PHT as the secondary traits. Multi-trait within a single
304 environment GBLUP model was fitted with the linear model:

305 $$\begin{bmatrix} \mathbf{BLUE}_{\mathbf{g}_1} \\ \vdots \\ \mathbf{BLUE}_{\mathbf{g}_\mathrm{t}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_{\mathrm{n}_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{1}_{\mathrm{n}_t} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_t \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{Z}_t \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_t \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_t \end{bmatrix} \quad \text{(Eq. 3)}$$

306 where, for each of t traits, $\mathbf{BLUE}_\mathbf{g}$ is the vector of accession fixed effects estimate, $\mathbf{1}_\mathrm{n}$ is a
307 vector of 1s, μ is the mean estimate, $\mathbf{Z}$ is the design matrix, $\mathbf{u}$ is the vector of GEBVs and $\mathbf{e}$ is
308 the vector of residuals. $[\mathbf{u}_1 \cdots \mathbf{u}_t]' \sim \mathrm{N}([0_1 \cdots 0_t]', \mathbf{T} \otimes \mathbf{A})$, where $\mathbf{A}$ is the SNP-based
309 relationship matrix and $\mathbf{T}$ is the $t \times t$ unstructured genetic variance-covariance matrix

310 $$\mathbf{T} = \begin{bmatrix} \sigma^2_{\mathrm{u}_1} & \cdots & \sigma_{\mathrm{u}_{1,t}} \\ \vdots & \ddots & \vdots \\ \sigma_{\mathrm{u}_{t,1}} & \cdots & \sigma^2_{\mathrm{u}_t} \end{bmatrix}$$

311 with σ^2_{u} and σ_{u} are the trait variance and the covariance between trait pairs, respectively.
312 $[\mathbf{e}_1 \cdots \mathbf{e}_t]' \sim \mathrm{N}([0_1 \cdots 0_t]', \mathbf{R})$, where $\mathbf{R}$ is as described above, with $\mathbf{R} = \bigoplus_1^t \mathbf{W}_1$, where $\mathbf{W}$ is a
313 diagonal matrix with elements $\mathbf{w}$, described above.
314

The multi-trait multi-environment model, and was fitted as follows:

$$\begin{bmatrix}\mathbf{BLUE}_{\mathbf{g}_1}\\ \vdots \\ \mathbf{BLUE}_{\mathbf{g}_p}\end{bmatrix} = \begin{bmatrix}\mathbf{1}_{\mathrm{n}_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{1}_{\mathrm{n}_p}\end{bmatrix}\begin{bmatrix}\mu_1\\ \vdots \\ \mu_p\end{bmatrix} + \begin{bmatrix}\mathbf{Z}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{Z}_p\end{bmatrix}\begin{bmatrix}\mathbf{u}_1\\ \vdots \\ \mathbf{u}_p\end{bmatrix} + \begin{bmatrix}\mathbf{e}_1\\ \vdots \\ \mathbf{e}_p\end{bmatrix}$$

where, for p trait-environment combinations, $\mathbf{BLUE_g}$ is the vector of accession fixed effects estimate, $\mathbf{1}_{\mathrm{n}}$ is a vector of 1s, μ is the mean estimate, $\mathbf{Z}$ is the design matrix, $\mathbf{u}$ is the vector of GEBVs and $\mathbf{e}$ is the vector of residuals. $[\mathbf{u}_1 \cdots \mathbf{u}_p]' \sim \mathrm{N}([0_1 \cdots 0_p]', \mathbf{G}\otimes\mathbf{A})$, where $\mathbf{A}$ is the SNP-based relationship matrix and $\mathbf{G}$ is a factor analytic variance-covariance matrix

$$\mathbf{G} = \mathbf{\Lambda}\mathbf{\Lambda}' + \boldsymbol{\psi}$$

where $\mathbf{\Lambda}$ is a $p \times k$ factor matrix of loadings and $\boldsymbol{\psi}$ is a $p \times p$ diagonal matrix of specific variance for each trait-environment combination. $[\mathbf{e}_1 \cdots \mathbf{e}_p]' \sim \mathrm{N}([0_1 \cdots 0_p]', \mathbf{R})$, where $\mathbf{R}$ is as described above, with $\mathbf{R} = \bigoplus_1^p \mathbf{W}_1$, where $\mathbf{W}$ is a diagonal matrix with elements $\mathbf{w}$, described above.

All the data analyses were done in the R programming language and environment (R Core Team, 2024). The $\mathbf{A}$ matrix was calculated using the *A.mat* function in *sommer* package (Covarrubias-Pazaran, 2016). All linear mixed models, including both univariate and multivariate GBLUP models were fitted using *ASReml-R* (Gilmour et al., 2009).

Heritability and Genetic Correlation Estimates

Broad-sense (H^2) and narrow-sense (h^2) heritability estimates were calculated for each environment following the linear mixed-model

$$\mathbf{y} = \mathbf{X_b}\mathbf{b} + \mathbf{Z_a}\mathbf{u_a} + \mathbf{Z_d}\mathbf{u_d} + \mathbf{Z_r}\mathbf{u_r} + \mathbf{Z_c}\mathbf{u_c} + \mathbf{e} \quad \text{(Eq. 4)}$$

Here, the accession effect $\mathbf{g}$ was fitted as random effect and $\mathbf{g} = \mathbf{u_a} + \mathbf{u_d}$, where $\mathbf{u_a}$ is the vector of GEBVs and $\mathbf{u_d}$ is the vector of non-additive genetic effects with $\text{var}(\mathbf{g}) = \sigma_a^2\mathbf{A} + \sigma_d^2\mathbf{I}$, where σ_a^2 as the additive genetic variance estimate, $\mathbf{A}$ is the SNP-based relationship matrix and σ_d^2 as the non-additive genetic variance estimate. The terms $\mathbf{y}$, $\mathbf{X_b b}$, $\mathbf{Z_r u_r}$, and $\mathbf{Z_c u_c}$ and corresponding variances were same as in Eqn. 1. Following Cullis et al. (2006), broad-sense heritability estimate was calculated as:

$$H^2 = 1 - \frac{\bar{v}_{\Delta..}^{GV}}{2\sigma_a^2 + 2\sigma_d^2} \text{ (Eq. 5)}$$

where $\bar{v}_{\Delta..}^{GV}$ is the mean pairwise prediction error variance of total genetic values (GV), while narrow-sense heritability was calculated following:

$$h^2 = 1 - \frac{\bar{v}_{\Delta..}^{GEBV}}{2\sigma_a^2} \text{ (Eq. 6)}$$

where $\bar{v}_{\Delta..}^{GEBV}$ is the mean pairwise prediction error variance of GEBVs.

Heritability estimates were also calculated across all environments based on the following linear mixed model:

$$\begin{bmatrix}\mathbf{y_1}\\ \vdots \\ \mathbf{y}_l\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1 & \cdots & 0\\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{X}_l\end{bmatrix}\begin{bmatrix}\mathbf{b}_1\\ \vdots \\ \mathbf{b}_l\end{bmatrix} + \begin{bmatrix}\mathbf{Z}_{\mathbf{a}_1} & \cdots & 0\\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{Z}_{\mathbf{a}_l}\end{bmatrix}\begin{bmatrix}\mathbf{u}_{\mathrm{a}_1}\\ \vdots \\ \mathbf{u}_{\mathrm{a}_l}\end{bmatrix} + \begin{bmatrix}\mathbf{Z}_{\mathbf{d}_1} & \cdots & 0\\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{Z}_{\mathrm{d}_l}\end{bmatrix}\begin{bmatrix}\mathbf{u}_{\mathrm{d}_1}\\ \vdots \\ \mathbf{u}_{\mathrm{d}_l}\end{bmatrix}$$

$$+ \begin{bmatrix}\mathbf{Z}_{\mathbf{r}_1} & \cdots & 0\\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{Z}_{\mathrm{r}_l}\end{bmatrix}\begin{bmatrix}\mathbf{u}_{\mathrm{r}_1}\\ \vdots \\ \mathbf{u}_{\mathrm{r}_l}\end{bmatrix} + \begin{bmatrix}\mathbf{Z}_{\mathbf{c}_1} & \cdots & 0\\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{Z}_{\mathrm{c}_l}\end{bmatrix}\begin{bmatrix}\mathbf{u}_{\mathrm{c}_1}\\ \vdots \\ \mathbf{u}_{\mathrm{c}_l}\end{bmatrix} + \begin{bmatrix}\mathbf{e}_1\\ \vdots \\ \mathbf{e}_l\end{bmatrix}$$

where $[\mathbf{b}_1 \cdots \mathbf{b}_l]'$ was the vector of fixed effects which included environment main effects and replicate effects within an environment, $[\mathbf{u}_{\mathrm{a}_1} \cdots \mathbf{u}_{\mathrm{a}_l}]'$, $[\mathbf{u}_{\mathrm{d}_1} \cdots \mathbf{u}_{\mathrm{d}_l}]'$, $[\mathbf{u}_{\mathrm{r}_1} \cdots \mathbf{u}_{\mathrm{r}_l}]'$, $[\mathbf{u}_{\boldsymbol{c}_1} \cdots \mathbf{u}_{\boldsymbol{c}_l}]'$

and $[\mathbf{e}_1 \cdots \mathbf{e}_l]'$ were the vector of random effects for additive genetic effects or GEBVs, non-additive genetic effects, row and column random effects, respectively. $\left[\mathbf{u}_{\mathrm{a}_1} \cdots \mathbf{u}_{\mathrm{a}_l}\right]' \sim \mathrm{N}([0_1 \cdots 0_l]', \mathbf{L}_\mathrm{a} \otimes \mathbf{A})$ where $\mathbf{L}_\mathrm{a}$ was the unstructured additive genetic variance-covariance matrix. $\left[\mathbf{u}_{\mathrm{d}_1} \cdots \mathbf{u}_{\mathrm{d}_l}\right]' \sim \mathrm{N}([0_1 \cdots 0_l]', \mathbf{L}_\mathrm{d} \otimes \mathbf{A})$ where $\mathbf{L}_\mathrm{d}$ was the unstructured additive genetic variance-covariance matrix. $\left[\mathbf{u}_{\mathrm{r}_1} \cdots \mathbf{u}_{\mathrm{r}_l}\right]' \sim \mathrm{N}([0_1 \cdots 0_l]', \mathbf{L}_\mathrm{r} \otimes \mathbf{I})$, $\left[\mathbf{u}_{\mathrm{c}_1} \cdots \mathbf{u}_{\mathrm{c}_l}\right]' \sim \mathrm{N}([0_1 \cdots 0_l]', \mathbf{L}_\mathrm{c} \otimes \mathbf{I})$ and $[\mathbf{e}_1 \cdots \mathbf{e}_l]' \sim \mathrm{N}([0_1 \cdots 0_l]', \mathbf{R} \otimes \mathbf{I})$, where $\mathbf{L}_\mathrm{a}$, $\mathbf{L}_\mathrm{a}$ and $\mathbf{R}$ were diagonal heterogeneous variance covariance matrices for row, column and residual terms, respectively. Narrow-sense and broad-sense heritabilities were estimated using Eq. 5 and Eq. 6, respectively. However, the σ_a^2 and σ_d^2 were estimated from the average of covariances between environments $\mathbf{L}_\mathrm{a}$ and $\mathbf{L}_\mathrm{d}$, respectively.

Genetic correlations for the trait pairs YLD-DTF and YLD-PHT were obtained for each environment. These were calculated from variance and covariance estimates based on Eq. 3, that is $\rho_{\mathrm{YLD,DTF}} = \frac{\sigma_{\mathrm{YLD,DTF}}}{\sqrt{\sigma_{\mathrm{YLD}}^2 \sigma_{\mathrm{DTF}}^2}}$ and $\rho_{\mathrm{YLD,PHT}} = \frac{\sigma_{\mathrm{YLD,PHT}}}{\sqrt{\sigma_{\mathrm{YLD}}^2 \sigma_{\mathrm{PHT}}^2}}$. Likewise, to assess the degree of genotype-by-environment interactions, genetic correlations for each trait across environments were calculated by fitting Eq. 3. The genetic correlation between pairs of environments is thus $\rho_{ij,i\neq j} = \frac{\sigma_{ij}}{\sqrt{\sigma_i^2 \sigma_j^2}}$ where i and j were pair of environments.

Cross-Validation Schemes

Performances of GBLUP models were assessed using 5-fold cross-validation with different schemes. The schemes varied based on 1) whether the validation set had measurements for the secondary traits, 2) whether the validation set had been tested in other environments, and 3)

whether accessions were randomly assigned to training and validation sets or based on DAPC grouping.

For single-trait single-environment models, the standard scheme was used (Fig. 1a). In the case of multi-trait models, two cross-validation schemes tested whether measuring secondary traits (DTF, PHT) on validation lines improved YLD prediction. In the **SEC** scheme, the validation set had no YLD data but were measured for secondary traits. In the **NOSEC** scheme, validation set had no data for any trait.

