

Targeted genotyping (90K-SPET) facilitates genome-wide association studies and the prediction of yield-related traits in faba bean (*Vicia faba* L.)

Antonio Lippolis

Wageningen University & Research

Salvador A. Gezan

VSN International Ltd

Jorrit Zuidgeest

Wageningen University & Research

Valeria Cafaro

University of Catania

Bert-Jan Dinter

Vandinter Semo B.V

Geert Elzes

Vandinter Semo B.V

Maria-João Paulo

Wageningen University & Research

Luisa M. Trindade

`luisa.trindade@wur.nl`

Wageningen University & Research

Research Article

Keywords: Faba bean, GWAS, Genomic Selection, Yield, GxE, Genotyping, Genetic diversity

Posted Date: August 13th, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-4687335/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Plant Biology on April 29th, 2025. See the published version at <https://doi.org/10.1186/s12870-025-06546-2>.

Abstract

Background

Establishing faba bean (*Vicia faba* L.) as a major protein crop in Europe requires developing high-yielding varieties. However, the genetic regulation of yield-related traits is currently under-explored. These traits can be improved by exploiting the extensive but largely uncharacterized faba bean germplasm. Our study aimed to identify associations between 38,014 single nucleotide polymorphisms (SNPs) and flowering time (FT), plant height (PH), pod length (PL), seeds per pod (SP), and single seed weight (SSW) in 245 faba bean accessions (CGN population) via a Genome-Wide Association Study (GWAS). The accessions were grown in 2021 and 2022 in the Netherlands. Additionally, we developed genomic selection (GS) models to predict the genetic merit within large germplasm collections.

Results

The CGN population was an optimal panel for performing high-resolution GWAS, showing large phenotypic variation, high narrow-sense heritability for all traits, and minimal genetic relatedness among accessions. Population structure analysis revealed the presence of four genetic groups. GWAS uncovered 33 SNP-trait associations in 2021 and 17 in 2022. We identified one stable QTL for FT and four for SSW over the two years, representing key molecular markers for testing in breeding applications. Short linkage disequilibrium decay (~268 Kbp) facilitated the identification of several important candidate genes with interesting homologs in other crops. Ten SNPs in 2021 and five in 2022 were predicted to be intra-genic missense variants, potentially altering protein function. Moreover, modeling the SNP effect simultaneously via Bayesian GS showed promising predictive ability (PA) and prediction accuracy (ACC), supporting their potential application in germplasm-improvement programs. Predictive ability ranged from 0.58 to 0.81 in 2021, and 0.47 to 0.85 in 2022 for different traits. Additionally, across-year predictions showed stable PA.

Conclusion

GWAS revealed promising QTLs for use in molecular breeding and highlighted new candidate genes. Interestingly, the prediction of intra-genic SNPs categorized 15 SNPs as putatively affecting protein function. Moreover, we demonstrated for the first time in faba bean that GS has the potential to unlock untapped diversity in genebank collections and accelerate trait integration into faba bean breeding programs.

1 Introduction

Boosting legume production is a key step toward advancing the "protein transition" and a more sustainable food system in Europe. The protein transition is the shift from animal- to plant-based proteins to ensure a healthier and more environmentally friendly diet. Animal-based diets strongly contribute to greenhouse gas emissions (1, 2). Moving to a plant-based diet can reduce the food's GHG emissions by up to 50% (3). This transition can be facilitated by strategies that aim to increase the cultivation of protein crops in Europe. The European Union has advocated for increased cultivation of legumes and protein-rich oilseeds with the goal of enhancing Europe's protein self-sufficiency (4). Growing legumes locally in Europe provides environmental

benefits through improved soil fertility, reduced fertiliser use, and enriches the range of species in crop rotations (5). These aspects positively contribute to a more sustainable agriculture.

Among legumes, faba bean (*Vicia faba* L.) is a promising cool-season crop for European climates, but it has received limited attention in either breeding, cultivation or research. Faba bean stands out for its high protein content (29% on average) (6), yield potential (up to 8–9 tonnes/ha) (7), nitrogen-fixing efficiency (60–250 kg/ha/year), and wide adaptation to different soil and climatic conditions (8). Despite its benefits, faba bean's global cultivation area declined from 3.7 million hectares in 1980 to 2.1 million in 2014 (9), and it still occupies less than 1% of Europe's total cropland (10). This decline can be attributed to several reasons, including a lack of major efforts in research and breeding, that results in a limited choice of high-performing varieties for farmers. Breeding challenges in faba bean have included its variable outcrossing rates (between 4% and 84%), the absence of wild progenitors, cross-incompatibility with other *Vicia* species, and its particularly large (~ 13Gb) and a lack of a sequenced genome available until recently (11). The release of the first reference genome in 2023 is a significant breakthrough, offering a crucial tool for genetic analysis and breeding (12). Despite the extensive genetic diversity present in faba bean, the insufficient characterization of germplasm hampers the potential for its exploitation in breeding. Germplasm collections and local populations maintained by farmers preserve all the existing genetic diversity for this species, representing its primary gene pool (13). Approximately 50,000 faba bean accessions are available worldwide (11), most of which remain uncharacterized both phenotypically and genetically. In the Netherlands, the Center for Genetic Resources (CGN) maintains around 700 accessions. Ongoing research focusing on characterizing extensive genebank collections highlights the importance of further exploring the crop's genetic diversity (14). Göhl et al. (2017) suggest that faba bean's genetic variability remains largely intact and it has not been lost due to high selection and breeding pressure (15). Hence, germplasm collections are an important reservoir of allelic variation to exploit in breeding.

Breeding efforts to improve yield and its stability can considerably increase the economic attractiveness of this crop. Improved varieties could yield up to 9 tonnes per hectare under optimal agronomic practices and favourable weather. Nevertheless, a gap exists between potential and actual yields. Faba bean yield is highly unstable, as a result of large genotype-environment interaction and vulnerability to various abiotic and biotic stress (16). Enhancing adaptation and yield through breeding requires a deep understanding of the biological and environmental factors impacting plant growth. An important step is to identify yield components, given the limited current knowledge of the genetic architecture of yield-related traits.

Some studies have identified quantitative trait loci (QTL) controlling yield-related traits, traditionally through biparental QTL mapping, and recently through genome-wide association studies (GWAS). Multiple QTL mapping studies have already been conducted for yield-related traits including flowering time, pod length, and seed characteristics (17–21). All of these studies used mapping approaches based on biparental populations. Biparental QTL mapping is a low-resolution method limited by few recombination events which hinders gene discovery (22). Following the genome release, the first genome-wide association studies (GWAS) were published to improve mapping resolution. Skovbjerg et al. (2023) identified 238 marker-trait associations affecting yield-related traits using a 7-parent MAGIC population (23). Ohm, Åstrand (24) and Gutierrez, Pegard (25) used larger diversity panels to identify 51 (24) and 112 (25) marker-trait associations across various yield-related traits, demonstrating the benefits of exploiting broad diversity in genetic analysis.

Additional research is necessary to identify molecular markers and candidate genes for yield-related traits. Only a few studies have utilized a broad diversity panel for molecular markers detection via GWAS. The effectiveness of GWAS in uncovering genetic signals is influenced by factors such as the chosen population, marker allele frequencies in that population, the density of genetic markers, and the testing environment. It is essential to conduct further GWAS on yield-related traits in varied environments and with different genotyping tools to further dissect the genetic control of these traits and to identify robust molecular markers. However, the effectiveness of using GWAS-detected molecular markers remains uncertain for significant genetic gains in highly polygenic traits. Hence, it is necessary to explore alternative approaches to support the breeding of quantitative traits. Genomic selection (GS) has emerged as a more effective methodology for complex trait genetic improvement, and it has been widely applied in the breeding research of other legumes (26–28). However, its utilization is still largely untapped in faba bean breeding in both research and practical applications.

In this context, we assembled a diverse panel of 256 accessions and genetically characterized them by using a targeted genotyping assay. The objectives of our study were: (1) to evaluate the panel for traits such as flowering time, plant height, seeds per pod, pod length, single seed weight, and yield; (2) to dissect their genetic architecture via GWAS, aiming to identify new candidate genes and molecular markers; and (3) to investigate the implementation of Genomic Selection (GS) as a tool to estimate the genetic merit of unphenotyped seed bank resources.

2 Material and Methods

2.1 Plant material and field trials

A total of 250 faba bean accessions were collected from the Centre for Genetic Resources (CGN) located in Netherlands. In addition, six commercial varieties were provided by the breeding companies Limagrain (LG) and NPZ. This complete set of 256 accessions is referred to as the “CGN population”. The CGN population comprises accessions originated from 43 countries, including 158 landraces, 84 varieties, 3 research/breeding lines, and 9 accessions with an unknown genetic status (Supplementary Table 1).

The plants were grown in Winschoten in the Netherlands (53°08'28.7"N 6°59'33.7"E) in 2021 and 2022. They were sown on March 24 and harvested on August 17–18 in 2021, and sown on April 13 with a harvest on August 22–23 in 2022. The difference in sowing dates was due to differences in annual climatic conditions.

The experimental design was a resolvable row-column design (RRC) with two complete blocks (i.e., replicates), each featuring 11 rows and 25 columns. The six commercial varieties were planted in four plots within each block as controls to better evaluate the field and season conditions. The design was generated by using the FieldHub R package (29). The randomization within the RRC design was different in 2021 and 2022. The plots were 1.8 m² plots and hosted 20 plants per accession, maintaining a spacing between them of 0.3 m.

Field management adhered to the local farmers' practices and irrigated exclusively through rainfall. 45 kg of KAS fertilizer (N 27%, NH₄ 13.5%, NO₃ 13.5%, and MgO 4%), and 26 kg of KALI60 (K₂O 60%), were applied in

both years.

2.2 Phenotyping of agronomic traits

The phenotypic traits recorded included flowering time (FT), plant height (PH), pod length (PL), seeds per pod (SP), single seed weight (SSW), and yield (YLD). The trait PL was only recorded in 2021. The data were recorded on the six inner plants of each plot to minimize potential border effects.

The FT was recorded when 50% of the plants per plot exhibited at least one open flower, and quantified in day after sowing (DAS). The PH was measured from the collar to the top of the plant in centimetre (cm) by using a standard ruler at harvest. After drying in the greenhouse, pods from each plot were photographed at 50 cm by using a smartphone in a holder. The pictures were analysed using ImageJ software to determine PL in cm, using a known size label as a reference. The SP was estimated by dividing the total seed count by the total pod count. The SSW was calculated by dividing the total seed weight by the number of seeds, providing an average weight per seed in grams. The total yield referred to plot yield (grams/plot), corresponding to the total weight of seeds per plot.

2.3 DNA extraction, sequencing, and single nucleotide polymorphisms (SNPs) typing

Young leaves from ten plants per accession were collected, lyophilized separately, and pooled in equal weights to capture intra-accession allelic variation in the partially allogamous *Vicia faba* L. DNA was extracted by using the NucleoMag® DNA (Bioké, the Netherlands) extraction kit. The company IGAtch performed DNA quality check, normalization, and library preparation using the NuGEN Technologies protocol, followed by sequencing on the Illumina NovaSeq 6000 with pair ends (2x150bp) (Illumina, San Carlos, CA). The 90k-SPET® technology (Allegro® Targeted Genotyping) was adopted as a targeted method to select and sequence ~ 90,000 genomic areas.

The sequencing reads were aligned to the faba bean reference genome (Faba bean cv. Hedin) using BWA-MEM v0.7.17 (30), selecting uniquely aligned pairs of reads with a mapping quality > 10. Variant calling was executed with freeBayes v1.3.2 (31) within a 300 bp range from SPET probes, treating samples as haplotype mixes (--pooled-continuous option). Variant normalization was done using BCFtools (32).

Allelic frequencies at each biallelic locus were calculated using a custom Python script, assigning a score s to SNPs as $s = \frac{A_r}{A_r + A_a}$, where A_r and A_a are the reads counts of the reference and alternative alleles, respectively. To minimize errors and bias, s was calculated only for SNPs with an average coverage > 100x across all samples and > 30x in over 15% of the samples. Additionally, the least represented allele (either A_r or A_a) needed at least 2 reads per pool and presence in at least 5% of other samples.

The final SNPs retained for GWAS were selected based on a minor allele frequency (maf) filter of at least 0.04 and a maximum of 6% missing values (NA) per SNP. For Genomic Selection, a stricter maf of 0.07 was used. Missing values were imputed using the mean imputation method. The filtering and imputation were done with a custom R script.

The unique SNP identification code used in this study is: “chromosome_position” (e.g., chr1S_1370573844 indicates the SNP at position 1370573844 on chromosome 1S).

2.4 Genomic relationship matrix (G), PCA, and homozygosity level using SNPs data

The Genomic Relationship Matrix (**G**) and Principal Component Analysis (PCA) were computed using a pruned set of SNPs. SNP pruning was done using a custom R script, retaining only one SNP per pair that showed a Pearson correlation > 0.4 within a 200 Kbp window. In total, 2631 markers out of 38,014 were discarded.

The **G** matrix was computed by using the VanRaden method (33) implemented in the AGHmatrix R package (34), which is adapted for pool genotypes and allelic frequency as marker score. Subsequently, the **G** matrix was blended using the formula described by Nazarian and Gezan (35): $\mathbf{G}^* = \mathbf{G}(1 - p) + p\mathbf{I}$, where $\mathbf{G}^*$ is the new blended matrix, **G** is the original matrix, *p* is the proportion of the identity matrix **I** to be considered (0.02), and **I** is the identity matrix to be blended (36). Henceforth, the term **G** denotes its blended form $\mathbf{G}^*$, which is invertible.

PCA was performed by using the base stats R package. Four principal components (PCs) were retained based on the variance-explained criteria. The PCs loadings were inspected to ensure that they represented genome-wide variation, based on the assumption that if SNPs with high loadings are all concentrated in a single or few regions of the chromosome, they are likely representing local genomic structure rather than population structure (37). Four clusters were identified by an iterative K-means clustering on the four PCs using the factoextra R package (38). The goodness of clustering was assessed by examining the within-cluster sums of squares (WSS) of 1 to 15 clusters by using the elbow method in a scree plot.

Later, the CGN population’s homozygosity was assessed by defining a within-pool locus as ‘fully’ homozygous when the frequency of a unique allele > 0.95.

2.5 Statistical analysis of the field trials data

The phenotypic data analysis was performed by fitting different linear mixed models via ASReml 4.2-R package (39), which estimates variance components using the restricted maximum likelihood (REML) methodology.

In order to estimate the narrow sense heritability (h^2), the following single-trait and single-environment analysis was performed using raw data based on the following generic model:

$$\mathbf{y} = 1\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u} + \mathbf{Z}_2\mathbf{r}_{(b)} + \mathbf{Z}_3\mathbf{c}_{(b)} + \mathbf{e} \quad (1)$$

where **y** represents a vector of raw phenotypes, μ the overall mean, $\boldsymbol{\beta}$ the fixed effect coefficients for blocks, **u** the random genetic effects, with $\mathbf{u} \sim N(\mathbf{0}, \sigma_g^2\mathbf{G})$; $\mathbf{r}_{(b)}$ the random rows effects within blocks, with $\mathbf{r}_{(b)} \sim N(\mathbf{0}, \sigma_r^2\mathbf{I}_r)$; $\mathbf{c}_{(b)}$ the random columns effect within blocks, with $\mathbf{c}_{(b)} \sim N(\mathbf{0}, \sigma_c^2\mathbf{I}_c)$; **e** represent the random residual errors with $\mathbf{e} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, and $\boldsymbol{\Sigma} = \sigma_e^2(\mathbf{AR1}_r \otimes \mathbf{AR1}_c)$. The sign $\otimes$ is the Kronecker product of two autoregressive structures of the rows and columns coordinates, respectively. The matrix **AR1** accounts for micro-

environmental variations between residuals. $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices of their respective orders, $\mathbf{I}$ identity matrices of their respective orders, and $\mathbf{G}$ is the Genomic Relationship Matrix capturing the additive relationship among individuals, described earlier.

The residuals were assessed for normality by exploring diagnostic plots after model fitting. Outliers were defined as residuals exceeding three times the standardized residual's standard deviation. Excluding outliers, models were re-fitted to calculate the narrow sense heritability (h^2) as follows: $h^2 = 1 - \frac{[mean(PEV_g)]}{\sigma_g^2}$ (40), with $mean(PEV_g)$ being the average prediction error variance (PEV) for the genotypes and σ_g^2 the estimated additive genetic variance.

To estimate an adjusted mean for each genotype (Y_{adj}), a similar model to (1) was fitted with $\mathbf{u}$ as a fixed genetic effect.

The type-B additive genetic correlations (41) between 2021 and 2022 were estimated by fitting the following single-trait and multi-environment trial (MET) model using raw data:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}_1\mathbf{t} + \mathbf{X}_2\boldsymbol{\beta}_{(t)} + \mathbf{Z}_1\mathbf{u}_{(t)} + \mathbf{Z}_2\mathbf{r}_{(bt)} + \mathbf{Z}_3\mathbf{c}_{(bt)} + \mathbf{e} \quad (2)$$

where $\mathbf{y}$ is a vector of raw phenotypes and μ is their overall mean across years, $\mathbf{t}$ is a vector of fixed year effect, $\boldsymbol{\beta}_{(t)}$ is a vector of fixed blocks effects within the years, $\mathbf{u}_{(t)}$ is a vector of random genetic effects within years, with $\mathbf{u}_{(t)} \sim N(\mathbf{0}, \sigma_g^2 \mathbf{G} \otimes \mathbf{U}_0)$ with $\mathbf{G}$ representing the genomic relationship matrix and $\mathbf{U}_0$ is a 2x2 variance-covariance (vcov) matrix between the two years with a heterogeneous structure (corgh), allowing for year-specific σ_g^2 and genetic correlations (ρ_{type-B}) estimation; $\mathbf{r}_{(bt)}$ is the vector of random effects for rows within blocks within years, with $\mathbf{r}_{(bt)} \sim N(\mathbf{0}, \sigma_r^2 \mathbf{D}_r)$, $\mathbf{c}_{(bt)}$ is the vector of random effects for columns within blocks within years, with $\mathbf{c}_{(bt)} \sim N(\mathbf{0}, \sigma_c^2 \mathbf{D}_c)$; $\mathbf{e}$ is the vector of residual errors, with $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{D}_e)$ and with heterogeneity defined as $\boldsymbol{\Sigma} = \sigma_e^2 (\mathbf{AR1}_r \otimes \mathbf{AR1}_c)$. $\mathbf{D}_r$, $\mathbf{D}_c$, and $\mathbf{D}_e$ are block diagonal matrices where each year level has its own component, σ_r^2 , σ_c^2 , and σ_e^2 respectively. All other terms were previously described. A value of ρ_{type-B} near one (obtained from the matrix $\mathbf{U}_0$) indicates minimal Genotype-by-Year (GxY) interaction. The approximated standard errors of ρ_{type-B} were estimated by using the Delta method (42). The significance of the year as a covariate was assessed by the incremental Wald statistic.

To estimate the type-A additive genetic correlations among traits (41), the following multi-trait (multivariate) and single-environment model was fitted using raw data:

$$\mathbf{y} = \mathbf{1}\mu_m + \mathbf{X}_1\mathbf{t} + \mathbf{X}_2\boldsymbol{\beta}_{(t)} + \mathbf{Z}_1\mathbf{u}_{(t)} + \mathbf{Z}_2\mathbf{r}_{(bt)} + \mathbf{Z}_3\mathbf{c}_{(bt)} + \mathbf{e} \quad (3)$$

where $\mathbf{y}$ is the stacked vector of phenotypes including all traits $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n]$; μ_m is the multivariate version of intercept (μ); $\mathbf{X}_2\boldsymbol{\beta}_{(t)}$, $\mathbf{Z}_2\mathbf{r}_{(bt)}$, $\mathbf{Z}_3\mathbf{c}_{(bt)}$ are stacked design effects defined as previously in the model (2), but with 't' now denoting the trait, $\mathbf{u}_{(t)}$ is a stacked vector of random genetic effects within trait, with $\mathbf{u}_{(t)} \sim N(\mathbf{0}, \sigma_g^2 \mathbf{G} \otimes \mathbf{U}_0)$, $\mathbf{G}$ is previously described, and $\mathbf{U}_0$ is a 6x6 vcov matrix between all traits (5x5 in 2022 because one trait less was analyzed) with a heterogeneous structure (corgh), allowing for trait-specific σ_g^2 and unique genetic

correlations across traits ($\rho_{\text{type-A}}$), $\mathbf{e}$ is a stacked vector of random residual effects, with $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I} \otimes \mathbf{R}_0)$ and $\mathbf{R}_0$ is the vcov between errors that is heterogeneous (corgh), introducing residual correlations between traits and unequal σ_e^2 across traits. The approximated standard errors for correlations were obtained by using the delta method (42). Convergence in REML for the multivariate model was aided by using the starting values for σ_g^2 and σ_e^2 estimated from the univariate model (1) for each trait.

2.6 Linkage disequilibrium decay

Simple linear regression models were fitted between pairs of allelic frequency at each SNP with maf > 0.04, treating one SNP as the response variable and the other as the predictor variable. The strength of linkage disequilibrium (LD) was approximated by the ratio of regression sum of squares (SSR) to the total sum of squares (SST), expressed as r^2 (43). To represent the LD decay, r^2 between markers was plotted against their physical distance. The genome was segmented into 300 Kbp bins, and the within bins 99th percentile r^2 values and mean marker distances were chosen as representative data points to fit a Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) via the `scipy.interpolate` module in Python (44, 45). LD decay was defined as the physical distance where r^2 halved from its maximum ($\text{LD}_{\text{max}1/2}$).

2.7 Genome-wide association study (GWAS)

GWAS was performed using a linear mixed model approach implemented in the `statgenGWAS` R package (46), as follows: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$, where $\mathbf{y}$ represents the vector of the Y_{adj} , $\mathbf{X}$ is the design matrix for fixed effects including the intercept, SNPs, and the genetic clusters to which each accession belongs; $\boldsymbol{\beta}$ denotes the vector fixed effect coefficients, with β_{SNP} specifying the SNP-effect; $\mathbf{Z}$ is the incidence matrix; $\mathbf{u}$ is the random additive genetic effects, with $\text{var}(\mathbf{u}) = \sigma_g^2 \mathbf{G}$. $\mathbf{G}$ was defined before (section 2.4), and it captures the shared additive genetic effects among individuals. Finally, $\mathbf{e}$ is the vector of residual errors, with $\text{var}(\mathbf{e}) = \sigma_e^2 \mathbf{I}$.

StatgenGWAS estimates σ_g^2 and σ_e^2 using the Efficient Mixed Model Association (EMMA) algorithm (47). Initially, a base model calculates the REML-variance components excluding the SNP fixed effects. Then variance components are fixed, and the null hypothesis $\beta_{\text{SNP}} = 0$ is tested for each SNP individually using a general F-statistic within the generalized least squares (GLS) framework.

The model's fit was gauged using the genomic inflation factors from `statgenGWAS`. SNPs exceeding a 4.5 LOD threshold were deemed significant. Furthermore, adjusted p-values were obtained using the Benjamini-Hochberg method (48).

2.8 Candidate gene identification and putative SNP effects

The candidate genes were identified near significant SNPs considering the LD decay. The genes' function were obtained from the public available annotation data generated using Prot-Scriber (<https://github.com/usadellab/prot-scriber>). For intragenic SNPs, SNPeff tool (49) was used to predict the SNPs effect on the protein produced. SNPeff categorizes SNP impacts as “high” (disruptive), “moderate” (potentially altering protein function), “low” (likely harmless), or “modifier” (affecting non-coding regions or genes).

2.9 Genomic selection based on a Bayesian modelling approach

Genomic Selection (GS) models were fit using the Bayes B method. The models were simply formulated as $\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{y}$ represents the vector of Y_{adj} ; μ is the intercept; $\mathbf{Z}$ denotes the marker incidence matrix, and $\boldsymbol{\beta}$ the corresponding vector of marker effects; lastly, $\mathbf{e}$ stands for the vector of residual effects.

In Bayes B, each marker's effect (β for maker j) has a prior that is a mixture of point of mass at zero and scaled-t density slab. The zero-point mass suggests no effect with probability π , while the scaled-t slab allows for an effect with probability $1 - \pi$. The hyperparameter π follows a Beta distribution $B(\pi | p_0, \pi_0)$, with shape $p_0 > 0$ and $\pi_0 \in [0, 1]$. The variance of the marker effects β_j are unknown and modelled by assuming a prior distribution from an inverse chi-squared (χ^{-2}) with 5 degrees of freedom and scale parameters $S\beta$. The $S\beta$'s uncertainty is represented by a gamma density with rate (r) and shape (s) parameters, $G(s_\beta | r, s)$. Bayes B model was fit using the BGLR R package with default hyperparameters (50). 30,000 iterations were performed for sampling and the initial 6,000 sampled values were discarded as burn-in.

The GS models were evaluated through a 5-fold cross-validation (CV), repeated 10 times, for the 2021 and 2022 data separately. The predictive ability (PA) and prediction accuracy (ACC) were used to evaluate the models' performance. The PA is the correlation between the Y_{adj} and predicted genomic estimated breeding values (GEBV), denoted as $\text{corr}(y_{\text{adj}}, \hat{g})$. For each CV round, the heritability (h^2) was estimated from the GS models as $h^2 = 1 - \frac{\sigma_e^2}{\sigma_g^2}$, where σ_e^2 and σ_g^2 represent the error variance and genetic variance, respectively. Then, h^2 was used to calculate the ACC as: $\text{ACC} = \frac{\text{PA}}{\sqrt{h^2}}$. The ACC is the correlation between the true genomic breeding value and the predicted breeding value, denoted as $\text{corr}(g, \hat{g})$. Furthermore, GS models were evaluated for cross-year prediction by calculating the PA, using 2021 data to predict 2022 and vice versa.

3 Results

3.1 Phenotypic data analysis

3.1.1 Large phenotypic variation and narrow sense heritability for all traits

Phenotypic data were recorded in the Netherlands in 2021 and 2022. After fitting the linear mixed models, none of the traits deviated importantly from normality, but some observations were regarded as outliers (Table 1).