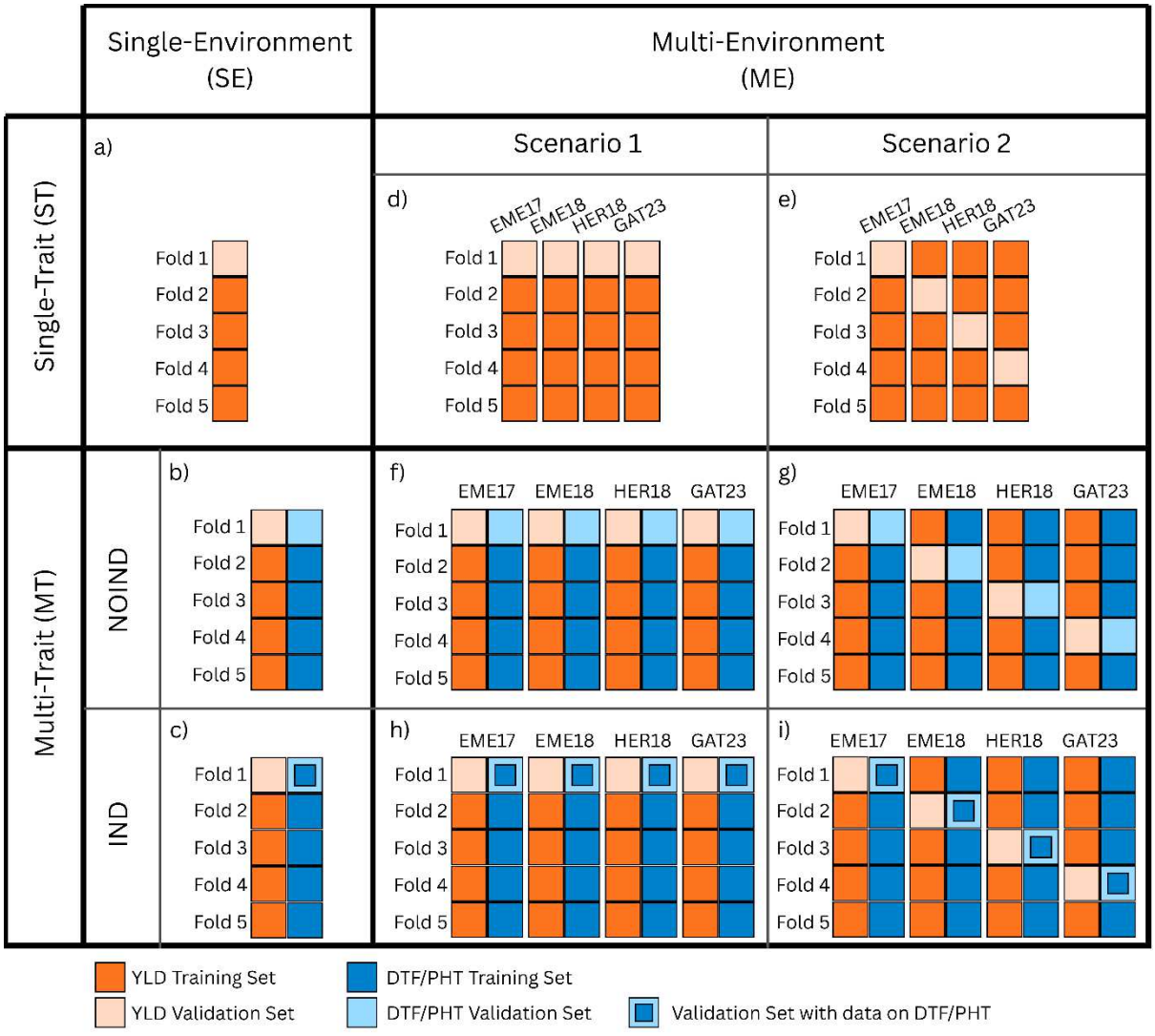


Fig. 1 Cross-validation schemes for GBLUP models. (a) Single-Trait Single-Environment: target trait masked in target environment only. (b–c) Multi-Trait Single-Environment: validation sets with all traits masked (NOSEC) or only target trait masked (SEC). (d–e) Single-Trait Multi-Environment: validation sets masked in all environments (Scenario 1) or target environment only (Scenario 2). (f–i) Multi-Trait Multi-Environment: factorial combinations of NOSEC/SEC and Scenario 1/2.

The schemes for multi-environment models tested whether performance data from other environments could improve prediction in a target environment. **Scenario 1** represented prediction for untested accessions. The validation set had no performance data from any environment, so predictions relied entirely on the training set. **Scenario 2** represented prediction of accessions tested elsewhere. The validation set had been evaluated in all non-target environments, allowing predictions to use both training set performance across all the environments and validation set performance in other environments.

NOSEC and **SEC** were implemented on multi-trait single-environment models (Fig. 1b and 1c), while **Scenario 1** and **Scenario 2** were used on single-trait, multi-environment models (Fig. 1d and 1e). The combination of these schemes, that is, **NOSEC-Scenario 1**, **NOSEC-Scenario 2**, **SEC-Scenario 1** and **SEC-Scenario 2** were employed on multi-trait, multi-environment models (Fig. 1f, 1g, 1h and 1i, respectively).

In all the schemes described above, the assignment of accessions into training and validation sets were random. To assess genomic prediction when the genetic distance between the training set and validation set is at a maximum, DAPC cluster-based assignment of accessions to training and validation sets was also implemented. Comparison of random and cluster-based fold assignment was evaluated using single-trait single environment, multi-trait single-

environment (under **NOSEC**), single-trait multi-environment (under **Scenario 1**), and multi-trait multi-environment (**NOSEC-Scenario 1**) models.

The GBLUP models were assessed based on their prediction accuracy calculated as the Pearson's correlation between GEBVs $\hat{\mathbf{u}}$ and $\mathbf{BLUE_g}$, $cor(\hat{\mathbf{u}}, \mathbf{BLUE_g})$. The performance of multi-trait single-environment, single-trait multi-environment, and multi-trait multi-environment GBLUP models was compared to single-trait single-environment models using percent change in prediction accuracy calculated as $100 \times \left(\frac{\text{pa}_{\text{M}_{tl}}}{\text{pa}_{\text{ST-SE}_{tl}}} - 1\right)$, where $\text{pa}_{\text{M}_{tl}}$ is the prediction accuracy of the multi-trait single-environment, single-trait multi-environment or multi-trait multi-environment models for trait t at environment l, and $\text{pa}_{\text{ST-SE}_{tl}}$ is the prediction accuracy of single-trait single-environment model for trait t at environment l.

Effect of Marker Density

The effect of marker density on GBLUP prediction accuracy was assessed using three marker sets: the full set of 3,310 SNPs (Full) and two randomly selected subsets of 2,000 SNPs (2K) and 1,000 SNPs (1K). These marker sets were used to construct genomic relationship matrix **A** for single-trait single-environment, multi-trait single-environment (NOSEC), single-trait multi-environment (Scenario 1), and multi-trait multi-environment (NOSEC-Scenario 1) models.

Results

Genetic Diversity and Population Structure of the Mini-Core Plus Panel

The optimal number of principal components for DAPC was determined to be 25 using cross-validation approach which explained ~50% of the total variation. Four discriminant functions were then derived from these 25 principal components. Using these discriminant functions, the 313 accessions in the Mini-Core Plus Panel were classified into five distinct genetic clusters (Fig. 2). Most of the accessions which were released as commercial varieties in Australia were in Cluster 5.

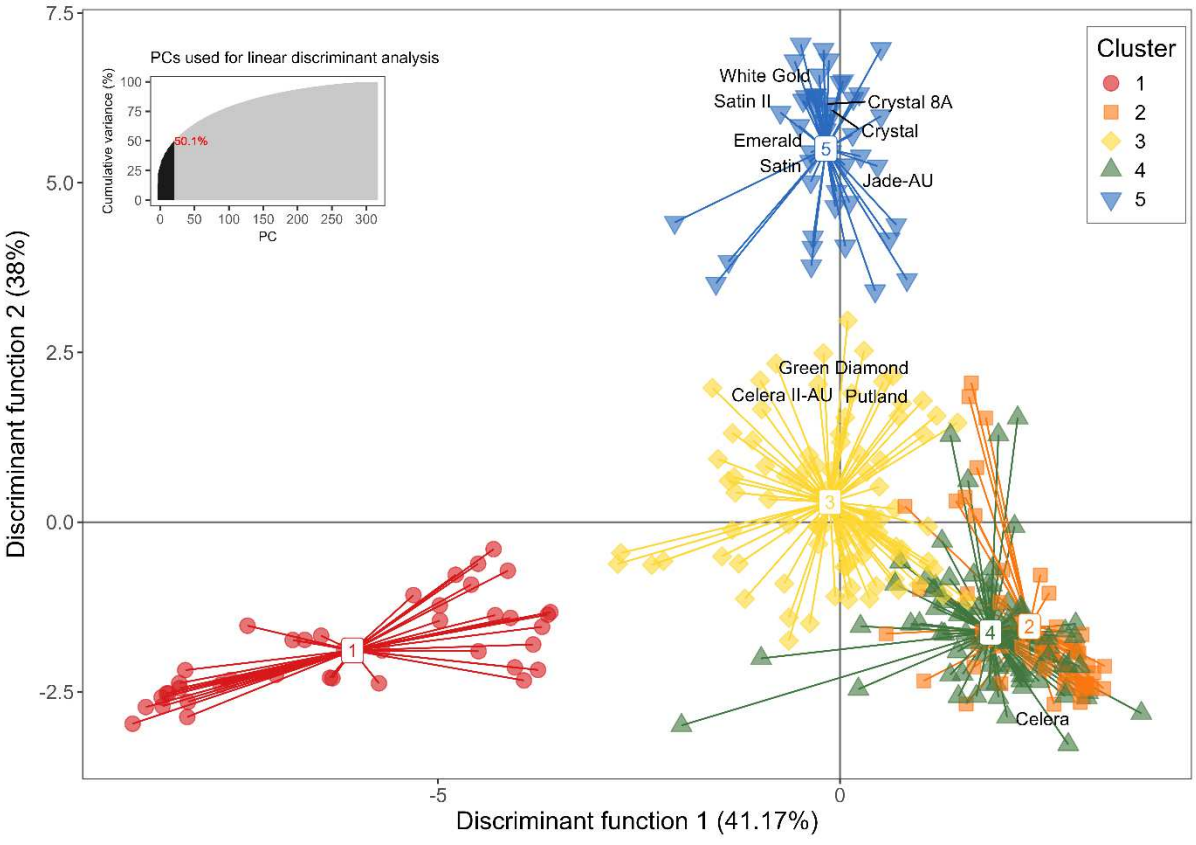


Fig. 2 DAPC-based genetic clustering of 313 mungbean accessions from the Mini-Core Plus Panel. Five distinct clusters were identified: Cluster 1 (n=42), Cluster 2 (n=52), Cluster 3 (n=90), Cluster 4 (n=81), and Cluster 5 (n=48). Labelled points were accessions released as commercial varieties in Australia

Genetic differentiation between clusters were quantified, using pairwise F_{ST}. F_{ST} values ranged from 0.06 to 0.40, with Cluster 1 and Cluster 5 showing the highest divergence (Supp Fig. 1). Mean pairwise F_{ST} values (averaged across comparisons with the four other clusters) were highest for Cluster 1 (0.34) and lowest for Cluster 3 (0.15).

Expected heterozygosity for each cluster was also calculated to assess genetic diversity within clusters. Expected heterozygosity (H_e) was lowest in Clusters 1 and 5 (0.17), intermediate in Clusters 2 (0.21) and 3 (0.22), and highest in Cluster 4 (0.26). Across clusters, H_e was 0.26.

To examine associations between cluster membership and phenotypic or geographic characteristics, clusters were compared based on geographic origin, seed size (seed length, seed width), thousand seed weight, and seed coat luster. Accessions from India were well represented across all clusters (Supp Fig. 2a). In addition, Clusters 3 and 5 had accessions from Australia, East and South-East Asian countries, while Clusters 1, 2 and 4 were composed of materials from Indo-Pakistani and European region. Cluster membership showed no clear association with seed coat luster (Supp Fig. 2b). However, seed size and weight differed among clusters, with Cluster 5 having the largest seeds and highest thousand seed weight and Cluster 1 having the smallest seeds and lowest thousand seed weight (Supp Fig. 2c).

The extent of linkage disequilibrium is a key parameter which determines the accuracy of genomic predictions and can be used to evaluate marker density necessary for genomic prediction. Depending on the threshold, in our mungbean population the LD decays at around 171kbp, 174kbp and 242kbp for $LD_{0.5max}$, LD_{20} and LD_{10}, respectively (Supp Fig. 3). In

addition, the LD calculated within each cluster varied (Supp Fig. 4). Based on LD_{10}, LD decays even faster at ~187kpb, ~206kbp, and ~192kbp for clusters 2, 3, and 4, respectively, compared to that of the entire Mini-Core Plus Panel. In contrast, clusters 1 and 5 have LD persisting at longer distances with LD_{10} at ~320kbp and ~323kbp, respectively.

Heritability and Genetic Correlation Estimates

Broad-sense heritability (H^2) for YLD ranged from 0.66 to 0.89 within environments and was 0.66 across environments, for DTF ranged from 0.77 to 0.96 within environments and 0.94 across environments, and for PTH ranged from 0.74 to 0.79 within environments and 0.74 across environments (Table 3). The lowest H^2 estimates within environments were consistently observed in EME17.

Table 3. Trait heritability estimates in each and across environments

Trait[a]	**Environment**[b]	**Narrow-sense Heritability** (h^2)	**Broad-sense Heritability** (H^2)
YLD	EME17	0.41	0.66
	EME18	0.58	0.87
	HER18	0.51	0.89
	GAT23	0.68	0.76
	Across Trials	0.62	0.66
DTF	EME17	0.77	0.77
	EME18	0.66	0.89
	HER18	0.96	0.96
	GAT23	0.93	0.93
	Across Trials	0.88	0.94
PHT	EME17	0.59	0.74
	EME18	0.45	0.79
	HER18	0.54	0.79
	Across Trials	0.64	0.74

[a] YLD, yield; DTF, days to 50% flowering; PHT, plant height
[b] EME17—Emerald 2017; EME18—Emerald 2018; HER18—Hermitage 2018; GAT23—Gatton 2023

Narrow-sense heritability (h^2) ranged from 0.41 to 0.68 within environments and was 0.62 across environments for YLD, from 0.66 to 0.96 within environments and 0.88 across environments for DTF, and from 0.45 to 0.59 within environments and 0.64 across environments for PHT (Table 3).