Overall, a large phenotypic variation was observed in both years, as shown by the coefficients of variation (CV%) ranging from $\sim 4.5\%$ to $\sim 59\%$ in 2021, and from $\sim 8\%$ to $\sim 44\%$ in 2022 (Table 1 and Supplementary Fig. 1). Flowering time had the lowest CV value ($\sim 4.5\%$ in 2021). Nevertheless, the earliest accessions

flowered two weeks earlier than the latest in this season. In contrast, SSW and YLD had the largest CV% in both years (Table 1), indicating a vast variability in seed size and total production per plot.

Table 1
Descriptive statistics for the adjusted phenotypic data (BLUEs estimate mean).

Trial	2021						2022					
Trait	FT	PH	PL	SP	SSW	YLD	FT	PH	SP	SSW	YLD	
Min	69.5	36.1	3.2	1.5	0.1	0.2	44.5	67.2	1.3	0.2	11.9	
Max	87	120.9	13.1	3.7	1.9	74.3	71	175.2	4.3	1.5	116.3	
Mean	75.4	64.2	5.7	2.6	0.6	20.2	56.3	112.6	2.7	0.7	50.7	
SD	3.4	14.4	1.6	0.4	0.3	12	4.5	24	0.5	0.3	17.8	
CV%	4.5	22.4	27.6	16.5	43.7	59.5	7.9	21.3	18.1	44.3	35.1	
Outliers	7	5	4	3	8	11	9	1	6	12	-	

The table details the minimum (min), maximum (max), mean, standard deviation (SD), coefficient of variation in percentage (CV%), outliers count (outliers), in 2021 and 2022.

The narrow sense heritability (h^2) values were positively high in both years with minimal differences across seasons (Table 2). In both years, YLD had the lowest h^2 , while FT and SSW the highest.

Table 2
Narrow sense heritability (h^2) in 2021 and 2022.

Trait	2021	2022
FT	0.89	0.80
PH	0.73	0.83
PL	0.77	-
SP	0.70	0.79
SSW	0.86	0.88
YLD	0.49	0.50

3.1.2 Lack of strong negative type-A correlations across traits

The additive genetic correlations across traits (type-A) were estimated using a single-year multi-trait linear mixed model (model 3). Overall, the correlation trends were consistent over the two years with only minor differences. Except for a moderately negative correlation between SSW and FT ($\rho_{\text{type-A}} = -0.25$ in 2021, and $\rho_{\text{type-A}} = -0.21$ in 2022), all other traits were not correlate negatively in either year. As shown in Fig. 1, YLD was weakly positively correlated with SSW and PL, and showed moderate to strong correlation with SP and PH in

2021 and 2022. Notably, SSW and PL strongly correlated in 2021 ($\rho_{\text{type-A}} = 0.77$), indicating a potential shared genetic control between these two traits.

3.1.3 Important Genotype-by-Year (GxY) interaction for yield

The additive genetic correlations across the two years (type-B) were estimated using a MET linear mixed model approach (model 2). The extent of genotype-by-year interaction (GxY) varied across traits. YLD was the least stable trait with a $\rho_{\text{type-B}}$ of 0.42 and an approximated standard error of ± 0.15 , followed by PH 0.70 ± 0.05 , SP 0.83 ± 0.05 , and FT 0.83 ± 0.02 . SSW was the most stable trait across years, as indicated by the particularly high additive genetic correlation of 0.95 ± 0.02 .

The year of cultivation had a significant effect on all the traits based on Wald test statistics (model 2). FT, PH, and YLD all exhibited p-values less than 2.2×10^{-16} , while SP and SSW of 5.39×10^{-6} and 7.9×10^{-6} , respectively.

3.4 Genotypic data analysis

3.4.1 A high-quality SNPs panel

The 90K-SPET assay effectively identified high-quality SNPs. Remarkably, 87% of the paired-end sequencing reads aligned uniquely and with high confidence to the genome, underscoring the effectiveness of the selection of specific regions for targeted sequencing.

DNA sequencing achieved a median depth of coverage of 240x. Only five samples with coverage below 100x were discarded. The sequencing generated 2,547,767 SNPs, 161,445 Indels, and 153,894 “other type” variants. For GWAS, 38,014 SNPs were selected based on $\text{maf} > 0.04$, of which $\sim 85\%$ were located within genes. Intragenic SNPs can increase the chance of finding candidate genes in a direct association. For GS, 21,935 SNPs were selected based on $\text{maf} > 0.07$.

Overall, the SPNs used in GWAS were well distributed across the genome (Table 3 and Supplementary Fig. 2). The average marker density was ~ 3 SNPs per 1 million base pairs (1 Mbp), with the highest local density on chromosome 2 and the lowest on chromosome 1L. Chromosome 6 had the maximum distance gap between adjacent SNPs (Table 3).

Table 3
SNPs distribution and coverage along the six faba bean chromosomes.

Chromosome	Total SNPs ^a	Covered distance (Mbp) ^b	Average density (SNPs/Mbp)	Average distance (Kbp) ^c	Max gap (Mbp) ^d
chr1L	6327	1802	3.51	285	31.2
chr1S	3959	1568	2.52	396	28.3
chr2	6125	1717	3.57	280	22.7
chr3	5720	1637	3.49	286	44.1
chr4	4954	1644	3.01	332	60.1
chr5	4413	1365	3.23	309	70.4
chr6	4512	1520	2.97	337	85.6
contig	2004	-	-	-	-
^a 38,014 SNPs selected for GWAS					
^b physical distance between the first and last SNP marker in Megabase pair (Mbp)					
^c average physical distance between each SNP and its nearest neighbour in Kilobase pair (Kbp)					
^d maximum physical distance between two neighbouring SNP markers in Megabase pair (Mbp)					

3.4.2 Four genetic groups

Population structure was analysed by using PCA, followed by a K-means clustering analysis. Four PCs were selected as remaining additional components contributed to explaining less than 1% additional variance. No genomic area displayed disproportionally high and clustered SNP loadings for the main PCs (Fig. 2-B) excluding the presence of genomic regions that could bias the population structure analysis. PC1 (~ 6.5% variance) differentiated Western/Northern European from South Asian accessions (mainly Pakistan and Afghanistan), while PC2 (~ 3.8%) distinguished Ethiopian from South Asian and European accessions (Fig. 2-A).

K-means clustering identified four main clusters based on the within sum of squares criteria (Supplementary Fig. 3). Cluster 4 was the biggest group with 102 accessions, followed by Cluster 2 with 101 accessions, Cluster 1 with 24, and Cluster 3 with 18.

Cluster 1 included predominantly South Asian accessions from Afghanistan and Pakistan. Cluster 2 comprised accessions from Northern Europe, Western Europe, Southern Europe and most Canadian accessions (Supplementary Table 2). This cluster also included five commercial varieties from breeding companies with the exception of LGCARTOUCHE. Cluster 3 was dominated by Ethiopian accessions, but it also included the commercial variety LGCARTOUCHE and one Swedish accession (CGN12328). Cluster 4 included accessions from Northern and Eastern Africa, Southern and Western Europe, Western, Eastern and Southern Asia (Supplementary Table 2).

3.4.3 Low level of genetic relatedness

The genetic relationship among accessions was investigated by inspecting the genomic relationship matrix (**G**). The diagonal values of **G** suggested a generally low level of inbreeding. The values ranged from ~ 0.57 to ~ 1.90, with an expected average value around 1 (Fig. 3). The average **G** off-diagonal values was as low as 0, and the third quartile distribution value was approximately 0.03, suggesting an overall limited genetic relatedness and family structure within the CGN population (Fig. 3). However, a marked presence of genetic relationship was found between individuals in Cluster 1 and Cluster 3, having a median off-diagonal above 0.23 and 0.25 respectively (Supplementary Fig. 4). Interestingly, certain accessions from Ethiopia showed important genetic relatedness with some Swedish accessions in Cluster 4, witnessing the use of Ethiopian material in breeding programs in Northern Europe.

3.4.4 Varying homozygosity level

The homozygosity level per accession was expressed as the proportion of homozygous loci out of the total number of genotyped loci. Different levels of homozygosity were observed, spanning from ~ 63% to ~ 97%, with a median value of 75% (Supplementary Fig. 5). As expected, accessions with lower homozygosity had smaller diagonal values in the **G** matrix. The four genetic clusters (section 3.4.2) showed similar levels of homozygosity, indicating that population structure does not affect homozygosity, and thus the reproductive biology of faba bean.

3.4.5 Linkage disequilibrium decay at 268 Kbp

The strength of linkage disequilibrium (LD) was estimated using the r^2 statistic derived from allelic frequencies in pooled samples. The $LD_{\max1/2}$ occurred at a physical distance of ~ 268.79 Kbp, with $LD_{\max1/2}$ indicating the distance at which the maximum r^2 (0.89) was halved (Fig. 4).

3.5 GWAS

3.5.1 SNP-trait associations in 2021 and 2022

A GWAS for FT, PH, PL, SP, SSW was performed in 2021, and repeated for all traits except PL in 2022. SNPs were declared significant when they had a logarithm of odds (LOD) value larger than 4.5. Furthermore, adjusted p-values were computed using a conservative multi-hypothesis correction framework. The genomic inflation factors for each GWAS model were all very close to the expected values of one (Supplementary Table 3).

GWAS uncovered as many as 33 SNP-trait associations in 2021. Specifically, five SNPs were significant for FT, five for PH, nine for PL, two for SP, and 12 for SSW (Fig. 5). PL and SP had a common genetic control sharing SNP chr5_488402698. Similarly, PL and SSW shared the SNPs chr1S_1370573844 and chr1S_1378555701 (Table 4 and Table 5).

In the 2022 experiment replication, GWAS revealed a total of 17 SNP-trait associations, two for FT, three for PH, four for SP, with SSW showing again the highest count of SNPs (eight) (Fig. 6). Positively, one SNP for FT (SNP chr5_46897321) and four SNPs for SSW (SNP chr1L_901724768, chr1S_1370573844,

chr1S_1378555701, SNP chr4_1165005706) correlated with these traits across years. This consistency highlights potential stable QTLs and interesting molecular markers to test in marker assisted selection. SNP chr5_468973212 explained ~ 9% of the flowering time phenotypic variance, while the four stable SNPs for SSW explained individually from ~ 3.5 to 7.1% of the trait's phenotypic variation.

As expected for complex quantitative traits, their genetic control was simultaneously regulated by multiple minor QTLs. In fact, no significant SNPs accounted for more than 10% of the phenotypic variation in either 2021 and 2022. Most of the minor QTLs were not consistent across years. For example, the traits PH and SP were associated with different chromosome segments in 2021 and 2022.

It is important to note that a total of 31 SNPs in 2021 and 15 in 2022 were intragenic (Table 4 and Table 5), increasing the likelihood of finding candidate genes with high resolution. However, these SNPs might also serve as a proxy for causal polymorphisms situated in the other adjacent regions. Ten intragenic SNPs in 2021 and five in 2021 were identified as missense variants. These variants were predicted to potentially impact protein sequences through amino acid substitution, thus representing candidates causal polymorphisms. As shown in Table 4 and Table 5, missense variants were located in the genes *GAT*, *DOF*, and *RIBA*, potentially influencing flowering time, in *PLATZ* affecting plant height, in *Kinesin-7*, *PRR*, and *CMU* regulating pod length, and in genes such as *TBL13*, *PRR*, *DCL1-HYL1*, *THG*, *RAD5*, *PHI-1*, and a likely *ZNFs* gene, putatively involved in seed size control.

The complete set of genes within the LD window from the significant SNPs are reported in Supplementary Table 4 (2021) and Supplementary Table 5 (2022).

3.6 Consistent GS predictions across 2021 and 2022

In addition to GWAS, all single SNP effects were modelled simultaneously via Genomic Selection (GS). GS was tested in 5-fold cross-validation (CV) repeated ten times for both 2021 and 2022 datasets (intra-trial predictions). A cross-year prediction was also performed using 2021 data to predict 2022 and vice versa.

Overall, GS performed robustly in 5-fold CV, with consistent and high average predictive ability (PA) and accuracy (ACC) over both years (Table 6 and Fig. 7A-B). In 2021, the average 5-fold CV's PAs varied based on the trait from 0.58 to 0.81, while the ACC from 0.72 to 0.87. The trait SSW was the most predictable and YLD the least predictable. In 2022, the average predictions remained stable, with minor changes being trait-specific (Table 6). Positively, the h^2 values estimated for each round of CV were exceptionally high, exceeding 0.65 for all the traits.

Table 6
Summary of the 5-fold cross-validation (CV) results in 2021 and 2022.

Trait	2021			2022		
	PA	ACC	h ²	PA	ACC	h ²
FT	0.69 ± 0.07	0.76 ± 0.08	0.82 ± 0.01	0.68 ± 0.04	0.75 ± 0.05	0.80 ± 0.01
PH	0.75 ± 0.08	0.83 ± 0.09	0.82 ± 0.01	0.75 ± 0.06	0.83 ± 0.07	0.82 ± 0.02
PL	0.77 ± 0.06	0.83 ± 0.06	0.85 ± 0.01	-	-	-
SP	0.53 ± 0.11	0.63 ± 0.14	0.72 ± 0.04	0.68 ± 0.09	0.76 ± 0.10	0.81 ± 0.02
SSW	0.81 ± 0.05	0.87 ± 0.06	0.87 ± 0.01	0.85 ± 0.02	0.91 ± 0.02	0.88 ± 0.01
YLD	0.58 ± 0.10	0.69 ± 0.13	0.71 ± 0.06	0.47 ± 0.12	0.59 ± 0.17	0.65 ± 0.06

The tables reports mean predictive ability (PA), prediction accuracy (ACC), and the heritability (h²) estimated from the GS models, with standard error (±).