From the multi-trait analysis, the YLD trait in most cases was negatively correlated with DTF and PHT, though the degree of correlation varied across trials (Fig. 3). The ranking based on the absolute genetic correlation followed the order GAT23>HER18>EME18>EME17 for both YLD-DTF (0.92±0.03, 0.61±0.10, 0.39±0.14, 0.10±0.24) and YLD-PHT (excluding GAT23, 0.40±0.15, 0.08±0.18, 0.02±0.28). On the average, the absolute value of genetic correlations between YLD-DTF was higher than that of YLD-PHT.

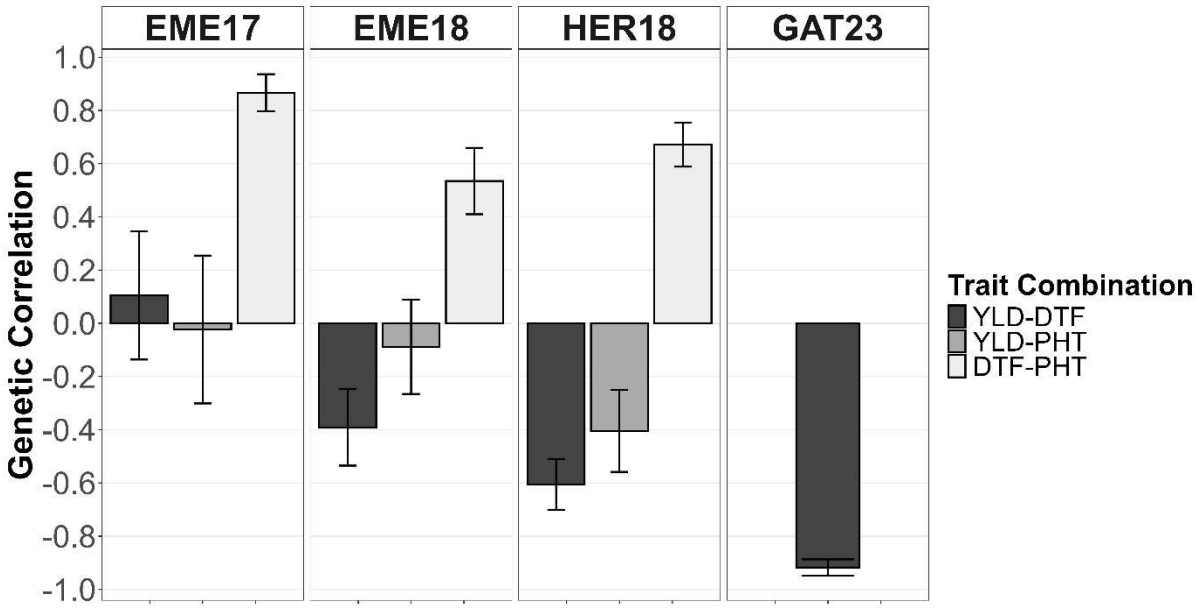


Fig. 3 Estimate of genetic correlations between trait pairs yield and days-to-50%-flowering (YLD-DTF), yield and plant height (YLD-PHT) and days-to-50%-flowering and plant height

(DTF-PHT) in each environment (EME17, Emerald 2017; EME18, Emerald 2018; HER18, Hermitage 2018; GAT23, Gatton 2023). Error bars are standard error of the estimates

Higher degree of genotype-by-environment interactions were observed for YLD compared to DTF and PHT. Low to moderate genetic correlations between environments were observed for YLD, which ranged from 0.03±0.16 to 0.44±0.25 (Fig. 4a). In contrast higher levels of genetic correlations between environments were estimated for DTF (Fig. 4b) and PHT (Fig. 4c), ranging from 0.83±0.06 to 0.94±0.04, and from 0.80±0.08 to 0.96±0.04, respectively.

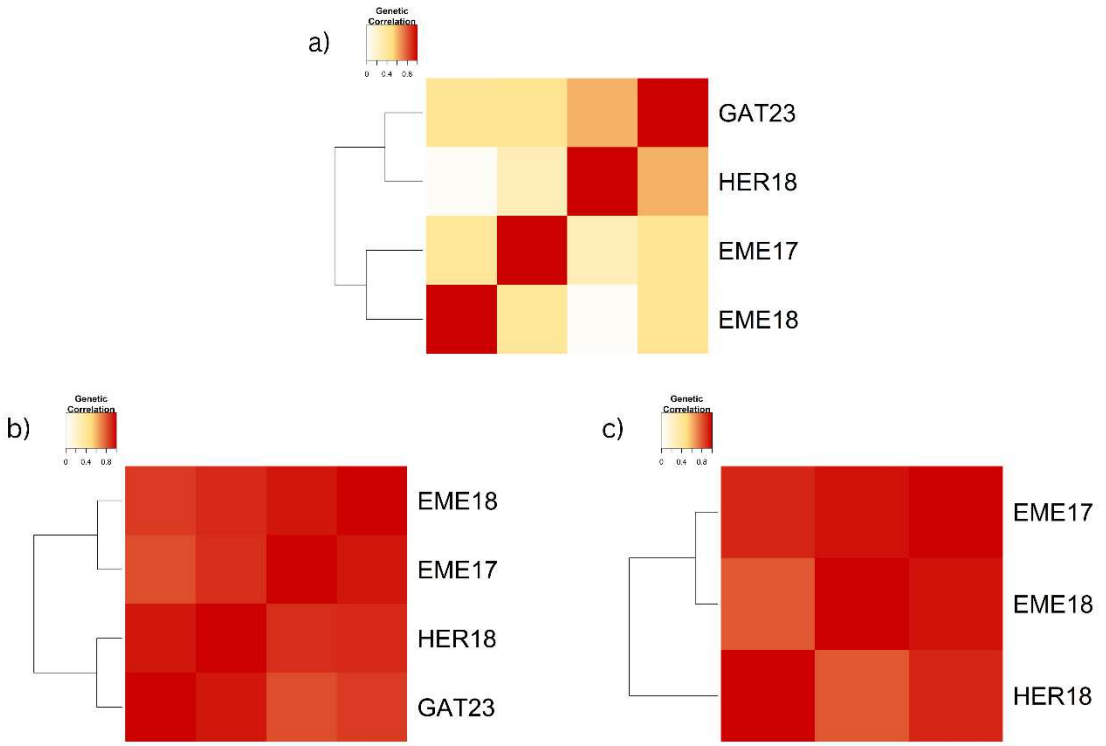


Fig. 4 Genetic correlations between environments for a) yield (YLD), b) days-to-50%-flowering (DTF) and c) plant height (PHT). Environments were Emerald 2017 (EME17), Emerald 2018 (EME18), Hermitage 2018 (HER18), and Gatton 2023 (GAT23)

In all the traits, hierarchical clustering based on genetic correlation between environments showed consistent grouping of EME17 with EME18, and HER18 with GAT23. Both EME17

and EME18 were rainfed and had higher average temperature, while HER18 and GAT23 were both irrigated and had lower average temperature.

Prediction Accuracies under Random Cross-Validation Schemes

With cross folds allocated at random, prediction accuracies of single-trait single-environment models were moderate to high. For YLD, it ranged from 0.34±0.04 to 0.60±0.02, 0.42±0.03 to 0.47±0.04 for DTF, and 0.40±0.03 to 0.63±0.02 for PHT (Fig. 5). Prediction accuracies were lowest for environments with the lowest narrow-sense heritability, except for PHT.

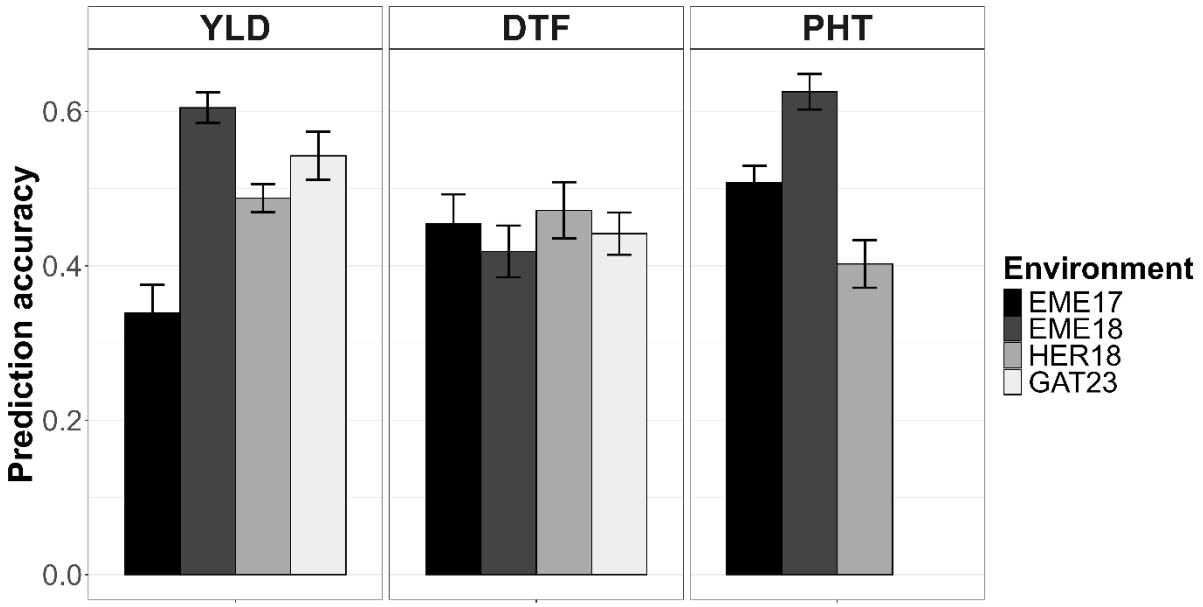


Fig. 5 Prediction accuracies of single-trait single-environment models for yield (YLD), days to 50% flowering (DTF), and plant height (PHT) across four environments: Emerald 2017 (EME17), Emerald 2018 (EME18), Hermitage 2018 (HER18), and Gatton 2023 (GAT23). Error bars represent standard errors of the mean across five cross-validation folds

Gains in prediction for YLD using multivariate models

The performance of multi-trait single-environment models varied between cross-validation schemes. Under the NOSEC scheme, where the validation set had no data on any traits at the time of prediction, multi-trait single-environment showed no improvement in YLD prediction accuracy compared to single-trait single-environment (Fig. 6). However, under the SEC scheme, where secondary trait data were available for the validation set at the time of prediction, multi-trait single-environment prediction accuracy increased substantially in GAT23 (+42.6%) and HER18 (+19.6%), while no advantage was observed in EME17 and EME18. Among multi-trait single-environment configurations, using PHT alone as secondary trait (MT-SE_YLD-PHT) resulted in lower prediction accuracy than using DTF alone (MT-SE_YLD-DTF) or both traits together (MT-SE_Full), which provided similar improvements.

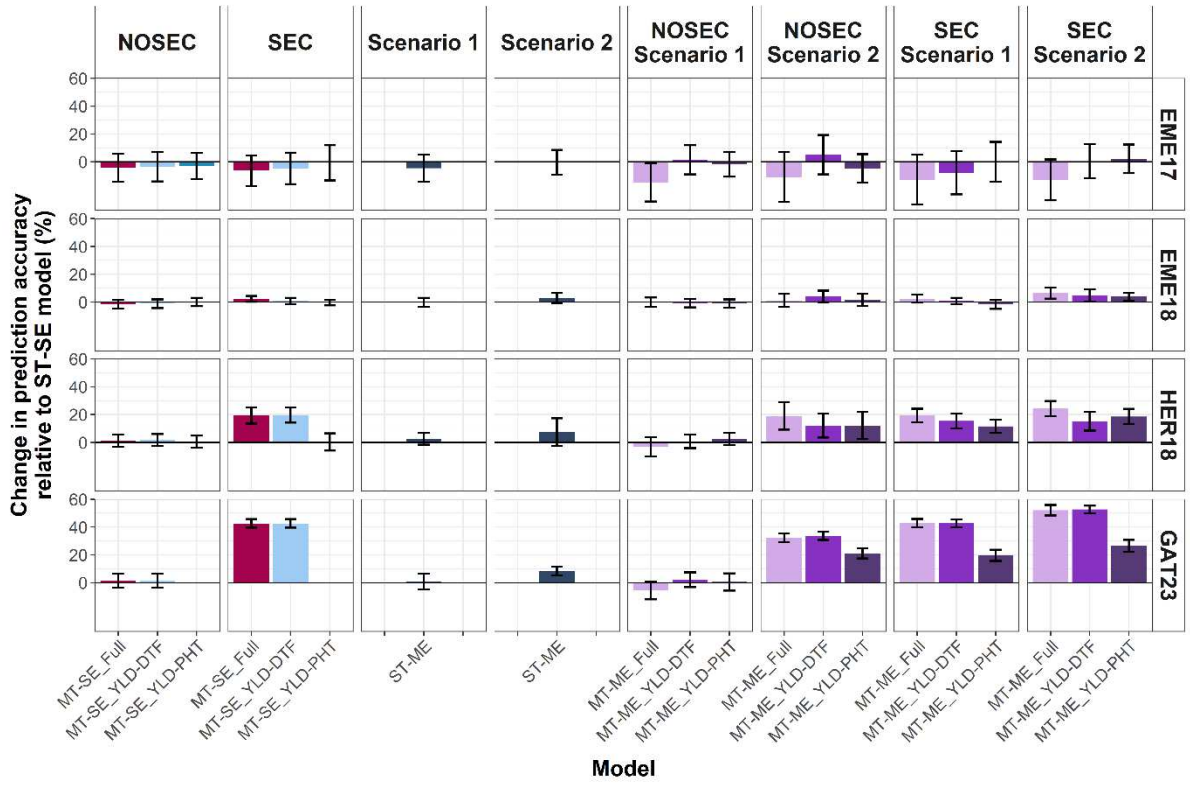


Fig. 6 Percent change in prediction accuracy relative single-trait single-environment (**ST-SE**) for yield (YLD), across multivariate models (**MT-SE**—multi-trait single-environment; **ST-ME**—single-trait single-environment; **MT-ME**—multi-trait multi-environment), cross-validation schemes (**NOSEC**—without secondary trait data on the validation set; **SEC**—with secondary trait data on the validation set; **Scenario 1**—validation set not tested on any environment; **Scenario 2**—validation set tested elsewhere) and environments (EME17—Emerald 2017; EME18—Emerald 2018; HER18—Hermitage 2018; GAT23—Gatton 2023). Error bars are the standard error of the mean.

In the case of single-trait multi-environment model, performance depended on cross-environment data availability. Under Scenario 1, where the validation set had no data from any environment, single-trait multi-environment showed no improvement over single-trait single-environment for YLD. Under Scenario 2, where validation sets in the target environment included data from non-target environments, single-trait multi-environment achieved modest YLD prediction gains ranging from ─0.4% (EME17) to +8.5% (GAT23).

For multi-trait multi-environment models, secondary trait availability (NOSEC vs SEC) and cross-environment data availability (Scenario 1 vs 2) were evaluated simultaneously, resulting to four distinct schemes.

NOSEC-Scenario 1. When validation sets lacked both secondary trait and cross-environment data, multi-trait multi-environment models provided negligible YLD prediction improvement. The maximum gain was only +2.6% in GAT23 using multi-trait multi-environment with DTF as secondary trait (MT-ME_YLD-DTF).

NOSEC-Scenario 2. When validation sets included cross-environment data but lacked secondary traits, multi-trait multi-environment performance improved substantially in GAT23 (+21.0% to +33.6%, depending on secondary trait) and moderately in HER18 (+12.1% to

+19.0%, highest with both secondary traits). EME17 and EME18 showed minimal gains (≤5.1%).

SEC-Scenario 1. When validation sets had secondary trait data across all environments but no YLD data, multi-trait multi-environment achieved large improvements in GAT23 (+19.7% to +42.8%, highest with both secondary trait) and HER18 (+11.6% to +19.4%, also highest with both secondaries). EME17 and EME18 showed no substantial improvement.

SEC-Scenario 2. With both secondary traits and cross-environment YLD data available in validation sets, multi-trait multi-environment reached the highest overall performance. GAT23 showed the largest gain (+52.7% with DTF as secondary trait), followed by HER18 (+24.3% with both secondary traits) and EME18 (+6.4% with both secondary traits). EME17 showed no improvement.

Ranking models and schemes by magnitude of YLD prediction gains, multi-trait multi-environment under SEC-Scenario 2 showed the largest improvements, followed by multi-trait multi-environment under SEC-Scenario 1 and multi-trait single-environment under SEC (similar levels), then multi-trait multi-environment under NOSEC-Scenario 2, then single-trait multi-environment under Scenario 2. Multi-trait single-environment under NOSEC, single-trait multi-environment under Scenario 1, and multi-trait multi-environment under NOSEC-Scenario 1 showed negligible differences from single-trait single-environment.

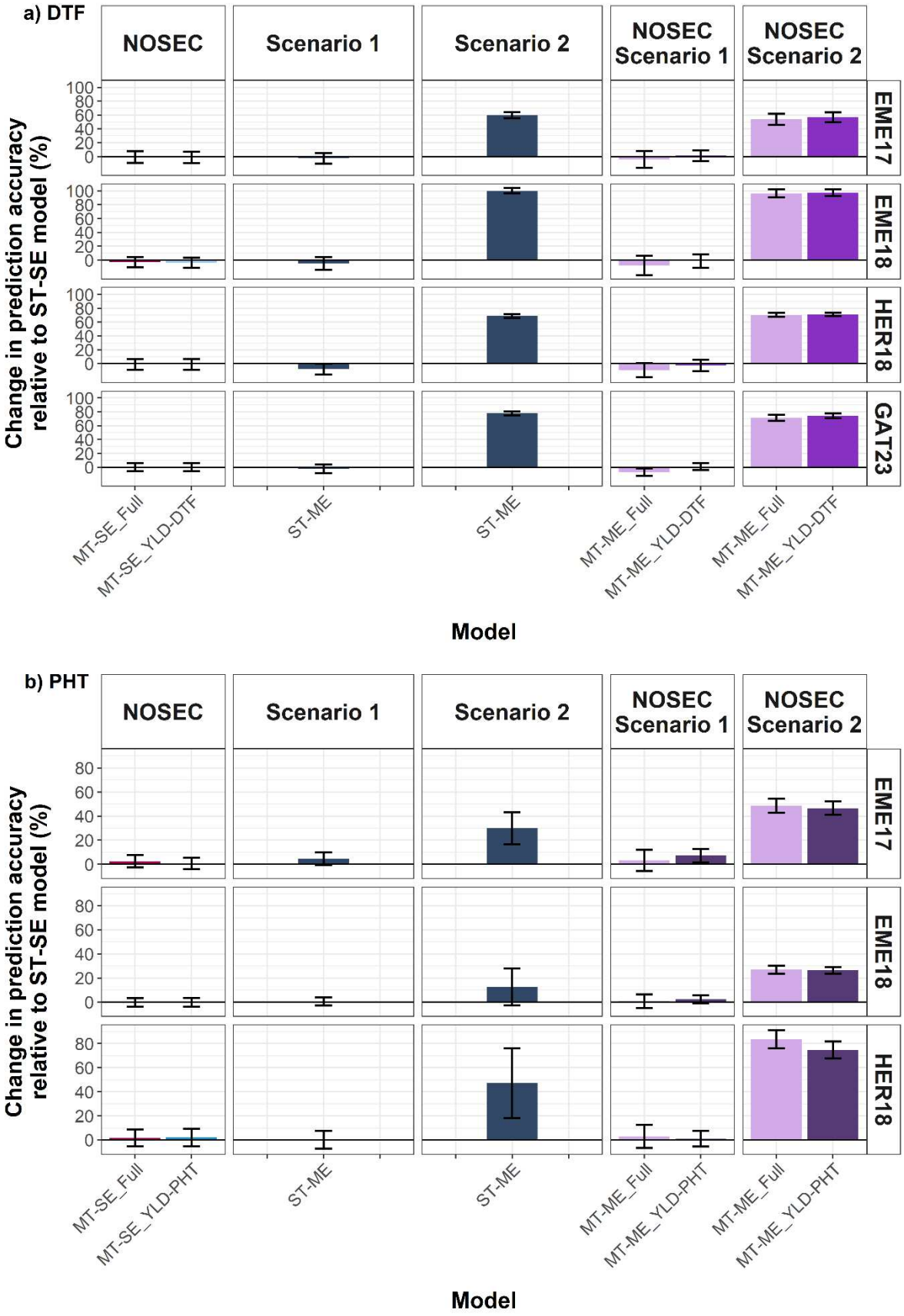
Environment-specific responses were consistent across all models and schemes. GAT23 consistently showed the highest prediction accuracy gains, followed by HER18. EME17 and EME18 showed minimal to no improvement regardless of model or scheme.

600

601 **Gains in prediction for DTF and PHT using multivariate models**

602 Prediction accuracies for DTF and PHT also varied based on models and schemes (Fig. 7).
603 Results are reported only for schemes where these traits were masked in validation sets
604 (NOSEC, Scenario 1, NOSEC-Scenario 1, Scenario 2, and NOSEC-Scenario 2). Schemes
605 where DTF or PHT were observed (SEC, SEC-Scenario 1, SEC-Scenario 2) represent in-
606 sample predictions and are not reported.

607



608

609 **Fig. 7** Percent change in prediction accuracy relative single-trait single-environment (ST-SE)
610 for a) days-to-50%-flowering (DTF) and b) plant height (PHT), across multivariate models

(**MT-SE**—multi-trait single-environment; **ST-ME**—single-trait single-environment; **MT-ME**—multi-trait multi-environment), cross-validation schemes (**NOSEC**—without secondary trait data on the validation set; **Scenario 1**—validation set not tested on any environment; **Scenario 2**—validation set tested elsewhere) and environments (EME17—Emerald 2017; EME18—Emerald 2018; HER18—Hermitage 2018; GAT23—Gatton 2023). Results shown are for NOSEC, Scenario 1, NOSEC-Scenario 1, Scenario 2, and NOSEC-Scenario 2 schemes, only. Error bars are the standard error of the mean

Similar with YLD, predictions for DTF and PHT of multi-trait single-environment models under NOSEC, single-trait multi-environment under Scenario 1, and multi-trait multi-environment models under NOSEC-Scenario 1 showed no improvement while substantial gains were observed from Scenario 2 and NOSEC-Scenario 2. In DTF, Scenario 2 (ranging from +59.53% to +100.22%) and NOSEC-Scenario 2 (ranging from +53.79% to +97.44%) provided similar levels of gains. For PHT, gains in prediction accuracy under NOSEC-Scenario 2 (ranging from +26.42% to +83.43%) were consistently higher than Scenario 2 (ranging from +12.67% to +47.12%).

Prediction accuracies under Cluster-based Cross-Validation Schemes

Prediction accuracies under cluster-based fold assignment were compared to random assignment for single-trait single-environment, multi-trait single-environment (under NOSEC), single-trait multi-environment (under Scenario 1), and multi-trait multi-environment (under NOSEC-Scenario 1) models. Cluster-based assignment consistently produced lower accuracy than random assignment across all models (Fig. 8), with reductions ranging from 0.02 to 0.34 (mean = 0.20).

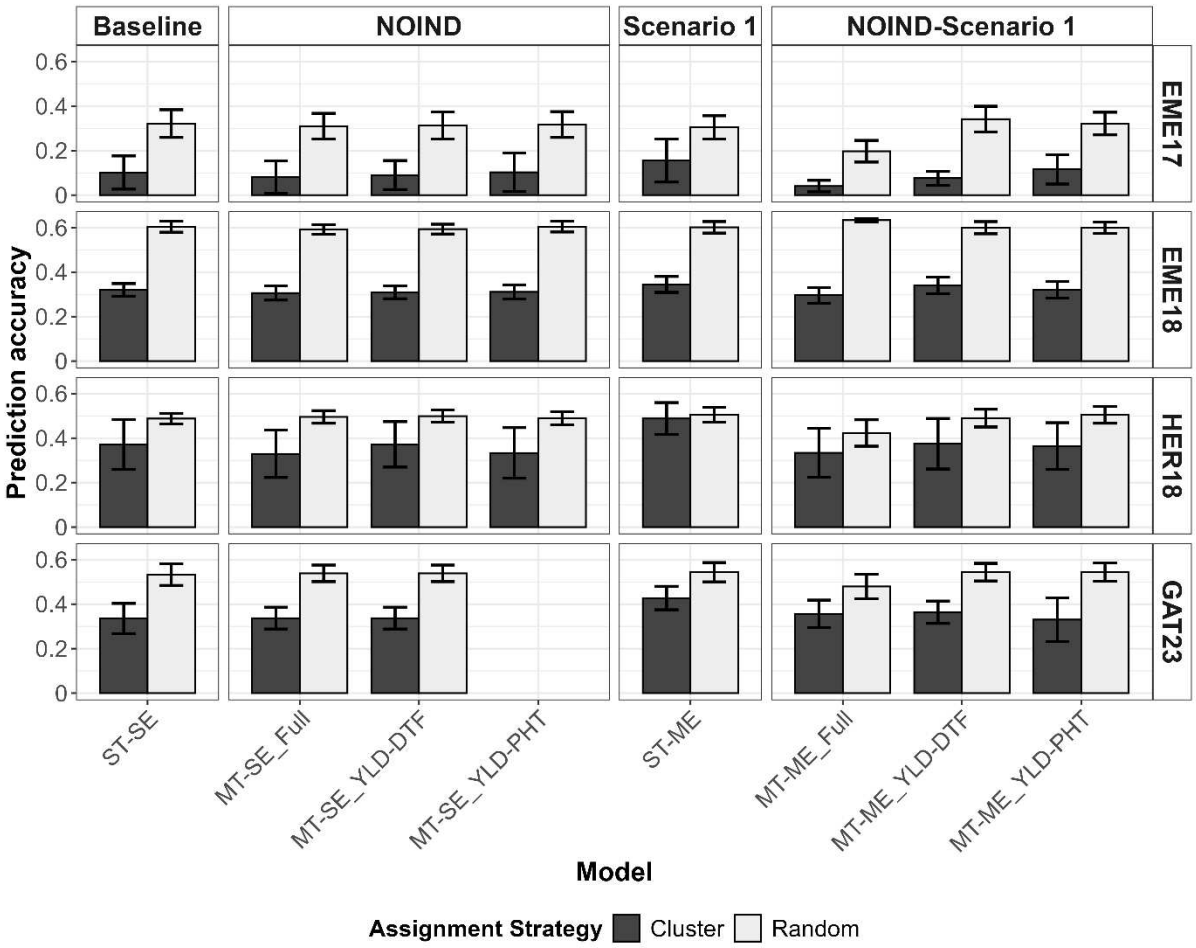


Fig. 8 Yield (YLD) prediction accuracy under cluster-based versus random fold assignment for single-trait single-environment (ST-SE), multi-trait single-environment (MT-SE) under NOSEC, single-trait multi-environment (MT-SE) under Scenario 1, and multi-trait multi-environment (MT-ME) under NOSEC-Scenario 1, across environments (EME17—Emerald 2017; EME18—Emerald 2018; HER18—Hermitage 2018; GAT23—Gatton 2023). Error bars represent standard errors of the mean across cross-validation folds

Analysis of single-trait single-environment predictions by validation cluster revealed substantial cluster-specific variation (Fig. 9). Prediction accuracy was lowest when Cluster 1 served as the validation set for YLD (0.08±0.07) and when Cluster 2 was the validation set for DTF (0.15±0.05) and PHT (0.06±0.08). Conversely, Cluster 5 consistently showed the highest accuracy across traits (0.32±0.17 to 0.48±0.03).