Across-year predictions showed stable PA, with 2021 predictions for 2022 ranging from 0.37 to 0.92. Specifically, PA was 0.77 for FT, 0.78 for PH, 0.72 for SP, 0.92 for SSW, and 0.37 for YLD. Similarly, when 2022 was used to predict 2021, PA ranged from 0.40 to 0.92, maintaining consistency across traits. In detail, PA was 0.78 for FT, 0.78 PH, 0.69 for SP, 0.92 for SSW, and 0.4 for YLD. Therefore, transitioning from within-trial to across-year predictions did not importantly reduce PA, as across-trial outcomes closely matched the CV results.

4 Discussion

Farmers require high-yielding and stable varieties to justify integrating faba bean into their cropping systems. Leveraging the allelic diversity in genebanks can facilitate breeding efforts, and developing molecular breeding tools can accelerate genetic gains beyond phenotypic selection. In this study, we screened genebank accessions (“CGN population”) for flowering time (FT), plant height (PH), pod length (PL), seed per pod (SP), single seed weight (SSW), and yield (YLD), in the Netherlands in 2021 and 2022. The Netherlands has seen faba bean cultivation expand importantly from 280 ha to 949 ha over a decade, witnessing growing interest among farmers. Our primary objectives were to perform a genome-wide association study (GWAS) and develop genomic selection (GS) models for these traits.

4.1 Large phenotypic variation, high heritability, and minimal genetic relationship

The CGN population encompasses much of the faba bean’s phenotypic variation when compared to other phenotypic data (23, 24, 51). This variation stems from including different botanical groups such as *Vicia faba minor*, *Vicia faba major*, and *Vicia faba equina* from 43 different countries. In 2021 and 2022, we observed significant differences in FT, PH, and YLD. The delayed flowering in 2021 was due to early sowing

and colder climate, slowing seed germination and early plant growth. The higher heights and yields in 2022 were due to improved field conditions and weather than in 2021.

Breeding high-yielding faba bean varieties suitable for broad mega-environments maybe challenging due to an important genotype-by-year interaction (GxY) for YLD, highlighted by a type-B additive genetic correlation of 0.42. Varying weather conditions or field management practices across locations could even increase GxY. Supporting this, Gutierrez et al. (2023) (25) found null genetic correlations for YLD between the UK and Spain. Positively, the type-B genetic correlations above 0.7 for the other traits indicates that GxY was not important across the two years, facilitating breeding.

Most of the type-A genetic correlations between traits were positive, suggesting a lack of significant genetic control where genes act in opposite directions. PL and SSW exhibited the strongest additive type-A correlation ($\rho_{\text{type-A}} = 0.77$), suggesting that many genes control both traits. Thus, SSW could be used as a predictor to select PL to some degree. Based on positive correlations for PH, SSW, and SP with YLD, indirect selection should focus on taller plants with larger and more seeds per pod to enhance YLD.

It is a common belief that smaller seeds do not impact yield potential in faba bean (52). However, we found a weak to moderate type-A correlation between SSW and YLD ($\rho_{\text{type-A}} = 0.25$ in 2021, and $\rho_{\text{type-A}} = 0.45$ in 2022). This population-specific positive correlation may be due to the fact that the smaller-seeded Asian and African landraces are less yield efficient than the larger-seeded European varieties in the Netherlands. This means that the genetic correlation may not necessarily be positive in other breeding populations, facilitating breeding for smaller seeds without reducing yield. Breeding for smaller seeds is advantageous because smaller size reduce breeding and production costs by improving the seed volume per unit area ratio. Moreover, developing small-seed varieties is currently supported by the rising demand for proteins in plant-based foods, as the breeding focus is shifting from seed size to nutrient profile and flavor (53).

The narrow sense heritability (h^2) in both years suggest that genetic gain can be achieved for the analysed traits via breeding. All the traits had $h^2 > 0.7$, except for yield ($h^2 \sim 0.5$). Although comparing h^2 across studies requires caution, our results align with previous studies reporting high h^2 , or high broad-sense heritability (H^2), for various agronomic traits in faba bean (23–25). Genotyping faba bean accessions is challenging due to the genetic heterogeneity within seed bank accessions. Similarly to Ohm et al. (2023) (24), we sequenced DNA from ten pooled plants per accession, estimating allelic frequency at each SNP locus rather than treating the pool as a single genotype, which usually introduces bias. The extensive set of high quality SNPs ($\sim 38,000$) that were produced enriches the available genomic resources for faba bean.

Dividing the CGN population into four main clusters enabled a geographical interpretation and traced back the migration routes of faba bean. Cluster 1 reflects the unique migratory path of the crop from the “Near East” to Asia, consisting mainly of accessions from Pakistan and Afghanistan. Cluster 3 included mainly accessions from Ethiopia, likely reflecting a distinct diversification in this country, which is the faba bean’s second centre of genetic diversity (54). The common clustering (Cluster 4) of accessions from Spain, Italy, North African countries, Turkey, Syria, and Lebanon is consistent with the migration route proposed by Cubero (55) from the Middle East via North Africa to Spain and other European regions. Cluster 2 traces the spread

from European Mediterranean regions to the broader continent, as evidenced by the accessions from Spain, the Balkans, and Italy, with a majority from Northern and Western Europe.

The faba bean's partial allogamy is mirrored in the homozygosity levels ranging from ~ 63% to ~ 97% in the CGN population. Varying heterozygosity levels in different faba bean populations have already been reported (23). The interpretation of our results should be cautious as it is complex to pinpoint accurately the homozygosity in pooled samples. Classifying a pool locus as homozygous if 95% of the sequencing reads match one allele may still mask potential heterozygosity in a few plants within the 10-plants accession.

Overall, the CGN population exhibited a low degree of genetic relatedness. Low relatedness enhances GWAS resolution due to more independent recombination events among unrelated accessions (56). The higher degree of genetic relatedness observed specifically in Cluster 1 may stem from distinct allelic frequencies within the cluster that could inflate genetic similarity when using the VanRaden method to compute the Genomic relationship matrix (G).

4.1 QTLs and genes controlling yield-related traits

Assessing a true linkage disequilibrium (LD) in pooled samples is complex due to potential multiple haplotypes within each accession (57). The LD calculated by using within-pool allelic frequencies decreased to half its maximum ($LD_{\max/2}$) at a relatively short distance of ~ 268 Kbp, as expected in a species with partial outcrossing and varying heterozygosity levels. This aligns with previous studies reporting $LD_{\max/2}$ between ~ 139 Kbp (25) and ~ 600 Kbp (23) in other diverse populations. Given this LD decay and an average distance of 307 kbp between genes, the average marker density of 3 SNPs per 1Mbp enabled high-resolution mapping.

We identified SNPs associated with FT (5), PH (5), PL (8), SP (2), and SSW (12) via GWAS in 2021. Similarly, we found SNPs for FT (2), PH (3), SP (2), and SSW (8) in 2022. No major QTLs were detected, aligning with quantitative traits typically being controlled by numerous genes with small effects. Detecting consistent minor QTLs poses a challenge as their signals can fluctuate across different years or environments due to QTL-environment interactions and the Beavis effect (58). Nevertheless, we identified one stable SNP for FT and four for SSW across the two years due to the high type-B additive genetic correlations (> 0.8). Similarly, the detection of shared SNPs between SSW and PL, and between PL and SP, reflects their positive type-A genetic correlations, supporting the presence of shared genetic control or pleiotropy.

The detected SNPs provide initial insights into candidate genes and the identified key molecular markers can be prioritized in a marker-assisted selection (MAS) validation strategy. Targeted genotyping (90K-SPET) located most of the significant SNPs within genes, aiding in directly pinpointing candidate genes or casual variants. We analysed intragenic SNPs for amino acid changes in proteins and identified multiple missense variants. This paves the way for future research to validate these SNPs as functional variants through functional studies. However, it must be said that not all intragenic variants directly impact gene function or represent the gene, thus leveraging LD in GWAS remains crucial, as GWAS focuses on recombination regions.

Flowering time (FT) was associated with a stable QTL (SNP chr5_468973212), a promising marker for MAS accounting for ~ 9% of the phenotypic variation annually. The SNP was located within a *pyrimidine reductase* (*PyrR*). *PyrR* has not been directly associated with flowering, but it ensures the availability of riboflavin, which

is a precursor of co-factors involved in critical cellular activities required for plant development. Interestingly, *PyrR* mutants exhibit delayed and abnormal seed development due to a disrupted abscisic acid (ABA) biosynthesis in maize (59). Moreover, two *cyclodeaminase* (*SARD4*) genes were found within ~ 100 Kbp distance from the SNP. *SARD4* is involved in the metabolism of pipecolic acid, which acts as a flower inducer in the aquatic plant *Lemna paucicostata* (60). This raises the question of pipecolic acid's role in faba bean flowering, meriting further investigations. SNP chr5_468973212 could also be a proxy for the regulatory elements affecting the expression of known major flowering genes present on chromosome 5 in faba bean (i.e., *SEPALLATA* at position 406,232,553 and *Flowering locus T* at 608,489,790).

Single Seed Weight (SSW) was consistently associated with four stable QTLs across years: SNP chr1L_901724768, SNP chr1S_1370573844, SNP chr1S_1378555701, and SNP chr4_1165005706. Notably, SNP chr4_1165005706 was predicted to alter the amino acid sequence of a Pentatricopeptide repeat-containing protein (PRR). PRR proteins are known to influence seed development, as seen in maize where *PPR18* dysfunction reduces seed size with an empty pericarp (61), and another *PPR* member affects kernel size via grain filling (62). Similarly, in peanuts, a *PPR* gene correlated with seed weight in QTL mapping, and the selection for a *PPR* variant potentially increased seed size in cultivated versus wild types (63). Additionally, we reported an *E3 ubiquitin ligase* gene (*DA2*) as a candidate for SNP chr1S_1370573844. The *DA2*'s role as a negative regulator of seed and organ size has been previously reported in other species (64, 65). Another interesting finding is the detection of SNP chr1L_901724768 within an *ABC transporter* gene, which belongs to a gene family crucial for seed development, germination, and plant growth (66). In tomato, an ABC transporter is the causal gene of the *Sw4.1* QTL, one of the most important QTLs underlying the evolution of seed size in the genus *Solanum* (67). This suggests that these genes may have a similar role in influencing seed size in faba bean, a hypothesis that requires further validation. Finally, SNP chr4_1165005706 is highlighted as a potential casual polymorphism within a TOUGH (TGH) component of the DCL1-HYL1-SERRATE complex. Given the *TGH*'s role in miRNA and siRNA biogenesis in *Arabidopsis* (68), it can be hypothesized that seed size may be influenced by complex gene expression regulation.

Pod length (PL) was associated with nine distinct genomic regions on chromosomes one, three, four, and five. Potential casual variants (missense variants) were found in a gene encoding a Kinesin-7 protein (SNP chr1L_1017352645), and in a gene encoding a microtubule regulatory protein (CMU) (SNP chr1S_983132287), both of which regulate microtubule movements and cytoskeleton dynamics. Interestingly, proteins associated with microtubules are key in determining plant organs' shape and size (69). For example, kinesins regulate the exponential cell production and enlargement of fruits during their early development in cucumber (70). Similarly, cellulose-microtubule uncoupling (CMU) affects cell growth when disrupted in *Arabidopsis* (71). Thus, cytoskeleton dynamics may regulate PL in faba bean, warranting further research to confirm this hypothesis. Additionally, a third putative causal SNP (chr1S_1378555701) was located within the previously mentioned *PRR*, which could be a gene that exerts pleiotropic control over SSW and PL.

Seed per pod (SP) was associated with SNPs on different chromosomes in 2021 and 2022, suggesting that multiple unlinked QTLs control this trait. Defining the functional roles of 2021's candidate genes proved challenging as most of them lacked annotations. In 2022, candidate genes on chromosome two included one for a TSUP transport protein, which plays a role in moving sulphur-containing compounds (72), and another gene encoding a sensory plastid-associated PsbP DOMAIN-CONTAINING PROTEIN 3 (PP3), which influences

plant growth (73). Additionally, we found a gene for the CRUMPLED LEAF (CRL) protein that when mutated significantly alters the morphogenesis of all plant organs in *Arabidopsis*, resulting in malformed structures (74).

Plant height (PH) is also likely controlled by multiple unlinked QTLs as SNPs were detected on different chromosomes in 2021 and 2022. Notably, SNP chr5_1006205542 was identified as a missense variant in a *PLATZ* transcription factor gene, belonging to the *PLATZ* family that plays a key role in plant growth, development, and senescence by hormones regulation. A recent study by Zhang et al. (2023) (75) pinpointed *PLATZ-A1* as the *RHT25* GA-sensitive dwarfing locus in wheat, highlighting its importance in regulation plant stature. We also identified candidate genes including *FIB4*, associated with a BPI/LBP family protein, and *MACPF domain-containing* genes, all of which are linked to plant resistance against biotic and abiotic stresses (76–78). This suggests that plants with enhanced stress response tend to be more robust and taller. Notably, within the LD window of SNP chr2_631026439, there were two *amino acid/auxin permease (AAP)* genes related to nitrogen (N) metabolism, which are responsible for encoding proteins involved in amino acid and hormone mobilization (79). Understanding the role of these genes in nitrogen distribution can provide insights into factors influencing plant vigour and height.

4.2. Genomic selection potential to enhance faba bean germplasm

Molecular markers help to identify and select faba bean genebank accessions with favourable alleles (80). As field phenotyping is costly, this approach focuses on testing only those accessions with promising molecular profiles. Yet, our GWAS indicates that numerous minor QTLs influence agronomic traits, complicating the targeted selection of accessions for phenotyping. For the first time in faba bean, we explored Genomic Selection (GS) based on SNP allele frequencies to characterize genebank accessions, as shown in white lupin germplasm collections (81).

The predictive ability (PA) and prediction accuracy (ACC) for six complex traits showed promising results, supporting the potential use of genomic selection in germplasm-improvement programs. The 5-fold cross-validations (CV) showed that PA ranged from 0.58 to 0.81 in 2021 and from 0.47 to 0.85 in 2022. Similarly, the ACC varied from 0.63 to 0.87 in 2021 and from 0.59 to 0.91 in 2022. The similar range of PA positively indicates the consistency of this approach for predictions made in two different years.

Heritability (h^2) is a major driver of PA in GS (82, 83). YLD and SP were the least predictable traits in both years, which aligns with their lower h^2 estimates. Moreover, variations in PA from 2021 to 2022 corresponded to changes in h^2 estimates, confirming its impact on predictive outcomes. The GS models explained a reasonable proportion of the additive genetic variance for a very complex traits such as YLD (estimated as PA^2/h^2). Specifically, GS models explained approximately 46% of the additive genetic variance with an average cross-validation h^2 of 0.70 in 2021. In 2022, GS models explained around 33% with an average h^2 of 0.65. Future developments should aim to enhance the proportion of additive variance explained, for example, by strategically designing training and validation sets based on genetic similarities within the germplasm, or by evaluating other GS methods. Moreover, model performance could be improved by leveraging existing genetic correlations between traits.

It is important to mention that CV's results often exceed true validation across trials or years due to potential overfitting and shared 'noise' in the data (84). A more reliable approach is to validate using independent or semi-independent data. To mimic actual validation, we predicted 2022 outcomes using 2021 data and vice versa. Encouragingly, both predictions showed similar PA, closely matching the within-trial predictions. This positive result may be due to the use of the same genotypes across years, and high type-B additive genetic correlations. As expected, YLD exhibited the lowest PA across trials due to its low genetic correlation.

5 Conclusion

In this study, for the first time in faba bean research, we combined the Genome-Wide Association Study (GWAS) approach with Genomic Selection (GS). A total of 33 SNP loci were significantly associated with five yield-related traits in 2021, and 17 SNP loci in 2022. The detection of intra-genic SNPs and short linkage disequilibrium decay (~ 268 Kbp) facilitated fine mapping. We proposed several new candidate genes involved in the control of these traits. Interestingly, we identified several missense variants that could represent causal polymorphisms directly linked to changes in phenotypes. In addition, the exploration of GS showed promising predictive ability and prediction accuracy. Across-year predictions closely matched the cross-validation results, suggesting that transitioning from within-trial to across-year predictions does not significantly reduce prediction accuracy. Thus, the development of GS models for gene bank accessions has the potential to unlock untapped diversity and accelerate trait integration into faba bean breeding programs. Once large germplasm collections are genotyped, the precise design of a GS strategy becomes feasible. However, the adoption of GS models will only be justified if genotyping costs continue to decrease in the coming years.

Abbreviations

SNP: Single nucleotide polymorphisms; FT: Flowering time; PH: Plant height; PL: Pod length; SP: Seeds per pod; SSW: Single seed weight; CGN: Center for Genetic Resources Netherlands; GWAS: Genome-Wide Association Study; GS: Genomic selection; QTL: Quantitative trait loci; PA: Predictive ability; ACC: Prediction accuracy; LD: Linkage disequilibrium; GHG: Greenhouse gas; MAGIC: Multiparent Advanced Generation Inter-Cross; CV%: Coefficient of variation in percentage; BLUE: Best linear unbiased estimates; PCA: Principal component analysis; LOD: Logarithm of the odds; PC: Principal component; G: Genomic relationship matrix; MAS: Marker assisted selection; LG: Limagrain; RRC: Resolvable Row Column design; DAS: Days after sowing; SPET: Single Primer Enrichment Technology; REML: Restricted maximum likelihood; MET: Multi-environment trial; VCOV: Variance-covariance; CORGH: General Heterogeneous Correlation structure; GxY: Genotype by Year; GEBV: Genomic estimated breeding value.

Declarations

Ethics approval and consent to participate

The study complies with relevant institutional, national, and international guidelines and legislation for plant ethics.

Consent for publication

Not applicable

Availability of data and materials

The datasets containing sequencing data generated and analysed during the current study are available in the Genome Sequence Archive (GSA) repository under the accession number CRA017806, <https://ngdc.cncb.ac.cn/gsa/browse/CRA017806>.

Competing Interests

The authors declare that they have no competing interests.

Funding

This research is part of project 'Pulses optimized for flavour and functionality' co-financed by the Top Consortium for Knowledge and Innovation Agri & Food by the Dutch Ministry of Economic Affairs under contract number LWV19028.

Authors' contributions

AL designed the study, generated the molecular and phenotypic data, performed the statistical analysis, implemented the pipeline for GWAS and GS and drafted the manuscript; SAG supervised the data analysis, contributed to results' interpretation and reviewed the manuscript; JZ contributed to phenotypic data collection; VC contributed to phenotypic data collection; BJVD contributed to field trial management and data collection; GE contributed to field trial management and data collection; JP contributed data analysis and results' interpretation and reviewed the manuscript; LMT acquired the fundings, designed the study, provided supervision in all the research steps, edited and reviewed the manuscript.

All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank the partners of the TKI pulses project. A special thanks to the employees of Vandinter Semo B.V. and UniFarm for their help and assistance in trial maintenance and phenotyping, respectively.

References

1. Ciani M, Lippolis A, Fava F, Rodolfi L, Niccolai A, Tredici MR. Microbes: Food for the Future. *Foods*. 2021;10(5):971. <https://doi.org/10.3390/foods10050971>
2. Pyett S, Jenkins W, van Mierlo B, Trindade L, Welch D, van Zanten H. Our future proteins: A diversity of perspectives: VU University Press; 2023.
3. Poore J, Nemecek T. Reducing food's environmental impacts through producers and consumers. *Science*. 2018;360(6392):987-92. doi: 10.1126/science.aag0216

4. European Commission. A Farm to Fork Strategy for a Fair, Healthy and Environmentally-Friendly Food System. 2020; 381 Final Brussels: European Commission <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0381>.
5. Jensen HG, Elleby C, Domínguez IP. Reducing the European Union's plant protein deficit: options and impacts. 2021. *Agricultural Economics (AGRICECON)*; 2021; DOI: 10.17221/94/2023-AGRICECON
6. Warsame AO, O'Sullivan DM, Tosi P. Seed storage proteins of faba bean (*Vicia faba* L): Current status and prospects for genetic improvement. *Journal of Agricultural and Food Chemistry*. 2018;66(48):12617-26. doi: 10.1021/acs.jafc.8b04992
7. Metayer N. *Vicia faba* breeding for sustainable agriculture in Europe: Identification of regional priorities and definition of target genotypes. GIE Févérole, Paris. 2004;5.
8. Maalouf F, Hu J, O'Sullivan DM, Zong X, Hamwieh A, Kumar S, et al. Breeding and genomics status in faba bean (*Vicia faba*). *Plant Breeding*. 2018;138(4):465-73. <https://doi.org/10.1111/pbr.12644>
9. Karkanis A, Ntatsi G, Lepse L, Fernandez JA, Vagen IM, Rewald B, et al. Faba Bean Cultivation - Revealing Novel Managing Practices for More Sustainable and Competitive European Cropping Systems. *Front Plant Sci*. 2018;9:1115. <https://doi.org/10.3389/fpls.2018.01115>
10. van Loon MP, Alimagham S, Pronk A, Fodor N, Ion V, Kryvoshein O, et al. Grain legume production in Europe for food, feed and meat-substitution. *Global Food Security*. 2023;39:100723. doi: 10.1016/j.gfs.2023.100723
11. Khazaei H, O'Sullivan DM, Stoddard FL, Adhikari KN, Paull JG, Schulman AH, et al. Recent advances in faba bean genetic and genomic tools for crop improvement. *Legum Sci*. 2021;3(3):e75. <https://doi.org/10.1002/leg3.75>
12. Jayakodi M, Golicz AA, Kreplak J, Fechete LI, Angra D, Bednar P, et al. The giant diploid faba genome unlocks variation in a global protein crop. *Nature*. 2023;615(7953):652-9. doi: 10.1038/s41586-023-05791-5
13. Duc G, Bao S, Baum M, Redden B, Sadiki M, Suso MJ, et al. Diversity maintenance and use of *Vicia faba* L. genetic resources. *Field Crops Research*. 2010;115(3):270-8. DOI: 10.1016/j.fcr.2008.10.003
14. Zhang H, Liu Y, Zong X, Teng C, Hou W, Li P, et al. Genetic diversity of global faba bean germplasm resources based on the 130K TNGS genotyping platform. *Agronomy*. 2023;13(3):811. <https://doi.org/10.3390/agronomy13030811>
15. Göl Ş, Doğanlar S, Frary A. Relationship between geographical origin, seed size and genetic diversity in faba bean (*Vicia faba* L.) as revealed by SSR markers. *Molecular Genetics and Genomics*. 2017;292:991-9. doi: 10.1007/s00438-017-1326-0
16. Cernay C, Ben-Ari T, Pelzer E, Meynard J-M, Makowski D. Estimating variability in grain legume yields across Europe and the Americas. *Scientific reports*. 2015;5(1):11171. <https://doi.org/10.1038/srep11171>
17. Aguilar-Benitez D, Casimiro-Soriguer I, Maalouf F, Torres AM. Linkage mapping and QTL analysis of flowering time in faba bean. *Sci Rep*. 2021;11(1):13716. <https://doi.org/10.1038/s41598-021-92680-4>
18. Avila CM, Šatović Z, Sillero JC, Nadal S, Rubiales D, Moreno MT, et al. QTL detection for agronomic traits in faba bean (*Vicia faba* L.). *Agriculturae Conspectus Scientificus*. 2005;70(3):65-73.

19. Cruz-Izquierdo S, Avila CM, Satovic Z, Palomino C, Gutierrez N, Ellwood SR, et al. Comparative genomics to bridge *Vicia faba* with model and closely-related legume species: stability of QTLs for flowering and yield-related traits. *Theor Appl Genet.* 2012;125(8):1767-82. doi: 10.1007/s00122-012-1952-1
20. Patto MV, Torres A, Koblikova A, Macas J, Cubero J. Development of a genetic composite map of *Vicia faba* using F2 populations derived from trisomic plants. *Theoretical and Applied Genetics.* 1999;98:736-43. DOI: 10.1007/s001220051129
21. Zhao N, Xue D, Miao Y, Wang Y, Zhou E, Zhou Y, et al. Construction of a high-density genetic map for faba bean (*Vicia faba* L.) and quantitative trait loci mapping of seed-related traits. *Front Plant Sci.* 2023;14:1201103. <https://doi.org/10.3389/fpls.2023.1201103>
22. Zhu C, Gore M, Buckler ES, Yu J. Status and prospects of association mapping in plants. *The plant genome.* 2008;1(1). DOI: 10.3835/plantgenome2008.02.0089
23. Skovbjerg CK, Angra D, Robertson-Shersby-Harvie T, Kreplak J, Keeble-Gagnère G, Kaur S, et al. Genetic analysis of global faba bean diversity, agronomic traits and selection signatures. *Theoretical and Applied Genetics.* 2023;136(5):114. <https://doi.org/10.1007/s00122-023-04360-8>
24. Ohm H, Åstrand J, Ceplitis A, Bengtsson D, Hammenhag C, Chawade A, et al. Novel SNP markers for flowering and seed quality traits in faba bean (*Vicia faba* L.): Characterization and GWAS of a diversity panel. *Frontiers in Plant Science.* 2023;15:1348014. doi: 10.3389/fpls.2024.1348014
25. Gutierrez N, Pegard M, Solís I, Sokolović D, Lloyd D, Howarth CJ, et al. Genome-Wide Association Study for Yield Related Traits in Faba Bean (*Vicia faba* L.). *Frontiers in Plant Science.* 2023;15:1328690. <https://doi.org/10.3389/fpls.2024.1328690>
26. Bandillo NB, Jarquin D, Posadas LG, Lorenz AJ, Graef GL. Genomic selection performs as effectively as phenotypic selection for increasing seed yield in soybean. *The Plant Genome.* 2023;16(1):e20285.
27. Crosta M, Romani M, Nazzicari N, Ferrari B, Annicchiarico P. Genomic prediction and allele mining of agronomic and morphological traits in pea (*Pisum sativum*) germplasm collections. *Frontiers in Plant Science.* 2023;14:1320506. <https://doi.org/10.3389/fpls.2023.1320506>;14:1320506.
28. Nazzicari N, Franguelli N, Ferrari B, Pecetti L, Annicchiarico P. The Effect of Genome Parametrization and SNP Marker Subsetting on Genomic Selection in Autotetraploid Alfalfa. *Genes.* 2024;15(4):449. doi: 10.3390/genes15040449
29. Murillo DA, Gezan SA, Heilman AM, Walk TC, Aparicio JS, Horsley RD. FieldHub: A shiny app for design of experiments in life sciences. *The Journal of Open Source Software.* 2021. doi: 10.21105/joss.03122
30. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25(14):1754-60. doi: 10.1093/bioinformatics/btp324
31. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:12073907.* 2012. <https://doi.org/10.48550/arXiv.1207.3907>
32. Danecek P, Bon eld JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools *Gigascience.* 2021;10(2):10.1093. doi: 10.1093/gigascience/giab008
33. VanRaden PM. Efficient methods to compute genomic predictions. *Journal of dairy science.* 2008;91(11):4414-23. doi: 10.3168/jds.2007-0980