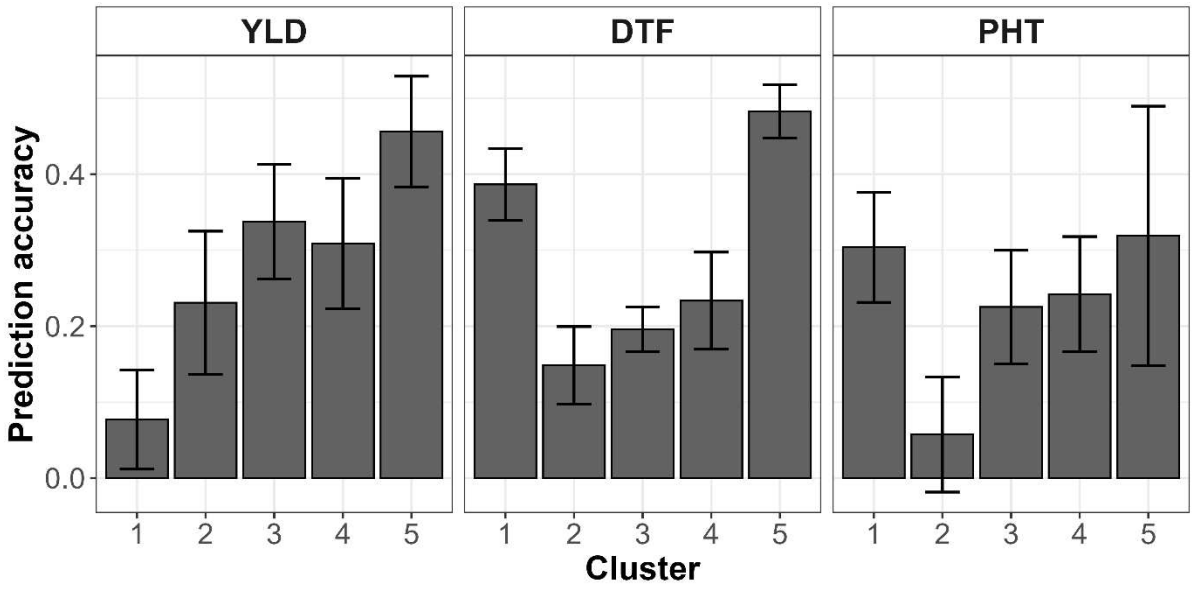


Fig. 9 Prediction accuracy of single-trait single-environment models according to DAPC cluster-based validation set. Error bars are the standard error of the mean across environments

Effect of Marker Density on Prediction accuracy

Prediction accuracies of GBLUP models using different marker densities were evaluated under the random NOSEC-Scenario 2 scheme. In general, as the marker density decreased, the prediction accuracy decreased slightly (Fig. 10). Against the full marker set (Full), the average change in prediction accuracy across environments and across GBLUP models were, on YLD -0.01 and -0.03, on DTF -0.02 and -0.07, and on PHT -0.02 and -0.04, for 2K and 1K, respectively.

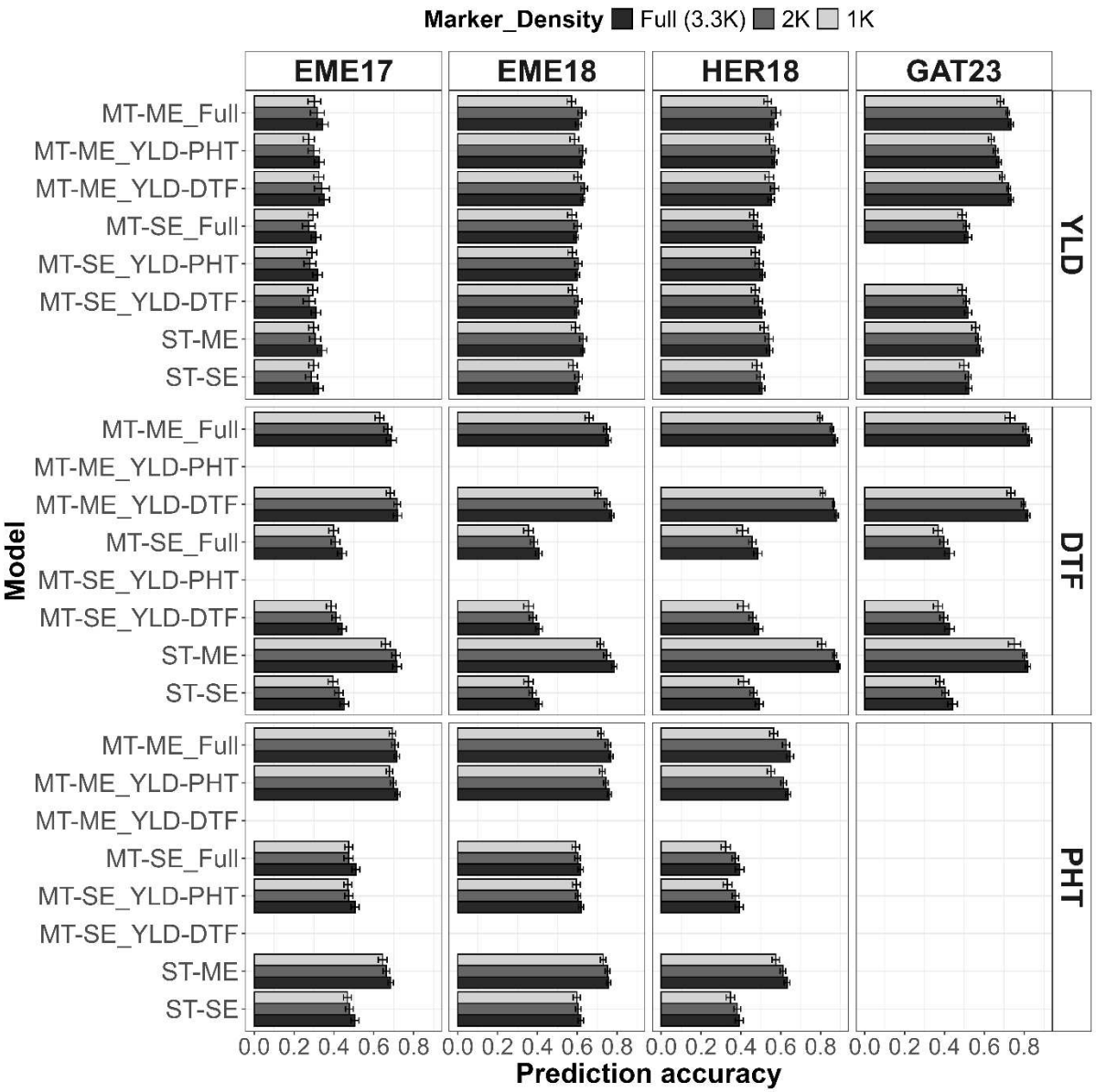


Fig. 10 Prediction accuracy of GBLUP models using different marker densities: 3, 310 (Full), 2000 (2K) and 1000 (1K) SNPs. Models were single-trait single-environment (ST-SE), multi-trait single-environment (MT-SE) under NOSEC, single-trait multi-environment (ST-ME) under Scenario 1, and multi-trait multi-environment (MT-ME) under NOSEC-Scenario 1

Discussion

Genetic diversity and population structure inform genomic prediction strategies

The Mini-Core Plus Panel exhibited substantial genetic diversity and clear population structure, with five distinct clusters showing varying degrees of differentiation (F_{ST} = 0.06-0.40). This genetic structure provided an ideal framework for evaluating genomic prediction across diverse germplasm. Cluster 1 was the most genetically divergent from other clusters (mean F_{ST} = 0.34), while Cluster 3 showed the greatest genetic similarity to other clusters (mean F_{ST} = 0.15). Within-cluster expected heterozygosity (H_e = 0.17-0.26) was lower than that of the entire panel (H_e = 0.26), indicating that within-cluster diversity is lower than panel-wide diversity.

The geographic distribution of accessions reveals the influence of different growing environment on the genetic differentiation of mungbean. Clusters 3 and 5 are characterized by an Indo-Pacific distribution, including materials from Australia and East/South-East Asia, which represent tropical conditions. Clusters 1, 2, and 4, primarily composed accessions from temperate Indo-Pakistani and European regions. While Indian accessions are well represented across clusters, it is worth noting that within Indian peninsula, diverse agro-climatic zone exists.

The strong differentiation in seed size and weight provides evidence of divergent selection. Cluster 5 contained genotypes with the largest seeds and highest thousand seed weight, while Cluster 1 had the smallest seeds and lowest weight. Most Australian commercial varieties and

advanced breeding lines grouped in Cluster 5, suggesting that Australian breeding programs have selected for larger seed size, which is a commercially important trait. This concentration of elite Australian germplasm in a single genetic cluster has important implications for genomic prediction in breeding programs.

Genomic prediction is feasible in mungbean with moderate to high accuracy

Single-trait single-environment GBLUP models achieved moderate to high prediction accuracies under random cross-validation (YLD: 0.34-0.60; DTF: 0.42-0.47; PHT: 0.40-0.63), demonstrating that genomic selection is feasible in mungbean breeding programs. These accuracies are comparable to or exceed those reported in other legume crops. For example, studies in common bean (*Phaseolus vulgaris*) reported prediction accuracies of 0.30-0.50 for yield (Keller et al., 2020), while chickpea (*Cicer arietinum*) studies found accuracies of 0.40-0.60 for various agronomic traits (Roorkiwal et al., 2018). The prediction accuracies observed in our study are sufficiently high to accelerate genetic gain in mungbean breeding programs, particularly when combined with shorter generation intervals enabled by speed breeding technologies.

The relationship between narrow-sense heritability and prediction accuracy was generally positive, with environments showing higher h^2 typically yielding higher prediction accuracies. However, this relationship was not absolute, as other factors including population structure, trait architecture, and genotype-by-environment interactions also influenced prediction performance. The moderate to high heritabilities observed (H^2 = 0.66-0.96; h^2 = 0.41-0.96)

indicate that the phenotypic data were of sufficient quality for genomic prediction. Environment EME17 consistently showed the lowest heritability estimates across all traits, likely due to environmental stress or management factors. Nevertheless, prediction accuracies in EME17 were still moderate (YLD = 0.34±0.04), which demonstrates the robustness of genomic prediction even under suboptimal phenotyping conditions.

Genotype-by-environment interactions differ substantially among traits

The magnitude of genotype-by-environment interactions varied among traits, with important implications for breeding strategies and prediction approaches. Yield showed strong G×E, with genetic correlations between environments ranging from only 0.03±0.16 to 0.44±0.25. In contrast, days to flowering and plant height exhibited weak G×E, with genetic correlations consistently high (DTF: 0.83-0.94; PHT: 0.80-0.96). Dudley et al. (2025) reported similarly low G×E for days to flowering using a nested association mapping population of mungbean. On one hand, these results reflect the complex biological nature of yield, which is greatly influenced by numerous environmental factors, while on the other hand, show that flowering time and plant height are less sensitive to environmental variation, at least for the environments we evaluated.

Hierarchical clustering based on genetic correlations revealed consistent environment groupings across all traits: EME17 and EME18 clustered together, as did HER18 and GAT23. These groupings correspond to management and climatic conditions. EME17 and EME18 were both rainfed trials with higher average temperatures, while HER18 and GAT23 were both

738 irrigated trials with lower average temperatures. This suggests that water availability and
739 temperature regime are the primary drivers of genotype-by-environment interactions in this
740 dataset. Understanding these environmental clusters is crucial for designing multi-environment
741 trial networks and for targeted breeding strategies that account for specific production
742 environments.

743

744 The strong G×E for yield poses both challenges and opportunities. It suggests that breeding for
745 broad adaptation may be difficult, and that environment-specific or region-specific breeding
746 targets may be necessary. However, it also indicates that selection in target environments or
747 representative test sites will be most effective. For traits with weak G×E like flowering time
748 and plant height, selection across environments is more straightforward, and data from any
749 environment can contribute effectively to predictions in other environments.

750

751 Multivariate approaches substantially improve prediction accuracy

752 **Multi-trait models leverage genetic correlations between traits**

753 Multi-trait single-environment models demonstrated that incorporating secondary trait data can
754 substantially improve prediction accuracy for target traits, in specific cases.