34. Amadeu RR, Garcia AAF, Munoz PR, Ferrão LFV. AGHmatrix: genetic relationship matrices in R. *Bioinformatics*. 2023;39(7):btad445. doi: 10.1093/bioinformatics/btad445
35. Nazarian A, Gezan SA. GenoMatrix: a software package for pedigree-based and genomic prediction analyses on complex traits. *Journal of Heredity*. 2016;107(4):372-9.
36. Gezan SA, de Oliveira AA, Galli G, Murray D. User's Manual for ASRgenomics v. 1.1. 0 An R package with complementary genomic functions. 2022. doi: 10.1093/jhered/esw020
37. Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MG. Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data. *Molecular biology and evolution*. 2016;33(4):1082-93. <https://doi.org/10.1093/molbev/msv334>
38. Kassambara A. Factoextra: extract and visualize the results of multivariate data analyses. R package version. 2016;1.
39. Butler D, Cullis B, Gilmour A, Gogel B, Thompson R. ASReml-R reference manual version 4. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK. 2017.
40. Cullis BR, Smith AB, Coombes NE. On the design of early generation variety trials with correlated data. *Journal of agricultural, biological, and environmental statistics*. 2006;11:381-93. doi: 10.1198/108571106X154443
41. Burdon RD. Genetic Correlation as a Concept for Studying Genotype-Environment Interaction in Forest Tree Breeding. *Silvae Genetica*. 1977;26(5-6):168-75.
42. Kendall MG, Stuart A. The Advanced Theory of Statistics. Vol. 3. *Biometrics*. 1969;25(2):435.
43. Adetunji I, Willems G, Tschoep H, Bürkholz A, Barnes S, Boer M, et al. Genetic diversity and linkage disequilibrium analysis in elite sugar beet breeding lines and wild beet accessions. *Theoretical and applied genetics*. 2014;127:559-71. doi: 10.1007/s00122-013-2239-x.
44. Fritsch FN, Butland J. A method for constructing local monotone piecewise cubic interpolants. *SIAM journal on scientific and statistical computing*. 1984;5(2):300-4. <https://doi.org/10.1137/0905021>
45. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*. 2020;17(3):261-72. ; doi: 10.1038/s41592-019-0686-2
46. van Rossum B-J, Kruijer W, van Eeuwijk F, Boer M, Malosetti M, Bustos-Korts D, et al. Package 'statgenGWAS'. R Packag version. 2022;1.
47. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. *Genetics*. 2008;178(3):1709-23. doi: 10.1534/genetics.107.080101
48. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*. 1995;57(1):289-300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
49. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *fly*. 2012;6(2):80-92. doi: 10.4161/fly.19695

50. Perez P, de los Campos G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics*. 2014;198(2):483-95. doi: 10.1534/genetics.114.164442
51. El-Sherbeeney MH, Robertson LD. Protein content variation in a pure line faba bean (*Vicia faba*) collection. *Journal of the Science of Food and Agriculture*. 1992;58(2):193-6. <https://dx.doi.org/10.1002/jsfa.2740580206>
52. Adhikari KN, Khazaei H, Ghaouti L, Maalouf F, Vandenberg A, Link W, et al. Conventional and Molecular Breeding Tools for Accelerating Genetic Gain in Faba Bean (*Vicia Faba* L.). *Front Plant Sci*. 2021;12:744259. doi: 10.1002/leg3.75
53. Lippolis A, Roland WSU, Bocova O, Pouvreau L, Trindade LM. The challenge of breeding for reduced off-flavor in faba bean ingredients. *Front Plant Sci*. 2023;14:1286803. <https://doi.org/10.3389/fpls.2023.1286803>
54. Nurgi N, Tana T, Dechassa N, Alemayehu Y, Tesso B. On-farm diversity of faba bean (*Vicia faba* L.) farmers' varieties in Eastern Hararghe Zone, Ethiopia. *Genetic Resources and Crop Evolution*. 2023;70(2):549-70. doi: 10.1007/s10722-022-01446-y
55. Cubero JI. On the evolution of *Vicia faba* L. *Theor Appl Genet*. 1974;45(2):47-51. doi: 10.1007/BF00283475
56. Flint-Garcia SA, Thornsberry JM, Buckler ESt. Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol*. 2003;54:357-74. doi: 10.1146/annurev.arplant.54.031902.134907
57. Schlötterer C, Tobler R, Kofler R, Nolte V. Sequencing pools of individuals—mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*. 2014;15(11):749-63. doi: 10.1038/nrg3803
58. Bernardo R. Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop science*. 2008;48(5):1649-64. <https://doi.org/10.2135/cropsci2008.03.0131>
59. Dai D, Tong H, Cheng L, Peng F, Zhang T, Qi W, et al. Maize Dek33 encodes a pyrimidine reductase in riboflavin biosynthesis that is essential for oil-body formation and ABA biosynthesis during seed development. *Journal of experimental botany*. 2019;70(19):5173-87. doi: 10.1093/jxb/erz268
60. Fujioka S, Sakurai A, Yamaguchi I, Murofushi N, Takahashi N, Kaihara S, et al. Isolation and identification of L-pipecolic acid and nicotinamide as flower-inducing substances in *Lemna*. *Plant and cell physiology*. 1987;28(6):995-1003. <https://doi.org/10.1093/oxfordjournals.pcp.a077402>
61. Liu R, Cao S-K, Sayyed A, Xu C, Sun F, Wang X, et al. The mitochondrial pentatricopeptide repeat protein PPR18 is required for the cis-splicing of nad4 intron 1 and essential to seed development in maize. *International Journal of Molecular Sciences*. 2020;21(11):4047.
62. Huang J, Lu G, Liu L, Raihan MS, Xu J, Jian L, et al. The kernel size-related quantitative trait locus qKW9 encodes a pentatricopeptide repeat protein that affects photosynthesis and grain filling. *Plant physiology*. 2020;183(4):1696-709.
63. Li Z, Zhang X, Zhao K, Zhao K, Qu C, Gao G, et al. Comprehensive Transcriptome Analyses Reveal Candidate Genes for Variation in Seed Size/Weight During Peanut (*Arachis hypogaea* L.) Domestication. *Front Plant Sci*. 2021;12:666483. doi: 10.3389/fpls.2021.666483

64. Xia T, Li N, Dumenil J, Li J, Kamenski A, Bevan MW, et al. The ubiquitin receptor DA1 interacts with the E3 ubiquitin ligase DA2 to regulate seed and organ size in Arabidopsis. *The Plant Cell*. 2013;25(9):3347-59. doi: 10.1105/tpc.113.115063
65. Li Q, Li L, Yang X, Warburton ML, Bai G, Dai J, et al. Relationship, evolutionary fate and function of two maize co-orthologs of rice GW2 associated with kernel size and weight. *BMC Plant Biology*. 2010;10:1-15. <https://doi.org/10.1186/1471-2229-10-143>
66. Do THT, Martinoia E, Lee Y. Functions of ABC transporters in plant growth and development. *Current opinion in plant biology*. 2018;41:32-8. doi: 10.1016/j.pbi.2017.08.003
67. Orsi CH, Tanksley SD. Natural variation in an ABC transporter gene associated with seed size evolution in tomato species. *PLoS Genet*. 2009;5(1):e1000347. doi: 10.1371/journal.pgen.100034.
68. Ren G, Xie M, Dou Y, Zhang S, Zhang C, Yu B. Regulation of miRNA abundance by RNA binding protein TOUGH in Arabidopsis. *Proceedings of the National Academy of Sciences*. 2012;109(31):12817-21. doi: 10.1073/pnas.1204915109
69. Lee Y-RJ, Qiu W, Liu B. Kinesin motors in plants: from subcellular dynamics to motility regulation. *Current opinion in plant biology*. 2015;28:120-6. doi: 10.1016/j.pbi.2015.10.003
70. Yang XY, Wang Y, Jiang WJ, Liu XL, Zhang XM, Yu HJ, et al. Characterization and expression profiling of cucumber kinesin genes during early fruit development: revealing the roles of kinesins in exponential cell production and enlargement in cucumber fruit. *Journal of experimental botany*. 2013;64(14):4541-57. doi: 10.1093/jxb/ert269
71. Lazzaro MD, Wu S, Snouffer A, Wang Y, Van der Knaap E. Plant organ shapes are regulated by protein interactions and associations with microtubules. *Frontiers in plant science*. 2018;9:1766. doi: 10.3389/fpls.2018.01766.
72. Shlykov MA, Zheng WH, Chen JS, Saier Jr MH. Bioinformatic characterization of the 4-Toluene Sulfonate Uptake Permease (TSUP) family of transmembrane proteins. *Biochimica et Biophysica Acta (BBA)-Biomembranes*. 2012;1818(3):703-17. doi: 10.1016/j.bbamem.2011.12.005
73. Jeh HE, Sanchez R, Beltrán J, Yang X, Kundariya H, Wamboldt Y, et al. Sensory plastid-associated PsbP DOMAIN-CONTAINING PROTEIN 3 triggers plant growth-and defense-related epigenetic responses. *The Plant Journal*. 2023;115(2):414-33. doi: 10.1111/tpj.16233
74. Asano T, Yoshioka Y, Kurei S, Sakamoto W, Sodmergen, Machida Y. A mutation of the CRUMPLED LEAF gene that encodes a protein localized in the outer envelope membrane of plastids affects the pattern of cell division, cell differentiation, and plastid division in Arabidopsis. *The Plant Journal*. 2004;38(3):448-59. doi: 10.1111/j.1365-313X.2004.0205
75. Zhang J, Li C, Zhang W, Zhang X, Mo Y, Tranquilli GE, et al. Wheat plant height locus RHT25 encodes a PLATZ transcription factor that interacts with DELLA (RHT1). *Proceedings of the National Academy of Sciences*. 2023;120(19):e2300203120. doi: 10.1073/pnas.2300203120
76. Singh DK, Maximova SN, Jensen PJ, Lehman BL, Ngugi HK, McNellis TW. FIBRILLIN4 is required for plastoglobule development and stress resistance in apple and Arabidopsis. *Plant Physiology*. 2010;154(3):1281-93. doi: 10.1104/pp.110.164095
77. Iizasa S, Iizasa Ei, Matsuzaki S, Tanaka H, Kodama Y, Watanabe K, et al. Arabidopsis LBP/BPI related-1 and-2 bind to LPS directly and regulate PR1 expression. *Scientific Reports*. 2016;6(1):27527. doi:

78. Ma S, Guo Y, Liu D, Zhang X, Guo J, Zhang T, et al. Genome-Wide Analysis of the Membrane Attack Complex and Perforin Genes and Their Expression Pattern under Stress in the Solanaceae. *International Journal of Molecular Sciences*. 2023;24(17):13193. <https://doi.org/10.3390/ijms241713193>
79. Yao X, Nie J, Bai R, Sui X. Amino acid transporters in plants: Identification and function. *Plants*. 2020;9(8):972. doi: 10.3390/plants9080972
80. Spooner DM, van Treuren R, De Vicente M. Molecular markers for genebank management: Bioversity International; 2005. <https://edepot.wur.nl/202151>
81. Annicchiarico P, Nazzicari N, Ferrari B. Genetic and genomic resources in white lupin and the application of genomic selection. *The Lupin Genome*. 2020:139-49. doi:10.1007/978-3-030-21270-4_10
82. Combs E, Bernardo R. Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. *The Plant Genome*. 2013;6(1):plantgenome2012.11.0030. <https://doi.org/10.3835/plantgenome2012.11.0030>
83. Bernardo R. Genomewide selection when major genes are known. *Crop Science*. 2014;54(1):68-75. doi:10.2135/cropsci2013.05.0315
84. Gezan SA, Osorio LF, Verma S, Whitaker VM. An experimental validation of genomic selection in octoploid strawberry. *Horticulture research*. 2017;4. <https://doi.org/10.1038/hortres.2016.70>

Tables

Tables 4 and 5 are available in the Supplementary Files section.