755

756 Under the NOSEC scheme, where validation sets had no phenotypic data on any trait, no multi-
757 trait single environment model outperformed single-trait single-environment. Several studies
758 implementing a similar cross-validation scheme reported similar results (Bhatta et al., 2020;
759 Pszczola et al., 2013; Rutkoski et al., 2016; Shahi et al., 2022), suggesting this is a general
760 pattern rather than a mungbean-specific phenomenon. Jia and Jannink (2012) provided

761 theoretical insight into this pattern through simulation studies. They observed that when the
762 primary trait has high heritability, multi-trait models provide negligible improvement
763 regardless of the genetic correlation magnitude. However, when heritability of the primary trait
764 was low (h^2 = 0.10), multi-trait GBLUP improved prediction accuracy even with modest
765 genetic correlations.

766

767 The threshold heritability below which multi-trait models begin to provide benefit under this
768 cross-validation scheme appears to be 0.31. Pszczola et al. (2013) found no improvement from
769 multi-trait analysis when primary trait heritabilities ranged from 0.31 to 0.44. In our study,
770 narrow-sense heritabilities for yield ranged from 0.41 to 0.68 across environments, which are
771 well above this threshold. This explains why multi-trait models provided no advantage under
772 NOSEC, despite substantial genetic correlations between yield and flowering time in some
773 environments (up to $\rho_{YLD,DTF}$ = -0.92 in GAT23). A possible explanation for this pattern is that
774 when the primary trait has high heritability, genetic markers already capture most of the genetic
775 variance, and incorporating a correlated trait adds little additional information. However, it is
776 the opposite when the primary trait has lower heritability (Calus & Veerkamp, 2011; Jia &
777 Jannink, 2012).

778

779 In contrast with NOSEC, under the SEC scheme where secondary traits (DTF and PHT) were
780 recorded in validation sets at the time of prediction, multi-trait single-environment models
781 achieved substantial improvements in yield prediction accuracy of up to +42.6% in GAT23 and
782 +19.6% in HER18. These environments also showed the highest absolute genetic correlations
783 between yield and flowering time (GAT23: $\rho_{YLD,DTF}$ = -0.92±0.03; HER18: $\rho_{YLD,DTF}$ = -
784 0.61±0.10). In contrast, EME17 and EME18, which showed weak or inconsistent genetic

correlations between yield and the secondary traits (YLD-DTF: -0.10±0.24 and -0.39±0.14; YLD-PHT: -0.02±0.28 and -0.08±0.18), gained no benefit from multi-trait modelling. This pattern strongly supports the notion that the value of multi-trait genomic prediction depends critically on the magnitude of genetic correlations between traits (Jia & Jannink, 2012).

Among the multi-trait single-environment model configurations, using DTF as secondary trait consistently outperformed or equalled using PHT, and performed similarly to the full model incorporating both secondary traits. This aligns with the observed genetic correlations that the absolute values of YLD-DTF correlations were consistently higher than YLD-PHT correlations across environments. From a practical breeding perspective, DTF is also easier and less expensive to measure than yield, requiring only observation rather than harvest and processing. This makes multi-trait genomic selection particularly attractive for early-generation selection, where phenotyping resources are limited but secondary traits can be measured inexpensively across large populations.

In summary, as suggested by Rutkoski et al. (2016) and confirmed here, when the primary trait has moderate to high heritability ($h^2 > 0.30$-0.35), including secondary traits in multi-trait models provides negligible benefit when validation sets lack phenotypic data on all traits. The substantial gains observed under the SEC scheme demonstrate that the value of multi-trait models emerges specifically when phenotypic observations on correlated traits are available for selection candidates. Multi-trait genomic prediction should therefore be prioritized for scenarios where secondary traits can be measured inexpensively on selection candidates while the primary trait remains unmeasured due to cost, timing, plot size, or other practical constraints.

Multi-environment models show trait-specific effectiveness

Single-trait multi-environment models revealed striking differences in the value of cross-environment information among traits. For yield, single-trait multi-environment models under Scenario 2 (where validation sets included data from non-target environments) provided only modest improvements over single-trait single-environment, ranging from -0.4% to +8.5%. This limited benefit directly reflects the low genetic correlations between environments for yield. When genotype performance in one environment is poorly predictive of performance in another (ρ = 0.03-0.44), cross-environment data contribute little to prediction accuracy.

In contrast, DTF and PHT showed dramatic improvements under single-trait multi-environment Scenario 2, with gains of +60-100% for DTF and +13-47% for PHT. These large improvements are explained by the high genetic correlations between environments for these traits (ρ = 0.80-0.96). When genotypes maintain consistent relative performance across environments, data from any environment effectively augment the training set for predictions in other environments. This demonstrates a fundamental principle that the value of multi-environment modelling scales directly with the magnitude of genetic correlations between environments (Burgueño et al., 2012; Jarquín et al., 2014).

The differential response of traits to multi-environment modelling has important implications for breeding program design. For traits with high genetic correlations across environments (like DTF and PHT in this study), conducting multi-environment trials and pooling data across sites provides substantial gains in prediction accuracy. For traits with strong G×E (like yield), environment-specific models or more sophisticated G×E modelling approaches such as factor

analytic models implemented here may be necessary to capture genotype-by-environment interaction patterns (Jarquín et al., 2014; Monteverde et al., 2019).

Multi-trait multi-environment models achieve the highest accuracy by combining information sources

Multi-trait multi-environment models, which simultaneously incorporate both secondary trait information and cross-environment data, achieved the highest prediction accuracies across all approaches tested. The factorial combination of schemes (NOSEC/SEC × Scenario 1/Scenario 2) revealed an additive pattern in the value of different data sources. For yield prediction, the hierarchy of model performance was SEC-Scenario 2 > SEC-Scenario 1 ≈ SEC > NOSEC-Scenario 2 > Scenario 2 >> NOSEC ≈ Scenario 1 ≈ NOSEC-Scenario 1.

Under SEC-Scenario 2, where validation sets had both the secondary trait data in the target environment and yield data from non-target environments, multi-trait multi-environment models achieved the highest gains of up to +52.7% in GAT23, +24.3% in HER18, and +6.4% in EME18. These improvements substantially exceeded those from either secondary trait alone (multi-trait single-environment under SEC: up to +42.6%) or cross-environment data alone (single-trait multi-environment under Scenario 2: up to +8.5%). The additive nature of these gains suggests that secondary trait correlations and cross-environment correlations contribute relatively independent information to genomic prediction.

Interestingly, multi-trait multi-environment models under NOSEC-Scenario 2 (cross-environment data but no secondary traits) achieved substantial gains (+21.0% to +33.6% in

GAT23), exceeding single-trait multi-environment under Scenario 2 despite both having access to the same cross-environment yield data. This indicates that even when secondary traits are not observed in validation sets, incorporating them into the model structure captures additional genetic correlations that improve prediction. The multi-trait framework leverages correlated responses and shared genetic architecture even when selection candidates lack phenotypes for all traits (Jia & Jannink, 2012).

The environment-specific pattern of improvement remained consistent across all multivariate approaches, that is GAT23 and HER18 consistently showed large gains, while EME17 and EME18 showed minimal to no improvement. This pattern reflects the underlying genetic correlation structure. GAT23 showed strong genetic correlations both between traits (YLD-DTF: ρ = -0.92) and with other environments (particularly HER18), while EME17 and EME18 showed weak correlations on both dimensions. This demonstrates that the effectiveness of multivariate genomic prediction depends on the genetic correlation network encompassing both traits and environments. When a target environment or trait is poorly correlated with available auxiliary information, multivariate approaches provide limited benefit.

From a practical standpoint, these results provide clear guidance for breeding programs. When phenotypic resources are limited, the priority should be: (1) measure inexpensive secondary traits on all selection candidates, (2) conduct multi-environment trials for traits with high cross-environment correlations, and (3) integrate both strategies through multi-trait multi-environment models when feasible. The substantial gains demonstrated here (up to +52.7% over baseline models) can translate to significantly accelerated genetic gain, particularly when combined with reduced generation intervals.

Genetic relatedness between training and validation sets is critical for prediction accuracy

The comparison between cluster-based and random fold assignment revealed that genetic relatedness between training and validation sets has a larger impact on prediction accuracy than training set size. When validation sets comprised entire genetic clusters, prediction accuracy decreased by an average of 0.20 compared to random assignments. This substantial reduction demonstrates that genomic prediction depends critically on having genetically related individuals in the training set (Daetwyler et al., 2012; Habier et al., 2007).

Random assignment distributed related individuals across training and validation sets, maintaining genetic connectivity between them. In contrast, cluster-based assignment isolated genetically homogeneous groups within validation sets, reducing their relatedness to training sets. The cluster-specific variation in prediction performance provides further insight into this relationship. Despite both having the smallest cluster sizes (n=42 and n=48, respectively) and thus the largest training sets when used as validation sets, Clusters 1 and 5 showed opposite prediction performance: Cluster 1 had the lowest prediction accuracy for yield, while Cluster 5 had the highest.

This divergence aligns perfectly with their genetic differentiation patterns. Cluster 1's high mean F_{ST} (0.34) indicates substantial genetic divergence from all other clusters, meaning its training set contained relatively few closely related individuals. Conversely, Cluster 5's lower mean F_{ST} (0.26), despite also having low within-cluster diversity (H_e = 0.17), reflects greater

genetic similarity to other clusters, providing more related individuals in the training set. These results demonstrate that genetic relatedness between training and validation sets outweighs training set size in determining prediction accuracy which is consistent with theoretical and empirical studies showing that prediction accuracy depends on kinship rather than just sample size (Daetwyler et al., 2012; Habier et al., 2007).

The concentration of Australian commercial varieties and breeding lines in Cluster 5 has important practical implications. Because Cluster 5 is genetically closer to other clusters in the panel (lower F_{ST}), genomic predictions for these elite materials benefit from the diverse genetic relationships available in the full panel. This suggests that the Mini-Core Plus Panel would serve effectively as a training population for genomic selection in Australian breeding programs. However, the poor prediction performance for Cluster 1, which represents genetically distinct germplasm, highlights the limitation that genomic prediction accuracy will be lower for selection candidates that are genetically distant from the training population.

These findings demonstrate that genomic selection can be successfully implemented even with genetically diverse germplasm, provided that training and validation sets maintain sufficient genetic relatedness. The challenge, therefore, is not diversity per se but ensuring adequate representation of target germplasm in training populations. Several strategies can address this challenge. First, training populations can be strategically designed to maximize genetic relationships with breeding program targets (Akdemir & Isidro-Sánchez, 2019; Rincent et al., 2012). Second, ongoing phenotyping of selection candidates can continually update and expand training sets to maintain genetic connectivity. Third, for highly diverse panels, subset selection

algorithms can identify optimal training sets that maximize prediction accuracy for specific validation targets (Akdemir & Isidro-Sánchez, 2019).

Marker density requirements are modest due to rapid LD decay

The number of markers required for effective genomic prediction is determined by the extent of linkage disequilibrium in the population. Calus et al. (2008) demonstrated that for accurate genomic prediction, adjacent markers should exhibit $r^2 \geq 0.2$. In the Mini-Core Plus Panel, LD (r^2) reaches 0.2 at approximately 174 kbp (based on the LD_{20} threshold). Given the mungbean genome size of approximately 475-500 Mbp distributed across 11 chromosomes, this LD decay rate implies that approximately 2,700-2,900 evenly spaced SNP markers would be needed to ensure adequate marker-QTL linkage across the genome. The full marker set of 3,310 SNPs used in this study therefore exceeded this theoretical minimum, providing good genome coverage.

However, marker density requirements may evolve as breeding programs progress. The current evaluation assessed prediction within the existing population structure, where LD patterns reflect historical recombination. In ongoing breeding programs, LD will decay over generations as selection advances and crosses generate new recombination events. Goddard et al. (2009) demonstrated that prediction accuracy can decline over generations as LD between markers and QTL breaks down, requiring periodic retraining or potentially higher marker densities to maintain long-term prediction accuracy. Additionally, if breeding objectives shift toward fine-mapping or marker-assisted selection for specific QTL, higher marker densities would be necessary to ensure markers are in tight linkage with causal variants. Nevertheless, for the

immediate implementation of genomic selection in mungbean, our results indicate that the current marker densities are sufficient.

Utilizing Mini-Core collection in breeding programs

Collections of diverse germplasm such as the Mini-Core panel are valuable sources of novel genetic variations. However, it is not straightforward to utilize them because often they exhibit substantial performance gap compared to elite germplasm and may not be well-adapted to the target production environments (Hölker et al., 2022; Kannenberg & Falk, 1995). Allier et al. (2020) addressed these challenges by demonstrating that bridging strategies, combined with genomic prediction models trained on both elite and diverse materials, can effectively integrate diversity sources into elite breeding programs while maintaining genetic gain. Specifically, their results showed that using a single training set combining elite and bridging progeny improved prediction accuracy for introduction crosses without compromising predictions of elite family, thus optimizing the conversion of novel genetic diversity into breeding progress.

Photoperiod sensitivity has been reported in mungbean landraces and wild mungbean relatives (Imrie & Lawn, 1990; Isemura et al., 2012; Kajonphol et al., 2012), although modern cultivars that are photoperiod insensitive have been developed by breeders (Somta et al., 2022). While we could not directly evaluate photosensitivity in our dataset, it is possible that the high genetic correlation between yield and flowering time is partly driven by photoperiod response and consequently temperature response. Currently, there is limited research available looking into how genetics of flowering time relates to yield in cultivated mungbean. However, in a study involving wild mungbean (*Vigna radiata ssp. Sublobata*), photoperiod extension indirectly

caused reduction in yield as the flowering period got delayed when temperature was lower (Rebetzke & Lawn, 2006), suggesting that adaptation to local photoperiod conditions can substantially shape yield performance. The strong negative correlation between yield and days to flowering observed in our dataset, specifically from HER18 and GAT23 trials, likely reflects the influence of photoperiod adaptation across the diverse accessions included. HER18 and GAT23 had lower average temperature than EME17 and EME18. Despite these limitations, our results present an opportunity for the utilization of Mini-Core in the commercial breeding program. The inclusion of released cultivars within the Mini-Core Plus dataset provides a direct genetic bridge to elite germplasm. By training genomic prediction models across both elite and diverse materials, leveraging the connectivity provided by these shared cultivars, as demonstrated by Allier et al. (2020), it may be possible to more effectively identify Mini-Core accessions harbouring favourable alleles for integration into commercial breeding programs. Nevertheless, this requires further investigation and validation in the actual breeding program.

Conclusion

This study demonstrates that genomic selection is feasible in mungbean, with moderate to high prediction accuracies achieved. Single-trait single-environment GBLUP models provided baseline prediction accuracies of 0.34-0.63, sufficient for accelerating genetic gain in breeding programs. Multivariate approaches, which incorporate either secondary traits (multi-trait models), cross-environment data (multi-environment models), or both (multi-trait multi-environment models), substantially improved prediction accuracy, with gains up to +52.7% over baseline models.

The effectiveness of multivariate genomic prediction depended critically on the genetic correlation structure among environments, and primary and secondary traits. Multi-trait models provided large gains only when genetic correlations between traits were strong, while multi-environment models succeeded only when genetic correlations between environments were high. Yield showed strong genotype-by-environment interactions, limiting the benefit of multi-environment approaches, while flowering time and plant height showed weak genotype-by-environment interactions and gained substantially from cross-environment prediction. These patterns provide clear guidance for choosing modelling approaches based on trait and environment characteristics.

Genetic relatedness between training and validation sets emerged as a critical determinant of prediction accuracy, outweighing the effect of training set size. Modest marker densities proved sufficient for genomic prediction in this diverse panel, making implementation economically feasible. Despite the inherent challenges of applying genomic prediction across diverse and elite germplasm, the inclusion of released cultivars within the Mini-Core panel provides a critical genetic bridge that could facilitate the identification of favourable alleles from unadapted accessions for integration into commercial mungbean breeding programs.

These findings establish a foundation for implementing genomic selection in mungbean breeding programs. By adopting multivariate modelling approaches tailored to specific traits and environments, maintaining genetic connectivity between training and selection populations, and using cost-effective genotyping strategies, breeding programs can substantially accelerate genetic gain for this important legume crop.

References

Akdemir, D., & Isidro-Sánchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. *Sci Rep*, *9*(1), 1446. https://doi.org/10.1038/s41598-018-38081-6

Allier, A., Teyssèdre, S., Lehermeier, C., Moreau, L., & Charcosset, A. (2020). Optimized breeding strategies to harness genetic resources with different performance levels. *BMC Genomics*, *21*(1), 349. https://doi.org/10.1186/s12864-020-6756-0

Australian Mungbean Association. (2022). *Australian Mungbean Industry Strategic Plan 2020–2025*

Bassi, F. M., Bentley, A. R., Charmet, G., Ortiz, R., & Crossa, J. (2016). Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.). *Plant Science*, *242*, 23-36. https://doi.org/https://doi.org/10.1016/j.plantsci.2015.08.021

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Science*, *48*(5), 1649-1664. https://doi.org/https://doi.org/10.2135/cropsci2008.03.0131

Bernardo, R. (2016). Bandwagons I, too, have known. *Theoretical and Applied Genetics*, *129*(12), 2323-2332. https://doi.org/10.1007/s00122-016-2772-5

Bernardo, R., & Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, *47*(3), 1082-1090. https://doi.org/https://doi.org/10.2135/cropsci2006.11.0690

Beyene, Y., Gowda, M., Olsen, M., Robbins, K. R., Pérez-Rodríguez, P., Alvarado, G., Dreher, K., Gao, S. Y., Mugo, S., Prasanna, B. M., & Crossa, J. (2019). Empirical comparison of tropical maize hybrids selected through genomic and phenotypic selections [Original Research]. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.01502

Beyene, Y., Semagn, K., Mugo, S., Tarekegne, A., Babu, R., Meisel, B., Sehabiague, P., Makumbi, D., Magorokosho, C., Oikeh, S., Gakunga, J., Vargas, M., Olsen, M., Prasanna, B. M., Banziger, M., & Crossa, J. (2015). Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. *Crop Science*, *55*(1), 154-163. https://doi.org/https://doi.org/10.2135/cropsci2014.07.0460

Bhatta, M., Gutierrez, L., Cammarota, L., Cardozo, F., Germán, S., Gómez-Guerrero, B., Pardo, M. F., Lanaro, V., Sayas, M., & Castro, A. J. (2020). Multi-trait genomic prediction model increased the predictive ability for agronomic and malting quality traits in barley (Hordeum vulgare L.). *G3 Genes|Genomes|Genetics*, *10*(3), 1113-1124. https://doi.org/10.1534/g3.119.400968

Burgueño, J., de los Campos, G., Weigel, K., & Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Science*, *52*(2), 707-719. https://doi.org/https://doi.org/10.2135/cropsci2011.06.0299

Calus, M. P., & Veerkamp, R. F. (2011). Accuracy of multi-trait genomic selection using different methods. *Genet Sel Evol*, *43*(1), 26. https://doi.org/10.1186/1297-9686-43-26

Covarrubias-Pazaran, G. (2016). Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. *PLOS ONE*, *11*(6), e0156744. https://doi.org/10.1371/journal.pone.0156744

Covarrubias-Pazaran, G., Schlautman, B., Diaz-Garcia, L., Grygleski, E., Polashock, J., Johnson-Cicalese, J., Vorsa, N., Iorizzo, M., & Zalapa, J. (2018). Multivariate GBLUP improves accuracy of genomic selection for yield and fruit weight in biparental populations of Vaccinium macrocarpon Ait [Original Research]. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.01310

Cowling, W. A., Castro-Urrea, F. A., Stefanova, K. T., Li, L., Banks, R. G., Saradadevi, R., Sass, O., Kinghorn, B. P., & Siddique, K. H. M. (2023). Optimal Contribution Selection Improves the Rate of Genetic Gain in Grain Yield and Yield Stability in Spring Canola in Australia and Canada. *Plants (Basel)*, *12*(2). https://doi.org/10.3390/plants12020383

Crain, J., Mondal, S., Rutkoski, J., Singh, R. P., & Poland, J. (2018). Combining High-Throughput Phenotyping and Genomic Information to Increase Prediction and Selection Accuracy in Wheat Breeding. *The Plant Genome*, *11*(1), 170043. https://doi.org/https://doi.org/10.3835/plantgenome2017.05.0043

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J., M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., & Varshney, R. K. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in Plant Science*, *22*(11), 961-975. https://doi.org/https://doi.org/10.1016/j.tplants.2017.08.011

Cullis, B. R., Smith, A. B., & Coombes, N. E. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics*, *11*(4), 381-393. https://doi.org/10.1198/108571106X154443

Daetwyler, H. D., Kemper, K. E., van der Werf, J. H. J., & Hayes, B. J. (2012). Components of the accuracy of genomic prediction in a multi-breed sheep population1. *Journal of Animal Science*, *90*(10), 3375-3384. https://doi.org/10.2527/jas.2011-4557

Dahiya, P. K., Linnemann, A. R., Van Boekel, M. A. J. S., Khetarpaul, N., Grewal, R. B., & Nout, M. J. R. (2015). Mung Bean: technological and nutritional potential. *Crit Rev Food Sci Nutr*, *55*(5), 670-688. https://doi.org/10.1080/10408398.2012.671202

Dudley, C., Van Haeften, S., Alahmad, S., Dinglasan, E., Hickey, L. T., Robinson, H., Beveridge, C. A., Udvardi, M., Noble, T., Massel, K., Dun, E. A., & Smith, M. R. (2025). Time to flowering and flowering duration in mungbean are unrelated physiological traits with independent genetic controls. *J Exp Bot*, *76*(18), 5528-5541. https://doi.org/10.1093/jxb/eraf222

Endelman, J. B., Atlin, G. N., Beyene, Y., Semagn, K., Zhang, X., Sorrells, M. E., & Jannink, J.-L. (2014). Optimal design of preliminary yield trials with genome-wide markers. *Crop Science*, *54*(1), 48-59. https://doi.org/https://doi.org/10.2135/cropsci2013.03.0154

Foyer, C. H., Lam, H.-M., Nguyen, H. T., Siddique, K. H. M., Varshney, R. K., Colmer, T. D., Cowling, W., Bramley, H., Mori, T. A., Hodgson, J. M., Cooper, J. W., Miller, A. J., Kunert, K., Vorster, J., Cullis, C., Ozga, J. A., Wahlqvist, M. L., Liang, Y., Shou, H.,…Considine, M. J. (2016). Neglecting legumes has compromised human health and sustainable food production. *Nature Plants*, *2*(8), 16112. https://doi.org/10.1038/nplants.2016.112

Fumia, N., Nair, R., Lin, Y.-P., Lee, C.-R., Chen, H.-W., von Wettberg, E. B., Kantar, M., & Schafleitner, R. (2023). Leveraging genomics and phenomics to accelerate improvement in mungbean: A case study in how to go from GWAS to selection. *The Plant Phenome Journal*, *6*(1), e20088. https://doi.org/https://doi.org/10.1002/ppj2.20088

Gill, H. S., Brar, N., Halder, J., Hall, C., Seabourn, B. W., Chen, Y. R., St. Amand, P., Bernardo, A., Bai, G., Glover, K., Turnipseed, B., & Sehgal, S. K. (2023). Multi-trait genomic selection improves the prediction accuracy of end-use quality traits in hard winter wheat. *The Plant Genome*, *16*(4), e20331. https://doi.org/https://doi.org/10.1002/tpg2.20331

Gill, H. S., Halder, J., Zhang, J., Brar, N. K., Rai, T. S., Hall, C., Bernardo, A., Amand, P. S., Bai, G., Olson, E., Ali, S., Turnipseed, B., & Sehgal, S. K. (2021). Multi-Trait Multi-Environment Genomic Prediction of Agronomic Traits in Advanced Breeding Lines of Winter Wheat [Original Research]. *Frontiers in Plant Science*, *Volume 12 - 2021*. https://doi.org/10.3389/fpls.2021.709545

Gilmour, A. R., Gogel, B. J., Cullis, B. R., & Thompson, R. (2009). ASReml user guide release 3.0. *VSN International Ltd, Hemel Hempstead, HP1 1ES, UK*. www.vsni.co.uk

Goddard, M. E., Hayes, B. J., & Visscher, P. M. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research*, *91*(1), 47-60. https://doi.org/10.1017/S0016672308009981

Goudet, J. (2005). hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, *5*(1), 184-186. https://doi.org/https://doi.org/10.1111/j.1471-8286.2004.00828.x

Granwal, L. (2024). *Yield of soybeans in Australia from financial year 2010 to 2024*. Statista. Retrieved 18/03/2024 from https://www.statista.com/statistics/631890/australia-soybean-yield/#:~:text=In%20the%20financial%20year%202023,tons%20per%20hectare%20in%202024.

Ha, J., Satyawan, D., Jeong, H., Lee, E., Cho, K. H., Kim, M. Y., & Lee, S. H. (2021). A near-complete genome sequence of mungbean (Vigna radiata L.) provides key insights into the modern breeding program. *Plant Genome*, *14*(3), e20121. https://doi.org/10.1002/tpg2.20121

Habier, D., Fernando, R. L., & Dekkers, J. C. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, *177*(4), 2389-2397. https://doi.org/10.1534/genetics.107.081190

Haile, J. K., N'Diaye, A., Clarke, F., Clarke, J., Knox, R., Rutkoski, J., Bassi, F. M., & Pozniak, C. J. (2018). Genomic selection for grain yield and quality traits in durum wheat. *Molecular Breeding*, *38*(6), 75. https://doi.org/10.1007/s11032-018-0818-x

Hayes, B. J., Panozzo, J., Walker, C. K., Choy, A. L., Kant, S., Wong, D., Tibbits, J., Daetwyler, H. D., Rochfort, S., Hayden, M. J., & Spangenberg, G. C. (2017). Accelerating wheat breeding for end-use quality with multi-trait genomic predictions incorporating near infrared and nuclear magnetic resonance-derived phenotypes. *Theor Appl Genet*, *130*(12), 2505-2519. https://doi.org/10.1007/s00122-017-2972-7

Heffner, E. L., Jannink, J.-L., & Sorrells, M. E. (2011). Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. *The Plant Genome*, *4*(1). https://doi.org/https://doi.org/10.3835/plantgenome2010.12.0029

Hölker, A. C., Mayer, M., Presterl, T., Bauer, E., Ouzunova, M., Melchinger, A. E., & Schön, C. C. (2022). Theoretical and experimental assessment of genome-based prediction in landraces of allogamous crops. *Proc Natl Acad Sci U S A*, *119*(18), e2121797119. https://doi.org/10.1073/pnas.2121797119

Hunt, C. H., Hayes, B. J., van Eeuwijk, F. A., Mace, E. S., & Jordan, D. R. (2020). Multi-environment analysis of sorghum breeding trials using additive and dominance genomic relationships. *Theor Appl Genet*, *133*(3), 1009-1018. https://doi.org/10.1007/s00122-019-03526-7

Imrie, B. C., & Lawn, R. J. (1990). Time to Flowering of Mung Bean (Vigna radiata) Genotypes and their Hybrids in Response to Photoperiod and Temperature. *Experimental Agriculture*, *26*(3), 307-318. https://doi.org/10.1017/S0014479700018470

Isemura, T., Kaga, A., Tabata, S., Somta, P., Srinives, P., Shimizu, T., Jo, U., Vaughan, D. A., & Tomooka, N. (2012). Construction of a Genetic Linkage Map and Genetic Analysis of Domestication Related Traits in Mungbean (Vigna radiata). *PLOS ONE*, *7*(8), e41304. https://doi.org/10.1371/journal.pone.0041304

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez, P., Calus, M., Burgueño, J., & de los Campos, G. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, *127*(3), 595-607. https://doi.org/10.1007/s00122-013-2243-1

Jia, Y., & Jannink, J. L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, *192*(4), 1513-1522. https://doi.org/10.1534/genetics.112.144246

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403-1405. https://doi.org/10.1093/bioinformatics/btn129

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(1), 94. https://doi.org/10.1186/1471-2156-11-94

Kajonphol, T., Sangsiri, C., Somta, P., Toojinda, T., & Srinives, P. (2012). SSR map construction and quantitative trait loci (QTL) identification of major agronomic traits in mungbean (Vigna radiata (L.) Wilczek). *SABRAO Journal of Breeding & Genetics*, *44*(1).