Figures

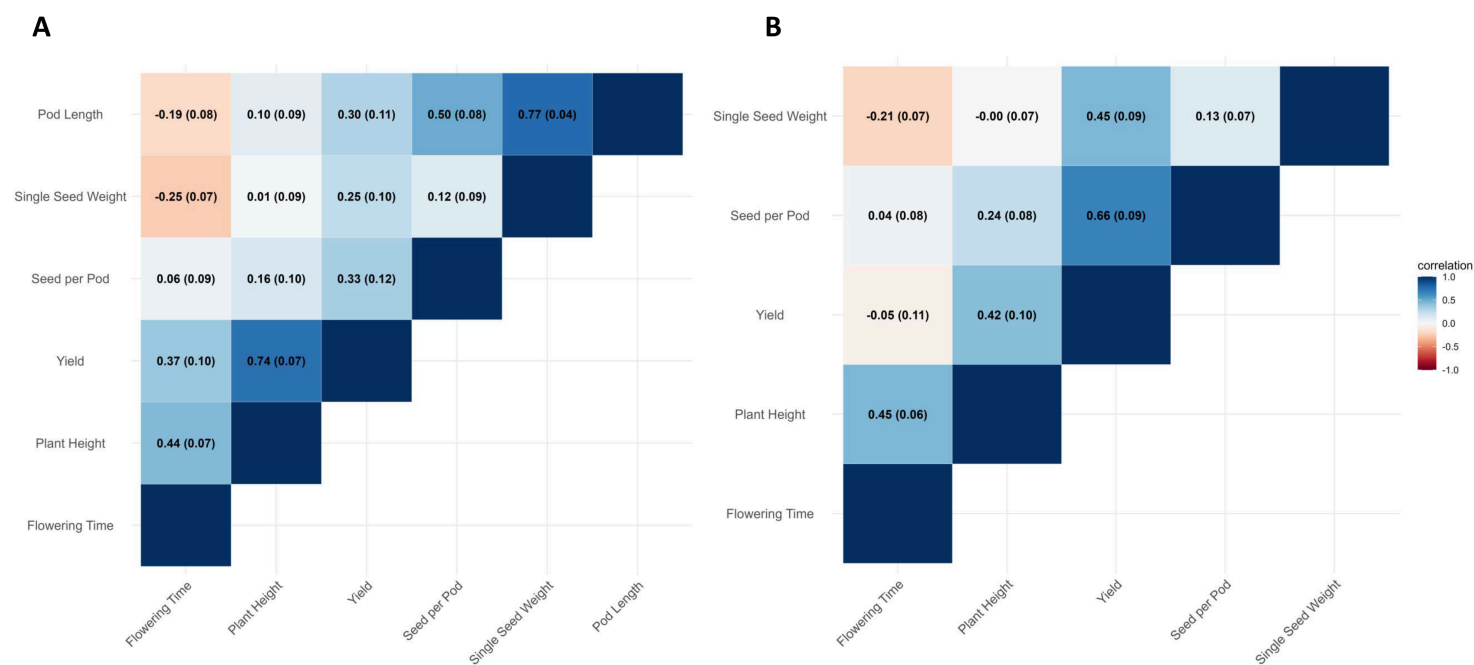


Figure 1

Heatmap showing type-A genetic correlations between traits for 2021 (A) and 2022 (B). The numbers in parentheses represent the approximated standard errors.

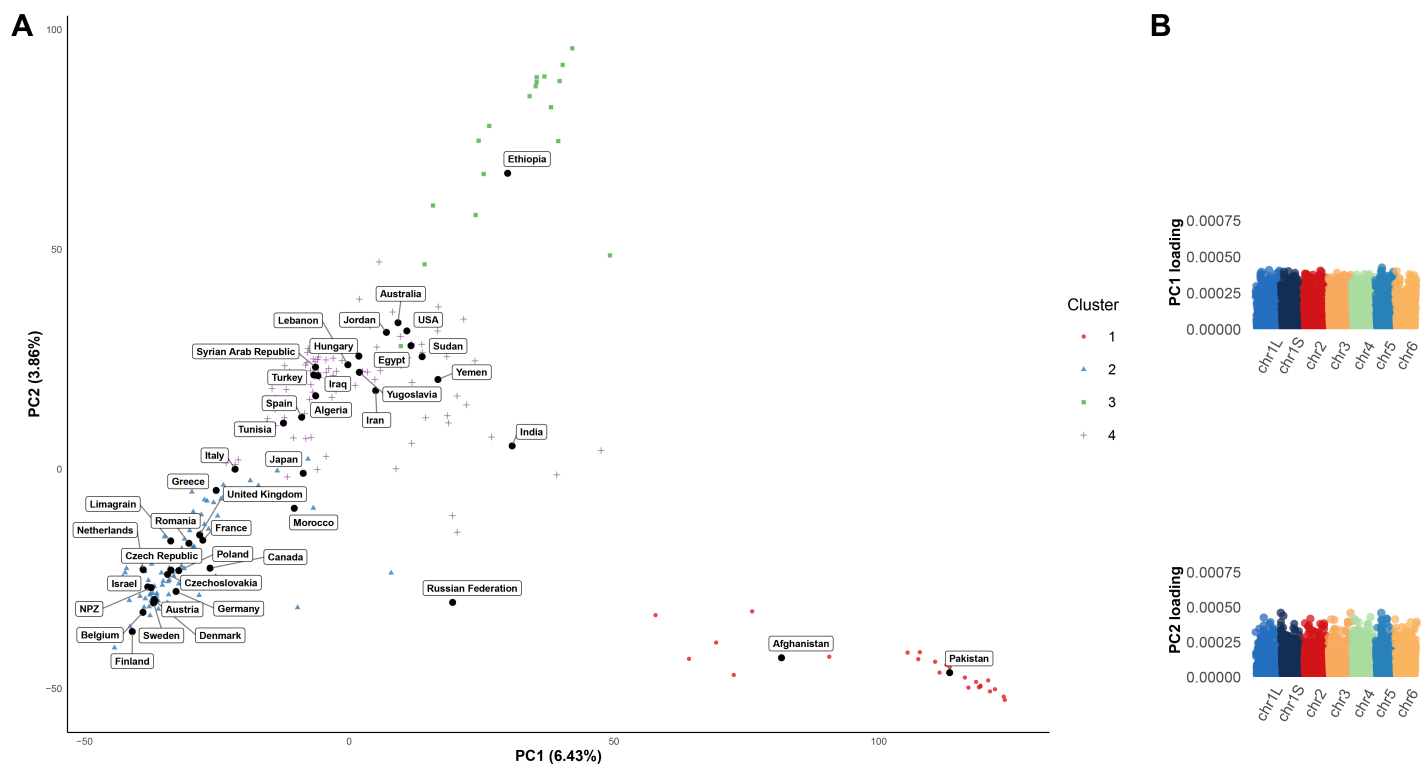


Figure 2

A) Genetic variation of the CGN population explained by two principal components. The black dots represent the median PC1 and PC2 values for accessions from the same country. B) Manhattan plot displaying the squared loadings for PC1 and PC2 of each SNP relative to its genomic positions.

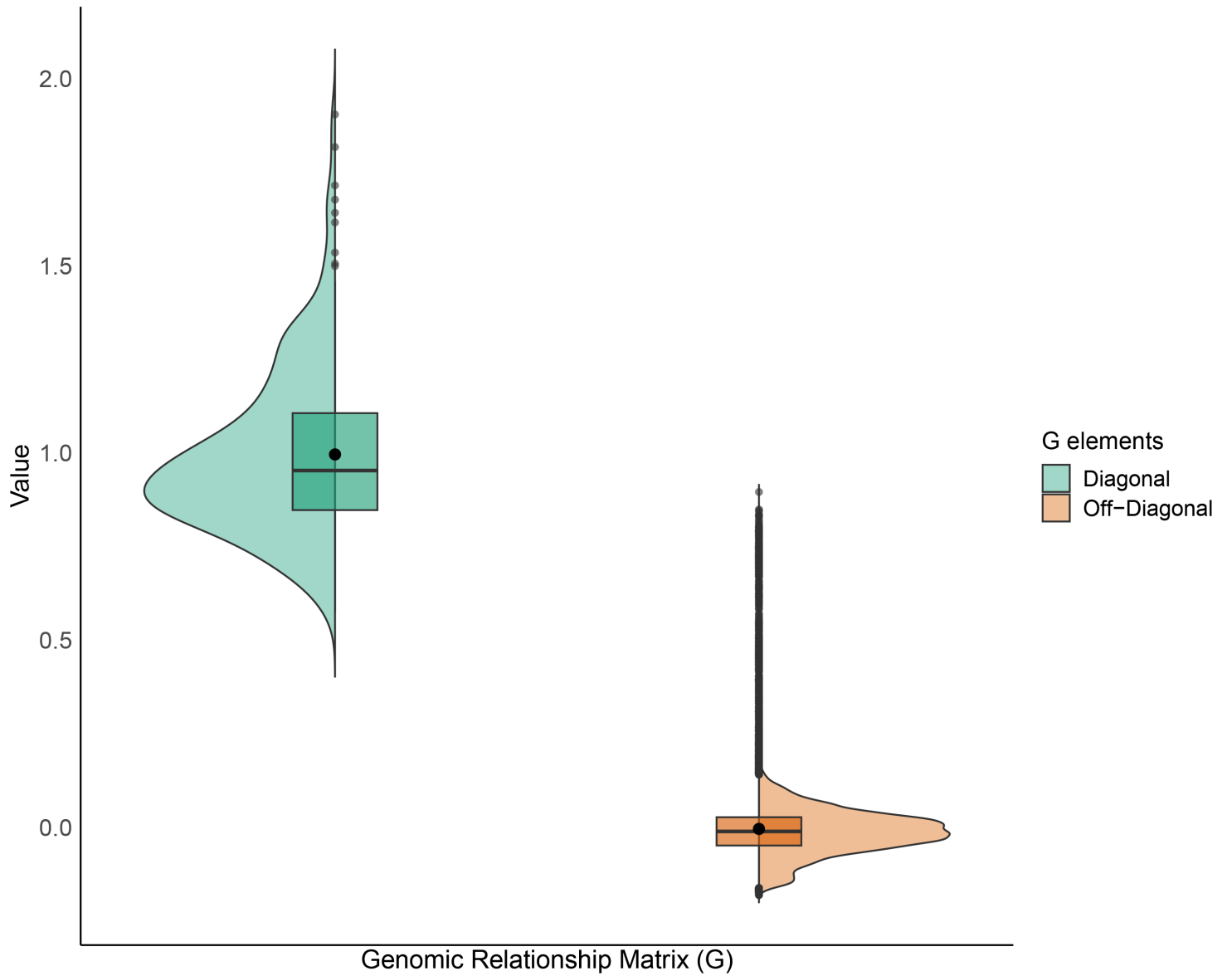


Figure 3

Box plots representing the diagonal and off-diagonal values of the Genomic Relationship Matrix (**G**). The horizontal lines in the box plots represent the median value, while the dots represent the mean value.

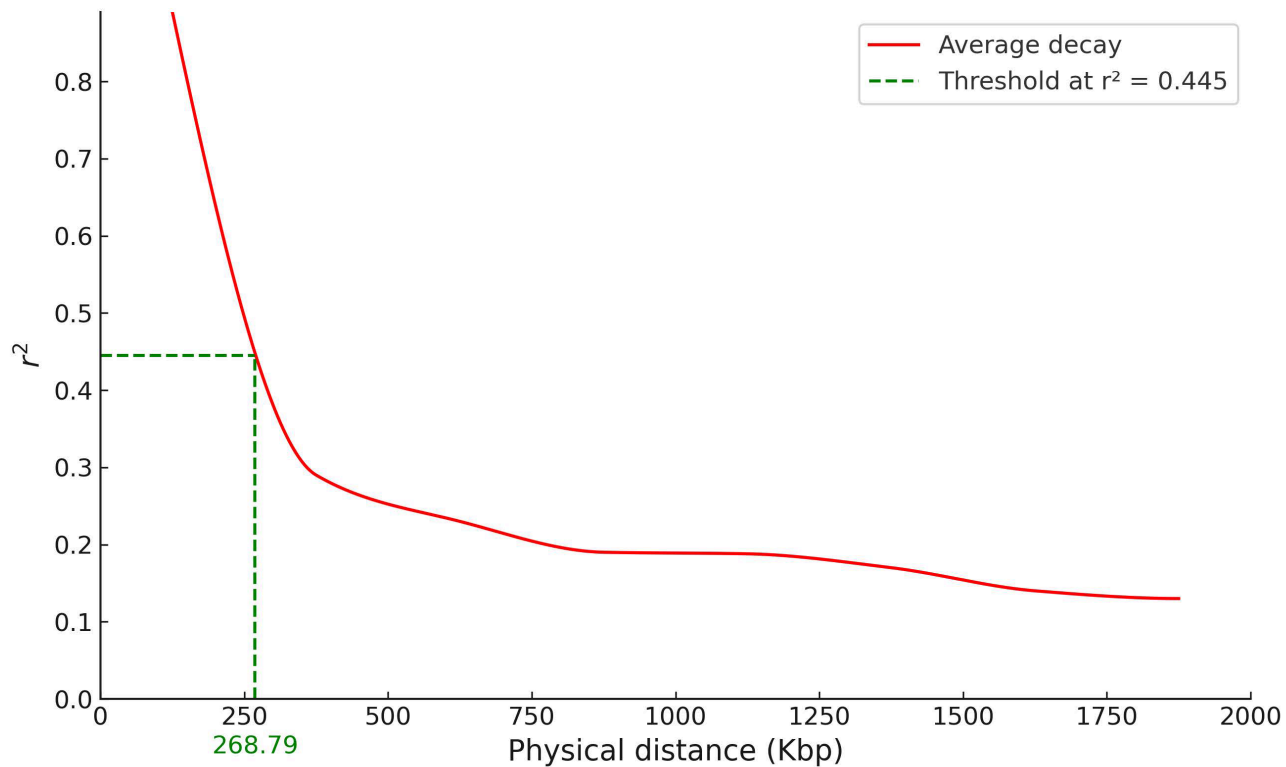


Figure 4

Average linkage disequilibrium (LD) decay along the faba bean genome. The green dashed line represents the physical distance (~268.79 Kbp) where the r^2 decreases to half of its maximum value.