Kang, Y. J., Kim, S. K., Kim, M. Y., Lestari, P., Kim, K. H., Ha, B.-K., Jun, T. H., Hwang, W. J., Lee, T., Lee, J., Shim, S., Yoon, M. Y., Jang, Y. E., Han, K. S., Taeprayoon, P., Yoon, N., Somta, P., Tanya, P., Kim, K. S.,…Lee, S.-H. (2014). Genome sequence of mungbean and insights into evolution within Vigna species. *Nature Communications*, *5*(1), 5443. https://doi.org/10.1038/ncomms6443

Kannenberg, L. W., & Falk, D. E. (1995). Models for activation of plant genetic resources for crop breeding programs. *Canadian Journal of Plant Science*, *75*(1), 45-53. https://doi.org/10.4141/cjps95-008

Keller, B., Ariza-Suarez, D., de la Hoz, J., Aparicio, J. S., Portilla-Benavides, A. E., Buendia, H. F., Mayor, V. M., Studer, B., & Raatz, B. (2020). Genomic Prediction of Agronomic Traits in Common Bean (Phaseolus vulgaris L.) Under Environmental Stress [Original Research]. *Frontiers in Plant Science*, *Volume 11 - 2020*. https://doi.org/10.3389/fpls.2020.01001

Krause, M. R., González-Pérez, L., Crossa, J., Pérez-Rodríguez, P., Montesinos-López, O., Singh, R. P., Dreisigacker, S., Poland, J., Rutkoski, J., Sorrells, M., Gore, M. A., & Mondal, S. (2019). Hyperspectral Reflectance-Derived Relationship Matrices for Genomic Prediction of Grain Yield in Wheat. *G3 Genes|Genomes|Genetics*, *9*(4), 1231-1247. https://doi.org/10.1534/g3.118.200856

Lado, B., Vázquez, D., Quincke, M., Silva, P., Aguilar, I., & Gutiérrez, L. (2018). Resource allocation optimization with multi-trait genomic prediction for bread wheat (Triticum aestivum L.) baking quality. *Theoretical and Applied Genetics*, *131*(12), 2719-2731. https://doi.org/10.1007/s00122-018-3186-3

Langemeier, M., & Zhou, L. (2022). *International Benchmarks For Soybean Production (2022)*. Retrieved 18/03/2024 from https://ag.purdue.edu/commercialag/home/resource/2022/03/international-benchmarks-for-soybean-production-2022/

Liu, C., Wang, Y., Peng, J., Fan, B., Xu, D., Wu, J., Cao, Z., Gao, Y., Wang, X., Li, S., Su, Q., Zhang, Z., Wang, S., Wu, X., Shang, Q., Shi, H., Shen, Y., Wang, B., & Tian, J. (2022). High-quality genome assembly and pan-genome studies facilitate genetic discovery in mung bean and its improvement. *Plant Communications*, *3*(6), 100352. https://doi.org/https://doi.org/10.1016/j.xplc.2022.100352

Lorenz, A., & Nice, L. (2017). Training Population Design and Resource Allocation for Genomic Selection in Plant Breeding. In R. K. Varshney, M. Roorkiwal, & M. E. Sorrells (Eds.), *Genomic Selection for Crop Improvement: New Molecular Breeding Strategies for Crop Improvement* (pp. 7-22). Springer International Publishing. https://doi.org/10.1007/978-3-319-63170-7_2

Massman, J. M., Jung, H.-J. G., & Bernardo, R. (2013). Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. *Crop Science*, *53*(1), 58-66. https://doi.org/https://doi.org/10.2135/cropsci2012.02.0112

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819-1829. https://doi.org/10.1093/genetics/157.4.1819

Michel, S., Kummer, C., Gallee, M., Hellinger, J., Ametz, C., Akgöl, B., Epure, D., Güngör, H., Löschenberger, F., & Buerstmayr, H. (2018). Improving the baking quality of bread wheat by genomic selection in early generations. *Theor Appl Genet*, *131*(2), 477-493. https://doi.org/10.1007/s00122-017-2998-x

Montesinos-López, A., Montesinos-López, O. A., Crossa, J., Burgueño, J., Eskridge, K. M., Falconi-Castillo, E., He, X., Singh, P., & Cichy, K. (2016). Genomic bayesian prediction model for count data with genotype × environment interaction. *G3 Genes|Genomes|Genetics*, *6*(5), 1165-1177. https://doi.org/10.1534/g3.116.028118

Montesinos-López, O. A., Montesinos-López, A., Tuberosa, R., Maccaferri, M., Sciara, G., Ammar, K., & Crossa, J. (2019). Multi-trait, multi-environment genomic prediction of durum wheat with genomic best linear unbiased predictor and deep learning methods [Original Research]. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.01311

Monteverde, E., Gutierrez, L., Blanco, P., Pérez de Vida, F., Rosas, J. E., Bonnecarrère, V., Quero, G., & McCouch, S. (2019). Integrating Molecular Markers and Environmental Covariates To Interpret Genotype by Environment Interaction in Rice (Oryza sativa L.) Grown in Subtropical Areas. *G3 Genes|Genomes|Genetics*, *9*(5), 1519-1531. https://doi.org/10.1534/g3.119.400064

Nair, R., & Schreinemachers, P. (2020). Global status and economic importance of mungbean. In R. M. Nair, R. Schafleitner, & S.-H. Lee (Eds.), *The Mungbean Genome* (pp. 1-8). Springer International Publishing. https://doi.org/10.1007/978-3-030-20008-4_1

Nair, R. M., Schafleitner, R., Kenyon, L., Srinivasan, R., Easdown, W., Ebert, A. W., & Hanson, P. (2012). Genetic improvement of mungbean [Staff Publications]. *44(2):177-190*, Article AVRDC Staff Publication. (SABRAO JOURNAL OF BREEDING AND GENETICS)

Neyhart, J. L., Lorenz, A. J., & Smith, K. P. (2019). Multi-trait improvement by predicting genetic correlations in breeding crosses. *G3 Genes|Genomes|Genetics*, *9*(10), 3153-3165. https://doi.org/10.1534/g3.119.400406

Noble, T. J., Tao, Y., Mace, E. S., Williams, B., Jordan, D. R., Douglas, C. A., & Mundree, S. G. (2018). Characterization of linkage disequilibrium and population structure in a mungbean diversity panel [Original Research]. *Frontiers in Plant Science*, *8*. https://doi.org/10.3389/fpls.2017.02102

Pszczola, M., Veerkamp, R. F., de Haas, Y., Wall, E., Strabel, T., & Calus, M. P. (2013). Effect of predictor traits on accuracy of genomic breeding values for feed intake based on a limited cow reference population. *Animal*, *7*(11), 1759-1768. https://doi.org/https://doi.org/10.1017/S175173111300150X

R Core Team. (2024). *R: A language and environment for statistical computing*. In R Foundation for Statistical Computing. https://www.R-project.org/

Rebetzke, G. J., & Lawn, R. J. (2006). Adaptive responses of wild mungbean (Vigna radiata ssp. sublobata) to photo-thermal environment. II.* Growth, biomass, and seed yield. *Australian Journal of Agricultural Research*, *57*(8), 929-937. https://doi.org/10.1071/ar05377

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., Rodríguez, V. M., Moreno-Gonzalez, J., Melchinger, A., Bauer, E., Schoen, C. C., Meyer, N., Giauffret, C., Bauland, C., Jamin, P., Laborde, J., Monod, H., Flament, P., Charcosset, A., & Moreau, L. (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). *Genetics*, *192*(2), 715-728. https://doi.org/10.1534/genetics.112.141473

Roorkiwal, M., Jarquin, D., Singh, M. K., Gaur, P. M., Bharadwaj, C., Rathore, A., Howard, R., Srinivasan, S., Jain, A., Garg, V., Kale, S., Chitikineni, A., Tripathi, S., Jones, E., Robbins, K. R., Crossa, J., & Varshney, R. K. (2018). Genomic-enabled prediction models using multi-environment trials to estimate the effect of genotype × environment interaction on prediction accuracy in chickpea. *Scientific Reports*, *8*(1), 11701. https://doi.org/10.1038/s41598-018-30027-2

Rutkoski, J., Poland, J., Mondal, S., Autrique, E., Pérez, L. G., Crossa, J., Reynolds, M., & Singh, R. (2016). Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. *G3 (Bethesda)*, *6*(9), 2799-2808. https://doi.org/10.1534/g3.116.032888

Saludares, R. A., Atanda, S. A., Piche, L., Worral, H., Dariva, F., McPhee, K., & Bandillo, N. (2024). Multi-trait multi-environment genomic prediction of preliminary yield trial in pulse crop. *The Plant Genome*, *17*(3), e20496. https://doi.org/https://doi.org/10.1002/tpg2.20496

Schafleitner, R., Nair, R. M., Rathore, A., Wang, Y.-w., Lin, C.-y., Chu, S.-h., Lin, P.-y., Chang, J.-C., & Ebert, A. W. (2015). The AVRDC – The World Vegetable Center mungbean (Vigna radiata) core and mini core collections. *BMC Genomics*, *16*(1), 344. https://doi.org/10.1186/s12864-015-1556-7

Shahi, D., Guo, J., Pradhan, S., Khan, J., Avci, M., Khan, N., McBreen, J., Bai, G., Reynolds, M., Foulkes, J., & Babar, M. A. (2022). Multi-trait genomic prediction using in-season physiological parameters increases prediction accuracy of complex traits in US wheat. *BMC Genomics*, *23*(1), 298. https://doi.org/10.1186/s12864-022-08487-8

Shanmugasundaram, S., Keatinge, J. D. H., & Hughes, J. d. A. (2009). The mungbean transformation, diversifying crops, defeating malnutrition. *International Food Policy Research Institute*, *00922*, 43. https://www.ifpri.org/publication/mungbean-transformation-diversifying-crops-defeating-malnutrition

Smith, A., Cullis, B., & Gilmour, A. (2001). Applications: The Analysis of Crop Variety Evaluation Data in Australia. *Australian & New Zealand Journal of Statistics*, *43*(2), 129-145. https://doi.org/https://doi.org/10.1111/1467-842X.00163

Smith, A., Cullis, B., & Thompson, R. (2001). Analyzing Variety by Environment Data Using Multiplicative Mixed Models and Adjustments for Spatial Field Trend. *Biometrics*, *57*(4), 1138-1147. https://doi.org/https://doi.org/10.1111/j.0006-341X.2001.01138.x

Smith, A. B., & Cullis, B. R. (2018). Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica*, *214*(8), 143. https://doi.org/10.1007/s10681-018-2220-5

Smith, A. B., Ganesalingam, A., Kuchel, H., & Cullis, B. R. (2015). Factor analytic mixed models for the provision of grower information from national crop variety testing programs. *Theoretical and Applied Genetics*, *128*(1), 55-72. https://doi.org/10.1007/s00122-014-2412-x

Somta, P., Laosatit, K., Yuan, X., & Chen, X. (2022). Thirty Years of Mungbean Genome Research: Where Do We Stand and What Have We Learned? *Front Plant Sci*, *13*, 944721. https://doi.org/10.3389/fpls.2022.944721

Sweeney, D. W., Rooney, T. E., & Sorrells, M. E. (2021). Gain from genomic selection for a selection index in two-row spring barley. *The Plant Genome*, *14*(3), e20138. https://doi.org/https://doi.org/10.1002/tpg2.20138

Tolhurst, D. J., Gaynor, R. C., Gardunia, B., Hickey, J. M., & Gorjanc, G. (2022). Genomic selection using random regressions on known and latent environmental covariates. *Theoretical and Applied Genetics*, *135*(10), 3393-3415. https://doi.org/10.1007/s00122-022-04186-w

Watson, A., Hickey, L. T., Christopher, J., Rutkoski, J., Poland, J., & Hayes, B. J. (2019). Multivariate genomic selection and potential of rapid indirect selection with speed rreeding in spring wheat. *Crop Science*, *59*(5), 1945-1959. https://doi.org/https://doi.org/10.2135/cropsci2018.12.0757

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the analysis of population structure. *Evolution*, *38*(6), 1358-1370. https://doi.org/10.2307/2408641

Statements and Declarations

Funding

This research was supported by the International Mungbean Improvement Network (ACIAR; CIM2014/079). MEF is supported by CSIRO AI for Missions Industry PhD Scholarship.

Conflict of Interests

All the authors declare no conflict of interest.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.docx
- Table2.docx
- Table3.docx
- MungbeanGPManuscriptTAGSubmissionFormat170326Supp.docx