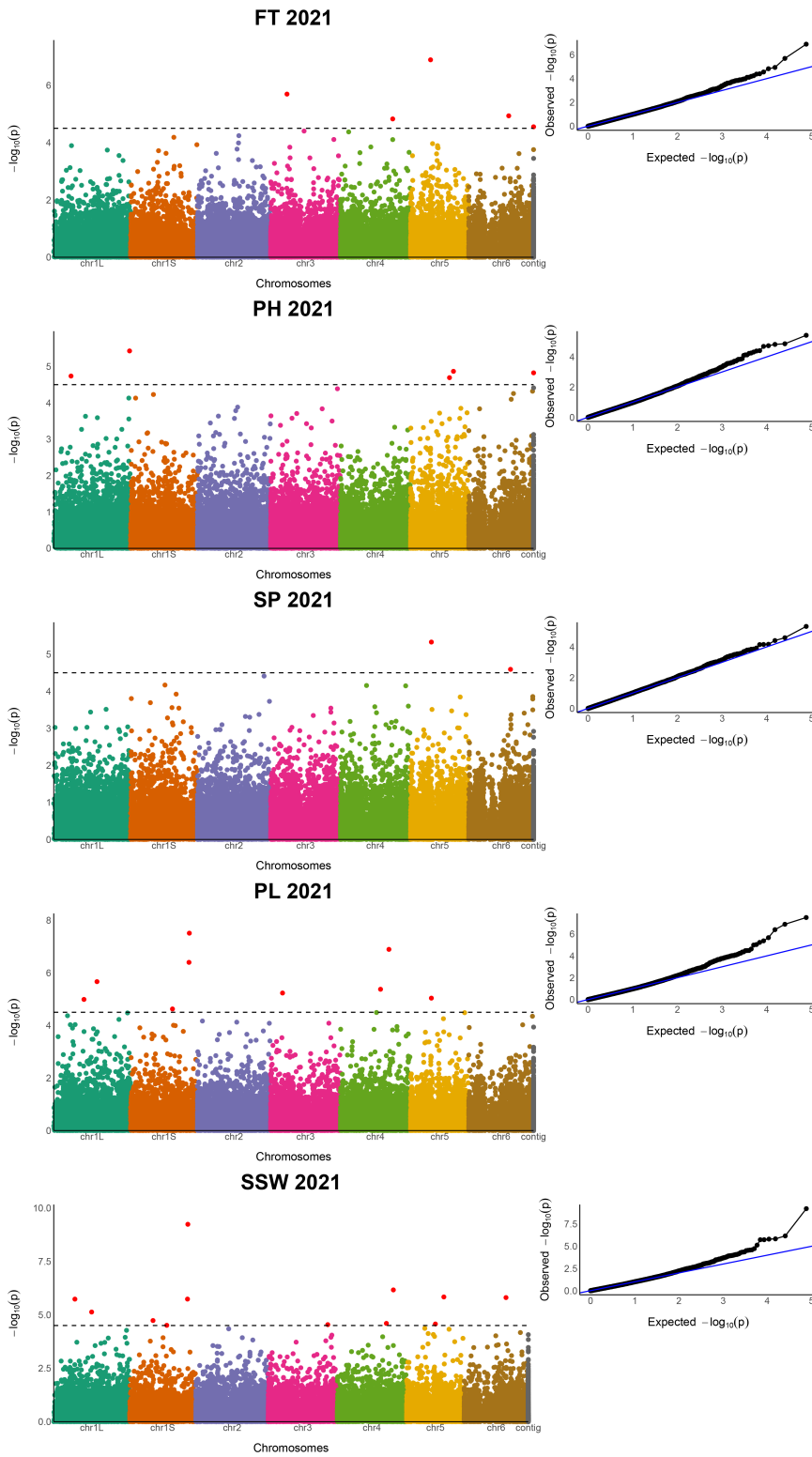


Figure 5

Manhattan and Quantile-Quantile (QQ) plots from the 2021 GWAS. The Manhattan plots (left) and the QQ plots (right size) are arranged vertically in the following order: flowering time (FT), plant height (PH), seeds per pod (SP), pod length (PL), and single seed weight (SSW). Significant SNPs exceeding the LOD threshold of 4.5 are highlighted with red dots.

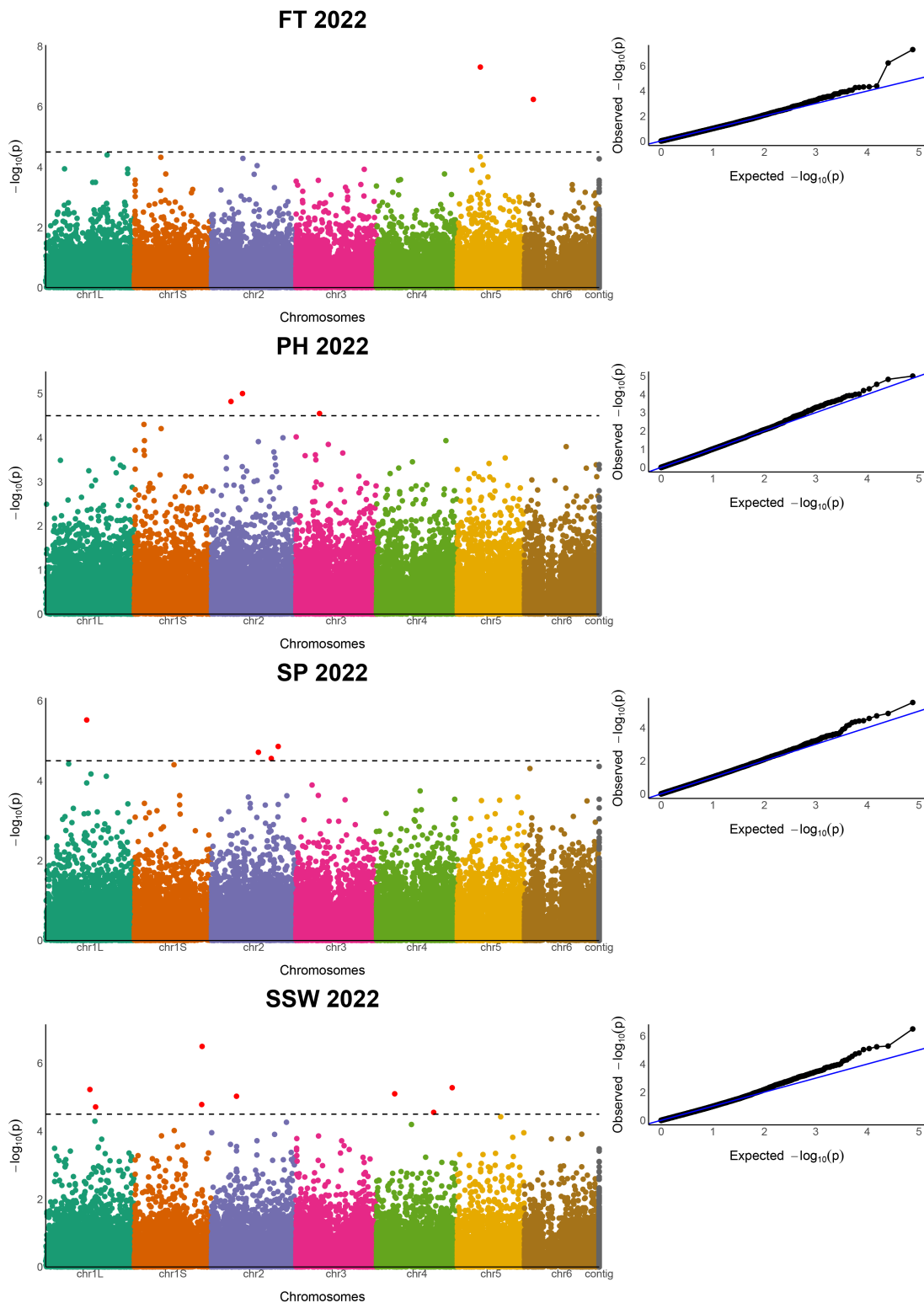


Figure 6

Manhattan and Quantile-Quantile (QQ) plots from the 2022 GWAS. The Manhattan plots (left) and the QQ plots (right size) are arranged vertically in the following order: flowering time (FT), plant height (PH), seeds per pod (SP), and single seed weight (SSW). Significant SNPs exceeding the LOD threshold of 4.5 are highlighted with red dots.

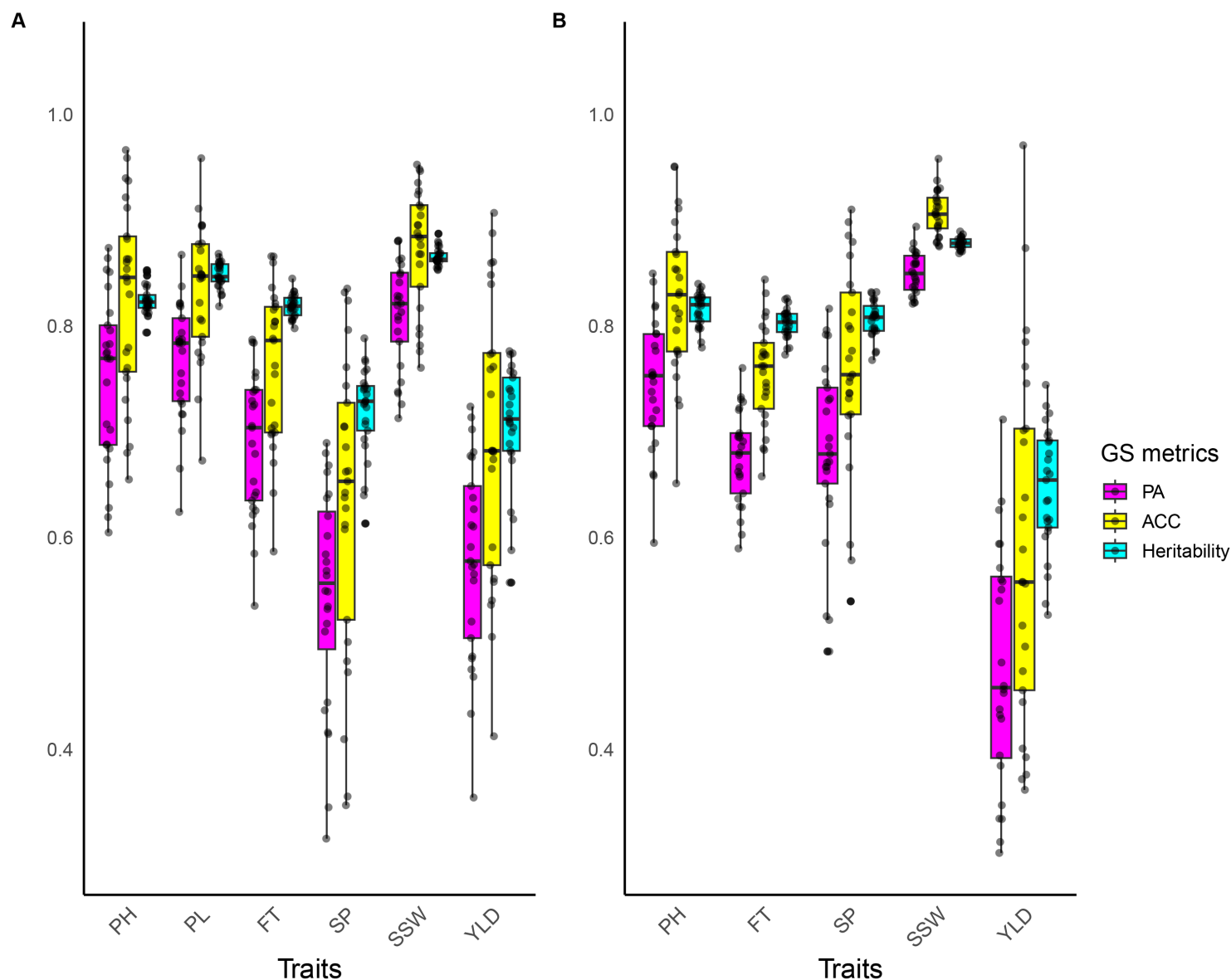


Figure 7

Box plots depicting the predictive ability (PA), prediction accuracy (ACC), and heritability of the cross-validation (CV) in 2021 (A) and 2022 (B). Each point in the box plot represents one round of the 5-fold CV repeated 10 times.

Supplementary Files

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

- [Table4.GWASResult2021bigtable.xlsx](#)
- [Table5.GWASResult2022bigtable.xlsx](#)
- [SupplementaryTable1.CGNPassportData.xlsx](#)
- [SupplementaryTable2.GeneticClustersbycountry.xlsx](#)
- [SupplementaryTable3.docx](#)

- [SupplementaryTable4.andTableS5Completegenelist20212022.xlsx](#)
- [SupplementaryFiguresS15.docx](#)